The SAGE Handbook of Spatial Analysis

Edited by
A. Stewart Fotheringham
Peter A. Rogerson

The SAGE Handbook of Spatial Analysis

The SAGE Handbook of Spatial Analysis

Edited by
A. Stewart Fotheringham and
Peter A. Rogerson

SAGE
Los Angeles • London • New Delhi • Singapore

First published 2009

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, Post Bag 7
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
33 Pekin Street #02-01
Far East Square
Singapore 048763

Library of Congress Control Number: 2008921399

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-4129-1082-8

Typeset by CEPHA Imaging Pvt. Ltd., Bangalore, India
Printed in Great Britain by The Cromwell Press Ltd, Trowbridge, Wiltshire
Printed on paper from sustainable resources

Acknowledgement: Research presented in Chapter 13 was supported by a grant to the National Centre for Geocomputation by Science Foundation ireland (03/RP1/1382) and by a Strategic Research Cluster grant (07/SRC1/1168) from Science Foundation ireland under the National Development Plan. The author gratefully acknowledges this support.

Contents

Notes on Contributors

Andrew B. Lawson is Professor in the Department of Biostatistics, Bioinformatics and Epidemiology at the Medical University of South Carolina. His research interests focus on the development of statistical methods in spatial and environmental epidemiology, disease surveillance, health data mining and space–time problems, directional statistics and environmetrics. Professor Lawson is a WHO advisor in Disease Mapping and Risk Assessment. He has a wide range of publications in the area of epidemiology, including 8 books and over 20 book chapters, and he gave numerous invited courses on spatial epidemiology in the UK, Sweden, Belgium, New Zealand, Australia, Canada and the USA.

Atsuyuki Okabe received his PhD from the University of Pennsylvania and the degree of Doctor of Engineering from the University of Tokyo. He is currently Professor of the Department of Urban Engineering at the University of Tokyo where he served as Director of the Center for Spatial Information Science (1998–2005). Professor Okabe's research interests include GIS, spatial analysis, and spatial optimization, and he has published many papers in journals, books and conference proceedings on these topics. He is a co-author of *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams* (John Wiley, 2000), the editor of *GIS-based Studies in the Humanities and Social Sciences* (Taylor & Francis, 2005). He serves on the Editorial Boards of seven international journals including *International Journal of Geographical Information Science.*

Chris Brunsdon is Professor of geographic information at the Department of Geography, Leicester University. His research interests include the methodologies underlying spatial statistical analysis and GIS. In particular he is interested in the analysis of crime patterns, house prices and health-related data. Professor Brunsdon has played a role in the development of Geographically Weighted Regression, a technique of analysis that models geographical variations in the relationships between variables. He is member of Editorial Board of *Computers Environment and Urban Systems.*

Christopher D. Lloyd is Lecturer in the School of Geography, Archaeology and Palaeoecology at the Queen's University, Belfast. His research interests focus on the analysis of spatial data (in both social and environmental contexts), geostatistics, spatial analysis, remote sensing and archaeology. His research key concern has been with the use and development of local models for spatial analysis. Dr Lloyd is author *of Local Models for Spatial Analysis* (Boca Raton: CRC

Press, 2006) and co-author of *The Atlas of the Island of Ireland: Mapping Social and Economic Change* (Maynooth: AIRO/ICLRD, 2008).

David Martin is Professor in the School of Geography, University of Southampton. He is director of the ESRC Census Programme and a co-director of the ESRC National Centre for Research Methods. His research interests include socioeconomic application of GIS, census population modeling, census geography design and geography of health. He is co-editor of *GIS and Geocomputation* (Taylor and Francis, 2000), *The Census Data System* (Wiley, 2002) and *Methods in Human Geography: a Guide for Students Doing a Research Project*, Second Edition (Pearson, 2005). Professor Martin is a member of the editorial advisory boards of *Transactions in GIS, Transactions of the Institute of British Geographers* and associate editor of the *Journal of the Royal Statistical Society – Series A: Statistics in Society.*

David W. Wong is Professor in the Department of Geography and Geoinformation Science, at George Mason University. His research interests fall into three main categories: investigating the modifiable areal unit problem (MAUP) effects; spatial dimensions of segregation and ethnic diversity; and GIS applications of spatial analytical techniques. He has co-authored two books: *Statistical Analysis with ArcView* (Wiley & Sons, 2001) and *Statistical Analysis and Modeling of Geographic Information* (Wiley & Sons, 2005). He has served as reviewer for various journals, funding agencies and organizations. He is on the editorial board of several journals: Computers, Environment and Urban Systems, Geographical Analysis, and Journal of Geographic Information Sciences.

Dimitris Ballas is a Senior Lecturer in the Department of Geography at the University of Sheffield. He received his PhD in Geography from the University of Leeds in 2001. His research interests include economic geography; spatial dimensions of socio-economic polarisation and income and wealth inequalities; socio-economic applications of GIS; geographical impact of area-based and national social policies; basic income policies; and social justice; geography of happiness and well-being. He is the lead author of the book "Geography matters: simulating the impacts of national social policies" and a co-author of the books "Post-Suburban Europe: Planning and Politics at the Margins of Europe's Capital Cities" and "Poverty, wealth and place in Britain, 1968 to 2005". He has fifteen papers in peer-reviewed international academic journals, eight peer-reviewed edited book chapters and over fifty conference papers.

Eric Delmelle is Assistant Professor in the Geography and Earth Sciences Department at the University of North Carolina (Charlotte) where he teaches GIS, geovisualization and spatial optimization. He received his PhD in geography from the State University of New York at Buffalo. His research interests focus on spatial sampling optimization and geostatistics, non-linear allocation problems, geovisualization and GIS.

Geoffrey M. Jacquez is President of BioMedware Incorporated, and Adjunct Associate Professor of Environmental Health Sciences at the University of Michigan. He received his PhD from the Department of Ecology and Evolution at State University of New York at Stony Brook. Dr. Jacquez develops and applies spatial statistics to elucidate underlying space–time processes in the environmental, biological and health sciences. His research includes applications in disease clustering, epidemiology, environmental monitoring and

population genetics. Dr. Jacquez is currently Principal Investigator on three grants from the National Cancer Institute to develop spatial statistical methods and software. He also publishes extensively in the fields of spatial statistics, GIS and epidemiology.

Graham Clarke is Professor in the School of Geography, Faculty of Environment at the University of Leeds. His research interests include GIS, urban services, retail and business geography, urban modelling and geography of crime, income and welfare. Dr Clarke is co-author of *Geography Matters* (Joseph Rowntree Foundation, 2005), and *Retail Geography and Intelligent Network Planning* (Wiley, Chichester, 2002). He is committee member of The Academy of Learned Societies for the Social Sciences, Executive Director of Regional Science Association International and serves on the Editorial Board *of European Journal of Geography.*

Harvey J. Miller is Professor and Chair of the Department of Geography at the University of Utah. His research and teaching interests include GIS, spatial analysis and geocomputational techniques applied to understanding how transportation and communication technologies shape individual lives and urban morphology. Since 1989, he has published approximately 50 papers in journals, books and conference proceedings on these topics. He is co-author of *Geographic Information Systems for Transportation: Principles and Applications* (Oxford University Press, 2001) and co-editor of *Geographic Data Mining and Knowledge Discovery* (Taylor and Francis, 2001) and *Societies and Cities in the Age of Instant Access* (Springer, 2007). Harvey serves on the editorial boards of several scientific journals and in 2005–2011 he is serving as co-Chair of the Transportation Research Board, Committee on Spatial Data and Information Science of U.S. National Academies.

Jaymie R. Meliker is Assistant Professor of Preventive Medicine in the Medical Center at State University of New York at Stony Brook. He received his PhD in 2006 from the Department of Environmental Health Sciences, University of Michigan School of Public Health. Dr. Meliker's research contributes to the fields of exposure science, GIScience, health geography, and environmental epidemiology by developing methodologies for integrating sources of spatial, temporal, and spatio-temporal variability in environmental health applications. Prior to joining Stony Brook, he worked as a Research Scientist at BioMedware, Inc., pioneering the development of spatio-temporal software and statistical algorithms for addressing public health concerns.

Luc Anselin is Faculty Excellence Professor and Director of the Spatial Analysis Laboratory in the Department of Geography at the University of Illinois, Urbana-Champaign. He is also a Senior Research Associate at the National Center for Supercomputing Applications at UIUC. Dr. Anselin's research deals with various aspects of spatial data analysis and geographic information science, ranging from exploratory spatial data analysis to geocomputation, spatial statistics and spatial econometrics. He has published widely on topics dealing with spatial and regional analysis, including a much cited book on *Spatial Econometrics* (Kluwer, 1988); over a hundred refereed journal articles and book chapters, as well as a large number of reports and technical publications.

Lance A. Waller is Professor in the Department of Biostatistics at the Rollins School of Public Health, Emory University. His research involves the development of statistical methods

to analyze spatial and spatio-temporal patterns. His recent areas of interest include spatial point process methods in alcohol epidemiology and conservation biology (sea turtle nesting patterns), and hierarchical models in disease ecology. Dr Waller is Chair of American Statistical Association – Section on Statistics and the Environment (2008). He is also President-Elect of International Biometric Society – Eastern North American Region (2008), and serves as Associate Editor of *Biometrics, Bayesian Analysis.*

Manfred M. Fischer is Professor of Economic Geography and Director of Institute for Economic Geography and GIScience at Vienna University of Economics and Business. His research spans a broad array of fields including regional and urban economics, housing and labor market research, transportation systems analysis, innovation economics, spatial behavior and decision processes, spatial analysis and spatial statistics, and GIS. He is one of the leading scholars in the field of GeoComputation. Based on the expertise and the scientific impact Dr. Fischer has gained a high reputation both nationally and internationally. In 1995 he was elected as a member of the International Eurasian Academy of Sciences, in 1996 as a corresponding member of the Austrian Academy of Sciences and in 1999 as a foreign member of the Royal Netherlands Academy of Arts and Sciences. Dr. Fischer has published over 250 scientific publications, including 28 monographs and edited books.

Marie-Josée Fortin is Professor in the Department of Ecology and Evolutionary Biology and Head of Landscape Ecology Laboratory at the University of Toronto, Canada. She has four main research areas: spatial ecology, forest ecology, landscape ecology and spatial statistics. She is co-author of *Spatial Analysis: A Guide for Ecologists* (Cambridge University Press, 2005), and has published over 150 research papers in peer-reviewed journals, book chapters, conferences, and invited lectures. Professor Fortin is assistant editor for *Theoretical Ecology,* subject editor for *Ecology* and *Ecology Monographs*, and Editorial Board member of *Ecosystems.*

Mark Dale is Professor in the Department of Biological Science and Dean in the Faculty of Graduate Studies and Research at the University of Alberta, Canada. He received his PhD from Dalhousie University, Canada. His current research interests involve methods for detecting and analyzing the spatial relationships of plants in populations and communities and spatial analysis and spatial statistics with applications in ecological systems. Professor Dale is co-author of *Spatial Analysis: A Guide for Ecologists.* (Cambridge University Press, 2005) and he served as an associate editor for *Canadian Journal of Botany.*

Michael F. Goodchild is Professor of Geography at the University of California, Santa Barbara. He also serves as Chair of the Executive Committee for the National Center for Geographic Information and Analysis (NCGIA), and Director of NCGIA's Center for Spatially Integrated Social Science. His current research interests center on GIS, spatial analysis, the future of the library, and uncertainty in geographic data. His major publications include *Geographical Information Systems: Principles and Applications* (1991); *Environmental Modeling with GIS* (1993); *Accuracy of Spatial Databases* (1989); *GIS and Environmental Modeling: Progress and Research Issues* (1996); *Scale in Remote Sensing and GIS* (1997); *Interoperating Geographic Information Systems* (1999); and *Geographical Information Systems: Principles, Techniques, Management and Applications* (1999); in addition he is author of some 300 scientific papers.

He is the recipient of numerous awards including the Educator of the Year Award from the University Consortium for Geographic Information Science, a Lifetime Achievement Award from Environmental Systems Research Institute, Inc., the American Society of Photogrammetry and Remote Sensing Intergraph Award and the Horwood Critique Prize of the Urban and Regional Information Systems Association.

Morton E. O'Kelly is Professor and Chair of the Department of Geography at the Ohio State University. His research interests include location theory, transportation, network design and optimization, spatial analysis and GIS. Dr. O'Kelly co-authored two books: *Geography of Transportation*, 2nd edition (Prentice Hall, 1996) and *Spatial Interaction Models: Formulations and Applications* (Kluwer Academic: Amsterdam, 1989), as well as over 75 research papers in peer-reviewed journals, book chapters and conference proceedings.

Peter M. Atkinson is Professor and Head of School of Geography and Director of the University Centre for Geographical Health Research at the University of Southampton. His research interests focus on geostatistics, spatial statistics, remote sensing, and spatially distributed dynamic modelling applications for environmental problems and hazards. He is co-editor of *International Journal of Remote Sensing Letters* and associate editor of *International Journal of Applied Earth Observation and Geoinformation.* Professor Atkinson is also Fellow of the Royal Geographical Society and Fellow of the Royal Statistical Society.

Peter A. Rogerson is Professor of Geography and Biostatistics at the University at Buffalo. His research interests are in the area of demography and population change, epidemiology, spatial statistics and spatial analysis. His current work is focused upon the development of new methods for the quick detection of newly emergent clusters in geographic data. Professor Rogerson has authored and co-authored four books with the most recent *Statistical Detection and Monitoring of Geographic Clusters* (Chapman and Hall/CRC, 2008), and over 85 research papers in peer-reviewed journals. He also developed *GeoSurveillance 1.1: Software for Monitoring Change in Geographic Patterns.*

Pusheng Zhang is currently with the Microsoft Virtual Earth team. He received his PhD in Computer Science from the University of Minnesota. His research interests include local search engine design, spatial and temporal databases, data mining and geographic information retrieval. Dr Zhang is a member of the IEEE Computer Society.

Ranga Raju Vatsavai received his PhD in Computer Science from the University of Minnesota where he also worked as Research Fellow in Remote Sensing Laboratory. Currently Dr Vatsavai is employed at the Oak Ridge National Laboratory. His broad research interests are centered on spatial and spatio-temporal databases and data mining.

Reginald G. Golledge is a Professor of Geography at the University of California, Santa Barbara. His research interests include behavioral geography, spatial cognition, cognitive mapping, individual decision-making, household activity patterns and the acquisition and use of spatial knowledge across the life-span. Professor Golledge has written or edited 14 books, more than 50 chapters in books, and over 120 papers in academic journals, and 80 miscellaneous

publications including technical reports, book reviews, and published research notes. He has presented more than 100 papers at local, regional, national, and international conferences in geography, regional science, planning, psychology, and statistics. Professor Golledge received an Association of American Geographers (AAG) Honors Award in 1981. He is an Honorary Life-Time Member of the Institutes of Australian Geographers and a Fellow of the American Association for the Advancement of Science. He received an International Geography Gold Medal Award from the LAG in 1999. In 1998 he was elected Vice President of the AAG; in 1999–2000 he was elected AAG President.

Robin A. Dubin is Professor of Economics in the Weatherhead school of Management at Case Western Reserve University. Her research interests involve urban and regional economics, real estate, and spatial econometrics. Professor Dubin has published numerous research papers in peer-reviewed journals, book chapters, conferences, and invited lectures.

Robert Haining is Professor of Human Geography at Cambridge University. Between 2002 and 2007 he was Head of Geography Department at Cambridge University. Professor Haining has published extensively in the field of spatial data analysis, with particular reference to applications in the areas of economic geography, medical geography and the geography of crime. His interests also include the integration of spatial data analysis with GIS and he developed SAGE, a software system for analysing area health data. His previous book, *Spatial Data Analysis in the Social and Environmental Sciences* (Cambridge University Press, 1993) was well received and cited internationally. Professor Haining is a member of the editorial board of *Journal of Geographical Systems and Computational Statistics.*

Shashi Shekhar is McKnight Distinguished University Professor in the University of Minnesota, Minneapolis, MN, USA. His research interests include spatial databases, spatial data mining, GIS, and intelligent transportation systems. He is a co-author of a textbook on *Spatial Databases* (Prentice Hall, 2003), co-edited the *Encyclopedia of GIS*, (Springer, 2008) and has published over 200 research papers in peer-reviewed journals, books, conferences, and workshops. He is co-Editor-in-Chief of *Geo-Informatica: An International Journal on Advance in Computer Science for GIS* and has served on the editorial boards of *Transactions on Knowledge* and *Data Engineering.*

A. Stewart Fotheringham is Science Foundation Ireland Research Professor and Director of the National Centre for Geocomputation at the National University of Ireland in Maynooth. His research interests include: the integration of spatial analysis and GIS; spatial statistics, exploratory spatial data analysis and spatial modelling. He is one of the originators of Geographically Weighted Regression. Professor Fotheringham is a founding editor of *Transactions in GIS* and is on a number of editorial boards, including *Geographical Analysis, Annals of the Association of American Geographers* and *Geographical Systems.* He has published six books, over 20 book chapters and over 100 journal articles.

Sudipto Banerjee is Associate Professor in the Division of Biostatistics at the University of Minnesota. He received his PhD in Statistics from the University of Connecticut, Storrs. Dr Banerjee's research focuses upon the analysis of data arising from spatial processes,

Bayesian interpolation and prediction (methods and smoothness of spatial processes. He is also interested in the application of Bayesian methodology in biostatistics. Dr. Banerjee has co-authored a textbook on spatial statistics, *Hierarchical Modeling and Analysis for Spatial Data* (CRC Press/Chapman and Hall, 2004), was a field editor for the *Encyclopedia of GIS* (Springer, 2008) and serves as Associate Editor for *Journal of the Royal Statistical Society Series C (Applied Statistics), Statistics in Medicine* and *Bayesian Analysis.*

Urška Demšar is a lecturer at the National Center for Geocomputation at the National University of Ireland, Maynooth. She has a PhD in Geoinformatics from the Royal Institute of Technology, Stockholm, Sweden. Her research interests include Geovisual Analytics and Geovisualisation. She is combining computational and statistical methods with geovisualisation for knowledge discovery from spatial data. Additionally, she is interested in spatial analysis and mathematical modelling of spatial phenomena. She has an established cooperation with researchers at the Helsinki University of Technology with whom she is working on spatial analysis of networks for crisis management.

Vijay Gandhi is Masters Student in Computer Science at the University of Minnesota, Twin Cities. After graduating from Computer Science and Engineering at Madras University he worked in the field of business intelligence and data warehousing. Currently he is involved in research on spatial databases and spatial data mining.

Vincent B. Robinson is Associate Professor in the Department of Geography and Planning at the University of Toronto at Mississauga. His research interests involve intelligent geographic information systems, geographical modelling, and remote sensing, land use/cover change, biogeography and landscape ecology. Professor Robinson published extensively on issues and challenges of incorporating fuzzy sets technique in ecological modeling. He is also Executive Committee Member of Project Open Source, Open Access at the University of Toronto.

Toshiaki Satoh is currently a researcher in Research & Development Center of PASCO Corporation, a surveying and GIS consulting company. He received a Bachelor's degree from Tohoku University in 1992, a Master's degree from Tokyo Institute of Technology in 1994 and Ph.D. of Eng. degree from the University of Tokyo in 2007, respectively. His main interests of research are network spatial analysis and computer visualization.

1

Introduction

A. Stewart Fotheringham
and Peter A. Rogerson

1.1. WHAT IS SPATIAL ANALYSIS?

Spatial data contain locational information as well as attribute information. That is, they are data for which some attribute is recorded at different locations and these locations are coded as part of the data. Spatial analysis is a general term to describe a technique that uses this locational information in order to better understand the processes generating the observed attribute values.

Spatial analysis is important because it is increasingly recognized that most data are spatial. Examples of common types of spatial data include census data, traffic counts, patient records, the incidence of a disease, the location of facilities and services, the addresses of school pupils, customer databases, and the distributions of animal, insect or plant species. Along with various attributes collected by hand or by different types of sensors, location is also being captured by an increasing variety of technologies such as GPS, WiMAX, LiDAR, and radio frequency identity (RFID) tags as well as by more traditional means such as surveys and censuses. Some of the resulting data sets can be extremely large. Satellites, for example, regularly record terrabytes of spatial data; LiDAR scanners can capture millions of geocoded data points in minutes; and GPS-encoded spatial video generally produces 24 frames per second each of which may have around a million geocoded pixels. The world is rapidly becoming one large spatial sensor with RFID tags, CCTV cameras and GPS linked devices recording the location of objects, animals and people.

Spatial data and the processes generating such data have several properties that distinguish them from their aspatial

equivalents. Firstly, the data are typically not independent of each other. Attribute values in nearby places tend to be more similar than are attribute values drawn from locations far away from each other. This is a useful property when it comes to predicting unknown values because we can use the information that an unknown attribute value is likely to be similar to neighbouring, known values. The subfield of geostatistics has grown up based on this premise. However, if data values do exhibit spatial autocorrelation, this causes problems for statistical techniques that assume data are drawn from independent random samples. Special statistical methods, such as spatial regression models, have been developed to overcome this problem. Equally, it is often hard to defend the assumption of stationarity in spatial processes. That is, it is often assumed that the process generating the observed data is the same everywhere. Spatial non-stationarity exists where the process varies across space. Again, special statistics, such as Geographically Weighted Regression, have been developed to handle this problem.

1.2. TYPES OF SPATIAL ANALYSIS

While there are many different techniques of spatial analysis, they can be categorized into four main types:

1 Those spatial analytical techniques aimed at reducing large data sets to a smaller amount of more meaningful information. Summary statistics, various means of visualizing data and a wide body of data reduction techniques are often needed to make sense of what can be extremely large, multidimensional data sets.

2 Those techniques collectively known as exploratory data analysis which consist of methods to explore data (and also model outputs) in order to suggest hypotheses or to examine the presence of unusual values in the data set. Often, exploratory data analysis involves the visual display of spatial data generally linked to a map.

3 Those techniques that examine the role of randomness in generating observed spatial patterns of data and testing hypotheses about such patterns. These include the vast majority of statistical models used to infer the process or processes generating the data and also to provide quantitative information on the likelihood that our inferences are incorrect.

4 Those techniques that involve the mathematical modelling and prediction of spatial processes.

This book will cover examples of all four types of spatial analysis.

1.3. SPATIAL ANALYSIS IN PERSPECTIVE

It is difficult to say exactly when spatial analysis began in earnest but the beginnings are generally cited in the late 1950s and early 1960s, although much earlier examples of individual pioneering work can be found (e.g, Spottiswoode, 1861). Certainly, the decades of the 1960s and 1970s were periods when quantitative methodologies diffused rapidly within disciplines such as geography and regional science and when much pioneering and fundamental research was carried out. There then followed a period of relative decline for various reasons as outlined by Fotheringham (2006) when many of the newer paradigms in human geography were starkly anti-quantitative. Perhaps also many of the early examples

of spatial analysis were overly concerned with form rather than with process and were rightly criticized for this focus. In addition, it is possible that expectations for quantitative methods may have initially been too high. For example, many believed that spatial modelling, when coupled with adequate data and rapidly increasing computing power, would lead society to 'solve' many of the pressing issues in urban and regional areas.

Significant advances in spatial analysis during the past two decades have brought about a new era of interest in the field. The period of relative decline has now been replaced by one of great enthusiasm for the potential of spatial analysis. This potential has been recognized and embraced by researchers from many fields, ranging from public health and criminal justice, to ecology and environmental studies, as evidenced by various contributions to this volume.

It is now widely recognized in a broad range of disciplines that spatial analysis has an important role to play in making sense of the large volumes of spatial data we now have available and the demand for spatial analysis has never been stronger. It thus is an important time to produce this *Handbook of Spatial Analysis* describing many of the major areas of spatial analysis.

1.4. OVERVIEW OF THE HANDBOOK

The book is designed to capture the state-of-the-art in a broad spectrum of spatial analytical techniques that can be applied to spatial data across a very wide range of disciplinary areas.

Our intent has been to provide a retrospective and prospective view of spatial analysis that covers:

- the reasons why the analysis of *spatial* data needs separate treatment;
- the main areas of spatial analysis;
- the key debates within spatial analysis;
- examples of the application of various spatial analytical techniques;
- problems in spatial analysis; and
- areas for future research.

Although there is inevitable (and desirable) variability in the structure and nature of the individual chapters, in a broad sense the contributions have the following aims:

- describe the current situation within the field, highlighting the main advances that have taken place, as well as current debates;
- describe the problems that still exist, indicating where future research may be best directed;
- indicate key works in the field and provide an extensive bibliography for the area;
- describe the use of the technique in several disciplines; and
- maintain a balance between concepts, theories and methods.

Rapid improvements in the development and availability of high-quality datasets, along with the power and features of geographic information systems that now increasingly provide capabilities for advanced forms of spatial analysis, have propelled the field forward. Consequently, the field of spatial analysis is currently in the midst of an exciting growth period,

where many new tools and methods for analyzing spatial data are being developed, and where applications are being made in an increasing number of fields. This Handbook represents a summary of these developments and applications, as well as a sense of the intense interest that the field now enjoys.

REFERENCES

Spottiswoode, W. (1861). 'On typical mountain ranges: an application of the calculus of probabilities to physical geography'. *Journal of the Royal Geographical Society of London*, **31**: 149–154.

2

The Special Nature of Spatial Data

Robert Haining

This chapter describes some of the special or distinguishing features of spatial data opening the way to methodological issues that will be treated in more depth in later chapters. The use of the term 'special' should not be taken to imply that no other types of data possess these features. Spatial data analysis is a sub-branch of the more general field of quantitative data analysis and has sometimes suffered from not paying sufficient attention to that fact. Many of the data properties that will be encountered are found in other types of (non-spatial) data but when found in spatial data, may possess a particular structure or properties may arise in particular combinations.

The chapter will first define what is meant by spatial data and then identify properties. It will be helpful, in order to put structure on this discussion, to distinguish 'fundamental' properties of spatial data from properties that are due to the chosen representation of geographical space and from properties that are a consequence of measurement processes by which data are collected for the purpose of storage in the spatial data matrix (SDM). The SDM is what the analyst works with. We conclude by considering the implications of these properties for the methodology of spatial data analysis.

Geographic Information Science (GISc) is the generic label that is frequently used, particularly by geographers, to define the area of science that involves the analysis of spatially referenced data – that is data where each case has some form of locational co-ordinate attached to it. Data is the lynch pin in the process of 'doing science' and it is essential that methodologies for spatial data analysis are tuned to the properties of spatial data.

The science undertaken with spatial data is usually 'observational' rather than 'experimental'. This is important. Much spatial data are not collected under controlled situations. We often cannot choose the values of independent variables in order to generate a satisfactory experimental design. There is no replication (in order, for example, to assess the effects of measurement error) and the analyst must take the world as he or she finds it. There may be further problems in specifying what the appropriate locational co-ordinate is when studying certain types of processes and outcomes. All this has implications for the quality of spatial data and for the methodologies that can be employed. We worry not only about the quality of our data but exactly what it is we are observing in any given situation. A consequence of this is that much of the data collected may be used to build a model of the situation under study which can then be used to estimate parameters and test hypotheses. We shall see that some of the fundamental properties of spatial data raise major problems in this regard.

2.1. SPATIAL DATA AND THEIR PROPERTIES

A spatial datum comprises a triple of measurements. One or more *attributes* (X) are measured at a set of *locations* (i) at *time* t, where t may be a point or interval of time. So, if k attributes are measured at n locations at time t, we can present the spatial data in the form:

$$\{x_j(i;t);\ j = 1, \ldots, k;\ i = 1, \ldots, n\}. \quad (2.1)$$

Equation (2.1) expresses in shorthand much of the content of the SDM. The record of when the observation was taken (t) may be suppressed if analysis is concerned with only a single time period but may be retained if there are to be a series of comparative studies through time or if different attributes were recorded at different times and the analyst needs to be aware of this. Such data may come from a variety of different sources including national censuses; public or private agency records (e.g., national health services, police force areas, consumer surveys); satellite imagery; environmental surveys; and primary surveys. The data may be collected from a census or from a sampling process. For the purposes of analysis data from different sources may be required. Studies in environmental epidemiology utilise health, demographic, socio-economic and environmental data. These data may come with differing degrees of quality and may not all be collected on the same areal framework (Brindley *et al.*, 2005).

To understand the properties of spatial data we need to understand the relationship between equation (2.1) and the 'real world' from which the data are taken. In order to undertake data analysis the complexity of the real world must be captured in finite form through the processes of conceptualization and representation (Goodchild, 1989; Guptill and Morrison, 1995; Longley *et al.*, 2001). We shall focus here only on the issues associated with capturing spatial variation, but the reader should note that there are conceptualization and representation issues associated with the way attributes and time are captured as well.

The first step in this process, which ultimately leads to the construction of the SDM, involves conceptualizing the geography of the real world. There are two views of the geographical world in GISc – the field and the object views. The field view conceptualizes space as covered by surfaces with the attribute varying continuously across the space. This is particularly appropriate for many types

of environmental and physical attributes. The object view conceptualizes space as populated by well-defined indivisible objects, a view that is particularly appropriate for many types of social, economic and other types of data that refer to populations. Objects are conceptualized as points, lines or polygons.

These two views constitute models of the real world. In order to reduce a field to a finite number of bits of data then the surface may be represented using a finite number of sample points at which the attribute is recorded or it may be represented using a raster grid. Pixels are laid down independently of the underlying field and its surface variation. Alternatively, the surface may be represented by polygons that partition the space into areas with uniform characteristics (e.g., vegetation zones). How well any field is captured by these different representations will depend on the density of the points or the size of the raster in relation to surface variability. There is a large theoretical and empirical literature on the efficiencies of different spatial sampling designs – for example the properties of random, systematic and stratified random sampling given the nature of variation in the surface to be sampled (see, e.g., Cressie, 1991; Ripley, 1981). The process of discretizing in this way involves a loss of information on surface variability.

This loss of detail on variability also arises when selecting a representation based on the object view. A city may comprise many households (points) but for confidentiality reasons information about households is aggregated into spatially defined groups (polygons) – output areas in the case of the 2001 UK census, enumeration districts prior to 2001 (Martin, 1998). Again aggregation into polygons involves a loss of information. There may be a further loss of information in capturing the polygon itself in the database. It may be captured using a representative point (such as its centroid) and its spatial relationship to other polygons captured using a neighbourhood weights matrix.

The conceptualization of a geographic space as a field or as an object is largely dictated by the attribute. However, representation – the process by which information about the geography of the real world is made finite using geometric constructs – involves making choices (Martin, 1999). These choices include the size and configuration of polygons, the location and density of sample points.

2.1.1. Fundamental properties

Fundamental properties are inherent to the nature of attributes as they are distributed across the earth's surface. There is a fundamental continuity (structure) to attributes in space that derives from the underlying processes that shape the human and physical geographical world. We shall discuss examples of these processes in section 2.2.2. The geographical world would be a strange place if levels of attributes changed suddenly and randomly as we moved from one point in space to another close by. Continuity is also a fundamental property of attributes observed in time. If we know the level of an attribute at one position in space (time) we can make an informed estimate of its level at adjacent locations (points in time). The information that is carried in a piece of data about an attribute at a given location provides information on what the level of the attribute is at nearby locations. However as distance increases then the similarity of attribute values weakens and in the GISc literature this is often referred to as Tobler's First Law of Geography (' . . . near things are more related than distant things'). Although Tobler's First Law is clearly an oversimplification, and in relation to some types of spatial variation just plain wrong, it is nonetheless a useful aphorism.

Testing for spatial autocorrelation was one of the high-profile research agendas in geography during the quantitative revolution. Geographers adapted spatial autocorrelation statistics based on the join-count statistic, the cross product statistic and the squared difference statistic that had been developed for quantifying spatial structure on regular areal frameworks (grids). These statistics were developed to test for statistically significant spatial autocorrelation on irregular areal frameworks (Cliff and Ord, 1973). The null hypothesis (no spatial autocorrelation) was assessed against a non-specific alternative hypothesis (spatial autocorrelation is present). We shall see how this argument was developed in later years with the introduction and use by geographers of models for spatial variation.

In the earth sciences, dealing principally with point data from surfaces, the quantification of structure was based on the use of the empirical semi-variogram which uses a squared difference statistic (Isaaks and Srivastava, 1989). The advantage of the latter route was that it led naturally to model specification and model fitting using theoretical semi-variograms. Of course these quantitative measures and tests of hypothesis depend on the scale of analysis. That is, they depend on the size of the polygons in terms of which data are reported, the inter-point distance between samples on a continuous surface. Thus the chosen representation has an important influence on the quantification of this fundamental property and hence its presence within any spatial dataset. If samples are taken at sufficient distances apart the level of spatial autocorrelation is likely to be much reduced relative to the case where samples are taken close together.

Autocorrelation statistics are also used to capture temporal structure in attribute values but there are important differences with the spatial situation. Time has a natural uni-directional flow (from past to present) whereas space has no such order. The two dimensional nature of space means that dependency structures might vary not just with distance but direction too giving rise to anisotropic dependency structures with structure along the north–south axis differing from the east–west axis. The presence of spatial autocorrelation, that attribute values are not statistically independent, has fundamental implications for the conduct of spatial analysis.

Spatial autocorrelation, in statistical terms, is a second order property of an attribute distributed in geographic space. In addition there may be a mean or first-order component of variation represented by a linear, quadratic, cubic (etc.) trend. We can think of these as two different scales of spatial variation although the distinction may be hard to make and quantify in practice. As Cressie (1991) remarks: 'What is one person's (spatial) covariance may be another persons mean structure' (p. 25). It has often been remarked that spatial variation is heterogeneous. This type of decomposition (plus a white noise element to capture highly localized heterogeneity) is one way of formally capturing that heterogeneity using what are termed 'global' models. Another approach is to only analyze spatial subsets, that is allow model structure to vary locally.

2.1.2. Properties due to the chosen representation

We have already noted that the extent to which our data retains fundamental properties depends on the chosen representation. We now turn to look at other properties that stem directly or indirectly from the chosen representation.

Representing spatial variation using polygons is employed in many branches of science that handle spatially referenced data. Two of the generic consequences of

working with data aggregates are: intra-areal unit heterogeneity and inter-areal unit heteroscedasticity.

Whether the data refer to a continuously varying phenomenon (field view) or aggregations of individuals like households (object view) the effect of bundling data into spatial aggregates has the effect of smoothing variation. In the case of environmental data and the use of pixels then the degree of smoothing will clearly depend on the size of the pixels. The larger the pixels the greater the degree of smoothing. A non-intrinsic partition, where the polygons are defined in terms of attribute variability with the aim of maximizing within unit homogeneity and maximizing between-unit heterogeneity will not produce this effect to the same extent. This second process shares common ground with the process of regionalization – to which it is sometimes compared.

Intra-unit heterogeneity is a particular problem for many types of social science data particularly in those cases where area boundaries are chosen arbitrarily as was the case with the UK census for example prior to 2001. Attributes reported for an area may represent percentages or means of attribute values associated with the individuals (people or households) that have been aggregated and the analyst may have no information on the variability around the mean. If an ecological or contextual attribute is calculated for an area (social capital say, or area deprivation) again the calculation is conditional on the chosen representation and the scale of the partition.

One of the conclusions that might be drawn from this is that it is better to have small areal aggregates rather than large ones. Assuming spatial structure, a reasonable supposition given the discussion in section 2.1.1, then smaller areas should be more homogeneous than larger areas and their mean values should be more representative of their area's population. But such spatial precision comes at the cost of statistical precision. Data errors or small random fluctuations in numbers of events (household burglaries; disease outcomes) will have a big effect on the calculation of rates when populations are small. Take the case of a standardized mortality ratio. If the expected count is small, for example 2.0, then the ratio itself (observed count divided by the expected count) rises or falls by 0.5 with each addition or subtraction of a single case. This will have implications for determining the statistical significance of counts – whether there are significantly more cases than would be expected on the basis of chance alone. It will also have implications for determining the statistical significance of differences in counts between areas which in turn raises problems for the detection of significant crime hotspots or disease clusters.

In summary, there is a trade-off that is linked to the number of individual elements in a polygon. A polygon containing few individuals will tend to be more homogeneous but statistical quantities, such as rates and ratios, tend to be unreliable in the sense that small errors and random fluctuations can impact severely on the calculated values. Polygons containing many individuals will generate robust rates and ratios but often conceal much higher levels of internal heterogeneity.

In practice an area is sometimes partitioned into polygons of varying size and this can yield a secondary effect on data properties. A rate calculated for a polygon where the denominator attribute is small has a larger variance than a rate computed for a polygon where the denominator attribute is large. Moreover there is a mean-variance dependence in the rate statistics. Take the case where the denominator is the number of households ($n(i)$). Rates are observed counts of some attribute (number of burglaries) in polygon i ($O(i)$) divided by the number of households. It follows from the binomial

model for $O(i)$ that:

$$E[O(i)/n(i)] = (1/n(i))\,E[O(i)] = p(i);$$

$$\begin{aligned}\mathrm{Var}[O(i)/n(i)] &= (1/n(i))^2\ \mathrm{Var}[O(i)] \\ &= p(i)(1-p(i))/n(i)\end{aligned} \tag{2.2}$$

where $E[\ \ldots\ ]$ and $\mathrm{Var}[\ \ldots\ ]$ denote mean and variance and $p(i)$ is the probability that any individual in area i (e.g., number of households) has the characteristic (e.g., been burgled) that is being counted. The mean and the variance in equation (2.2) are clearly not independent. It also follows from equation (2.2) that the standard error of the estimate of the rate $p(i)$ which is:

$$[p(i)\,(1-p(i))\,/n(i)]^{1/2}$$

is inversely related to the number of households. It follows that any real spatial variation in rates could be confounded by variation in $n(i)$ (the number of households) or alternatively spatial variation in rates could be an artifact of any spatial structure in $n(i)$ (see Gelman and Price, 1999, who give examples from disease mapping in the USA).

Standardized ratios provide an estimate of the true but unknown area-specific relative risk of the selected disease under the assumption of an independent Poisson model for the observed counts. It follows from the properties of the Poisson distribution that the standard error of the standardized ratio is $O(i)^{1/2}/E(i)$. Using a normal approximation for the sampling distribution of the standardized ratio, $SR(i)$, approximate 95% confidence intervals can be computed:

$$SR(i) \pm 1.96\left[O(i)^{1/2}/E(i)\right].$$

However there are problems here when making comparisons. The standard error tends to be large for areas with small populations and small for areas with large populations because of the effect of population size on $E(i)$. So extreme ratios tend to be associated with small populations but ratios that are significantly different from 1.0 tend to be associated with areas with large populations (Mollie, 1996).

These examples are intended to illustrate the way in which data properties can be induced by the chosen representation. In certain circumstances the geographical structure of the representation (for example the geography of which areas have large and which have small denominator values) could induce a geographical structure on the statistics which when mapped could then give rise to a misleading impression about trends or patterns in the data.

2.1.3. Properties due to measurement processes

The final step in the creation of the SDM involves obtaining measurements on the attributes of interest given the chosen representation.

Data quality can be assessed in terms of four characteristics: accuracy, completeness, consistency and resolution. As noted above, a spatial datum comprises a triple of measurements: the attributes, location and time. Thus the quality of each of these three measurements needs to be assessed against the four characteristics. What is of interest here, however, is how measurement problems might introduce certain properties into the data (Guptill and Morrison, 1995).

A common assumption in error analysis is that attribute errors are independent. This is likely to hold less often in the case of spatial data. Location error may lead to overcounts in one area and undercounts

in adjacent areas because the source of the overcount is the set of nearby areas that have lost cases as a result of the location error. So, count errors in adjacent areas may be negatively correlated (Haining, 2003, pp. 67–70). Location error can be introduced into a spatial data set as a result of having to put data, collected on different spatial frameworks, onto a common spatial framework. Areal interpolation methods are used but these are based on assumptions about how attributes are distributed within areal units and these assumptions often cannot be tested. The consequence is that further levels and patterns of error are introduced into the database (Cockings *et al*., 1997).

In the case of remotely sensed data, the values recorded for any pixel are not in one-to-one relationship with an area of land on the ground because of the effects of light scattering. The form of this error depends on the type and age of the hardware and natural conditions such as sun angle, geographic location and season. The point spread function quantifies how adjacent pixel values record overlapping segments of the ground so that the errors in adjacent pixel values will be positively correlated (Forster, 1980). The form of the error is analogous to a weak spatial filter passed over the surface so that the structure of surface variation, in relation to the size of the pixel unit, will influence the spatial structure of error correlation. Linear error structures also arise in remotely sensed data (Short, 1999). Finally, we note that the effects of error propagation may further complicate error properties when arithmetic or cartographic operations are carried out on the data and source errors are compounded and transformed via these operations (Haining and Arbia, 1993).

Data incompleteness may induce false patterns in spatial data. Data incompleteness refers to the situation where there are missing data points or values or where there are under or overcounts arising from the reporting process. 'Spatially uniform' data incompleteness raises problems for analysis but spatial variation in the level of data incompleteness with, for example, undercounting, more serious in some parts of the study area than others, can seriously affect comparative work and the interpretation of spatial variation. Missing or inaccurately located cases in a point pattern of events may result in failure to detect a local cluster of cases (Kulldorff, 1998).

Incompleteness in cancer data leads to forms of under or overcounting which give rise to spatial variation that is an artifact of how the data were collected. In the case of official crime statistics geographical differences between large counties in England may be due to differences in police investigative and reporting practices. On the intra-urban scale, burglaries in suburban areas will, on the whole, be well reported for insurance purposes, but in some inner-city areas there may be under reporting either because there is no 'incentive' or because of fear of reprisals. The Census provides essential denominator data for computing small area rates. However refusals to cooperate can lead to undercounting and the 1991 Census in the UK was thought to have undercounted the population by as much as 2% because of fears that its data would be used to enforce the new local 'poll tax'. Inner city areas show higher levels of undercounting than suburban areas where populations are easier to track. Finally, since there are 10-year gaps between successive censuses, population in- and out-flows in many areas may be such as to preserve the essential socio-economic and demographic characteristics of the areas. On the other hand some areas of a city, especially inner-city areas, may experience population mobility and redevelopment which result in marked shifts that have implications for the reliability of the data in the years following the Census.

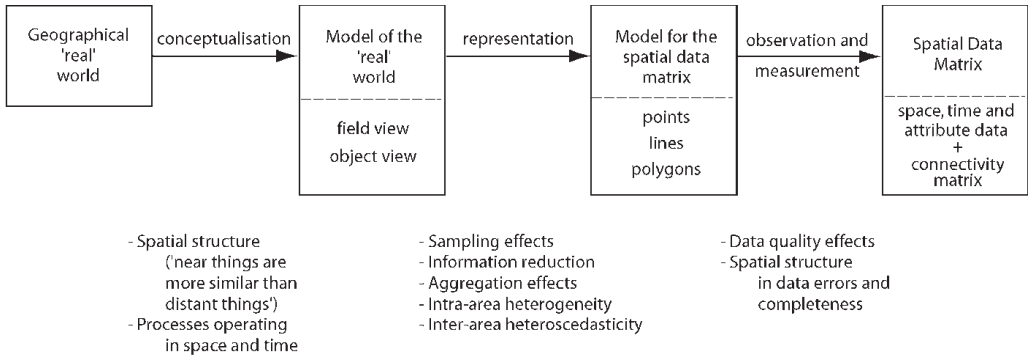


Figure 2.1 Processes involved in constructing the spatial data matrix and the data properties that are present or introduced at each stage.

Finally, in the case of some imagery, some areas of the image may be obscured because of cloud cover. A distinction should be drawn between data that are 'missing at random' from data that are missing because of some reason linked to the nature of the population or the area. Weather stations temporarily out of action because of equipment failure produce data missing at random. On the other hand, mountainous areas will tend to suffer from cloud cover more than adjacent plains and there will be systematic differences in land use between such areas. This distinction has implications for how successfully missing values can be estimated and whether the results of data analysis will be biased because some component of spatial variation is unobservable.

Figure 2.1 provides a summary of the points raised in this section.

2.2. IMPLICATIONS OF DATA PROPERTIES FOR THE ANALYSIS OF SPATIAL DATA

In this section we turn to a consideration of the implications of the properties of spatial data for the conduct of spatial analysis. Again we shall simply introduce ideas which will be taken up in more detail in later chapters. We divide this section into situations where spatial properties can be exploited to help solve problems and situations where spatial properties introduce complications for the conduct of data analysis.

2.2.1. Taking advantage of spatial data properties to tackle problems

Consider the following problems:

- Samples of attribute values have been taken across an area. The analyst would like to construct a map to describe surface variation using the information contained in the sample. Perhaps instead the analyst just wishes to estimate the surface at a point, or set of points, where no sample has been taken and estimate the prediction error.

- A spatial database has been assembled but the database contains data that are 'missing at random' in the sense that there are no underlying reasons (such as suppression or confidentiality) why the particular values are missing. The analyst wants to estimate these missing values.

In both these cases we might expect to exploit some formalized version of the notion that data points near together in space carry

information about each other. Both of these examples constitute a form of the spatial interpolation problem and solutions such as kriging exploit the spatial structure inherent in the surface as well as the configuration of the sample points to provide an estimate of surface values together with an estimate of the prediction error (Isaak and Srivastava, 1989). It is intuitive that any solution that did not use the information contained in the location co-ordinates of sample data values would be considered an inefficient solution.

Consider another group of problems:

- Aggregated data are obtained on race (black/white) and voting behaviour (did vote/did not vote). Counts in the 2×2 table are known but the real interest lies in the voting behaviour at the constituency level.

- Unemployment estimates have been obtained from a survey for each of a number of small areas in a region. The small area estimators are unbiased but, because of small sample sizes have low precision. Conversely the region wide estimator has high precision, but as an estimate for any of the small area levels of unemployment is biased. A similar situation arises when estimating relative risk levels across the small areas of a larger region using the standardized mortality ratio.

In both these cases there is again an opportunity to exploit some formalized version of the notion that data points nearby in space carry information about each other. One solution is to 'borrow information' or 'borrow strength' so that the low precision of small area estimates are raised by using data from nearby areas (Mollie, 1996; King, 1997). These nearby areas provide additional data (helping to improve precision) and because they are nearby should reflect an underlying situation that is close to the small area in question so will not introduce a serious level of bias.

2.2.2. Where spatial data properties introduce complications for data analysis

Spatial analysis is often called upon to address scientific *questions* relating to outcomes (numbers of cases of a disease, distribution of house prices, regional economic growth rates) that are a consequence of processes that by their nature are spatial. Haining (2003) identifies four generic groups of spatial processes. A *diffusion* process is one where some attribute is taken up by a population so that at any point in time some individuals have the attribute (e.g., an infectious disease) and some do not. If the diffusion process operates in ways that are constrained by distance then there is likely to be spatial structure in the geography of those who do and those who do not have the attribute in question. An *exchange and transfer* or *mixing* process is one where places become similar in attribute values (per capita income; employment) as a result of flows of goods or services that bind their economic fortunes together or where patterns of movement and mixing perhaps at different scales introduce a measure of spatial homogeneity into structures. A third type of spatial process is an *interaction* process in which outcomes at one location (e.g., the price of a commodity) are observed and as a result of the competition effect influence outcomes (prices) at another location. Finally, there is a *dispersal* process in which individuals spread across space (such as the dispersal of seeds around a parent plant) so that counts reflect the geography of the dispersal mechanism.

These generic spatial processes – processes that operate in geographic space – generate data where spatial structure emerges as a fundamental property of the data. Process shapes or at least influences attribute variation and the resulting data that are collected possess

dependency structures that reflect the way the process plays out across geographic space.

Not all processes of interest are 'spatial' in the sense described above. Many of the processes of interest to geographers play out across geographic space in response to the place-based characteristics of areas (the particular mix of attributes they possess) and the spatial relationships between those areas. Outcomes in places (whether for example economic, social, epidemiological or criminological) are not necessarily merely the consequence of the properties of those places – as places – but may also be the consequence of relational and contextual influences. The distance between places; the difference between adjacent places in terms of relevant attributes; the overall configuration of places across a region, are all facets of relation and context that may impact on outcomes and modify the role of 'place' in influencing outcomes. Two places may be identical in terms of their place-based characteristics but differ significantly in terms of their relational and contextual attributes with neighbouring areas and these differences may explain why (for example) two similarly affluent neighbourhoods experience quite different levels of assault and robbery; why two similarly deprived neighbourhoods experience quite different levels of health outcomes.

We now examine briefly how these features of how attribute values are generated impact on the choice of methodology for the purpose of data analysis. We distinguish between exploratory spatial data analysis and model-based forms of analysis that allow hypothesis testing and parameter estimation.

Exploratory spatial data analysis

Exploratory data analysis (EDA) comprises a collection of visual and numerically resistant techniques for summarizing data properties, detecting patterns in data, identifying unusual or interesting features in data including possible data errors and formulating hypotheses. Exploratory spatial data analysis (ESDA) undertakes these activities with respect to spatial data so that cases can be located on a map and the spatial relationships between cases assumes importance because they carry information that is likely to be relevant to the analysis (Cressie, 1984; Haining *et al.*, 1998; Fotheringham and Charlton, 1994). It is important to be able to answer questions such as: 'where does that subset of cases on the scatterplot or that subset of cases on the boxplot, occur on the map?' 'What are the spatial patterns and spatial associations in this geographically defined subset of the map?' In the case of regression modelling do the large positive residuals, for example, cluster in one area of the map?

ESDA and the software that supports ESDA needs to be able to handle the spatial index and be able to handle the special queries that arise because of the spatial referencing of the data. Thus the map becomes an essential visualization tool (Dorling, 1992). The linkage between a map window and other graphics windows, so that cases can be simultaneously highlighted in more than one window, becomes an essential part of the conduct of ESDA (Andrienko and Andrienko, 1999; Monmonier, 1989).

Visualizing spatial data raises particular problems, in part because of some of the properties discussed in earlier parts of this chapter. We highlight two here. First, it has been noted that data values, particularly rates and ratios, may not be strictly comparable because standard errors are population size dependent. So if areas vary substantially in terms of population counts (used as the denominators for a rate) then extreme values and even patterns detected by visual inspection might be associated with that effect rather than real differences between areas. Second, areas that partition a region might be very different in physical size.

This may mean that the viewer of a map has their attention drawn to certain areas of the cartographic display (those areas with physically large spatial units) whilst other areas are ignored. This may be particularly important if in fact it is the small areas that have the larger populations so that it is their rates and ratios (rather than the rates and ratios associated with the physically larger but less densely populated areas) that are the more robust. One solution to this problem is to use cartograms so that areas are transformed in physical extent to reflect some underlying attribute such as population size (Dorling, 1994). This comes at a cost because the individual areas in the resulting cartogram may be hard for the analyst to place. There may be a need for a second, conventional, map linked to the cartogram, so the analyst can highlight areas on the cartogram and see where they are on the conventional map.

Conventional visualization technology is often based on the assumption that all data values are of equal status so that the viewer can extract information from visual displays without worrying about the statistical comparability of the data values that are displayed. This assumption may break down when dealing with spatially aggregated data (Haining, 2003).

Model fitting and hypothesis testing

If n data values are spatially autocorrelated then one of the consequences of this for the application of standard statistical inference procedures is that the information content of the data set is less than would be the case if the n values were independent. This means that the degrees of freedom available for testing hypotheses is not a simple function of n. We shall take the example of testing for significant bivariate correlation between two variables to illustrate this point.

Suppose n pairs of observations, $\{(x(i), y(i))\}_i$ are drawn from a bivariate normal distribution. Pearson's product moment correlation coefficient (r) is the statistic used to measure the association between X and Y. If the observations on the two variables are independent (there is no spatial autocorrelation in either X or Y), then if the null hypothesis is of no association between X and Y then a test statistic is given by:

$$(n-2)^{1/2}\,|r|\,\left(1-r^2\right)^{-1/2} \tag{2.3}$$

which is t distributed with $(n-2)$ degrees of freedom.

These distributional results do not hold if X and Y are spatially correlated. The problem is that when spatial autocorrelation is present the variance of the sampling distribution of r, which is a function of the number of pairs of observations n, is underestimated by the conventional formula which treats the pairs of observations as if they were independent. The effect of spatial autocorrelation on tests of significance have been extensively studied (for reviews see Haining, 1990, 2003) and shown to be very severe when both X and Y have high levels of spatial autocorrelation.

Clifford and Richardson (1985) obtain an adjusted value for n (n') which they call the 'effective sample size'. This value, n', can be interpreted as measuring the equivalent number of independent observations so that the solution to the problem lies in choosing the conventional null distribution based on n' rather than n. An approximate expression for this quantity is:

$$n' = 1 + n^2\left(\text{trace}\left(\mathbf{R}_x\mathbf{R}_y\right)\right)^{-1} \tag{2.4}$$

where $\mathbf{R}_x$ and $\mathbf{R}_y$ are the estimated spatial correlation matrices for X and Y respectively. (For a discussion of estimators see Haining,

1990, pp.118–120.) The null hypothesis of no association between X and Y is rejected if:

$$\left(n' - 2\right)^{1/2} |r| \left(1 - r^2\right)^{-1/2} \qquad (2.5)$$

exceeds the critical value of the t distribution with $(n' - 2)$ degrees of freedom.

This illustrates a general problem. Since the n observations are positively spatially autocorrelated, the information content of the sample is over-estimated if n is used – it needs to be deflated. The sampling variance of statistics are underestimated leading the analyst to reject the null hypothesis when no such conclusion is warranted at the chosen significance level. For the effects of spatial dependency on the analysis of contingency tables see, for example, Upton and Fingleton (1989) and Cerioli (1997).

To make further progress in understanding the importance of spatial data properties and the complications they introduce we need to introduce models for spatial variation – or data generators for spatial variation. Such models are important. By specifying a model to represent the variation in the data (including the spatial variation), the analyst is able to construct tests of hypothesis with greater statistical power than is possible if testing is against a non-specific alternative. There are a number of possible formal models for spatial variation of which the simultaneous spatial autoregressive (SAR), the conditional spatial autoregressive (CAR) and the moving average (MA) models are probably the best known. We will briefly look at the first two but the interested reader will need to follow up the literature to gain a fuller understanding of these models and their properties (Whittle, 1954; Besag, 1974, 1975, 1978; Ripley, 1981; Cressie, 1991; Haining, 1978, 1990, 2003).

A multivariate normal CAR model which satisfies the first order (spatial) Markov property and thus might be thought of as the simplest departure from spatial independence can be written as follows (Besag, 1974; Cressie, 1991, p. 407):

$$E\left[X(i) = x(i) \mid \left\{X(j) = x(j)\right\}_{j \in N(i)}\right] = \mu + \sum_{j \in N(i)} \tau\, w(i, j)\,[X(j) - \mu]\,, \quad i = 1, \ldots, n \qquad (2.6)$$

and:

$$\mathrm{Var}\left[X(i) = x(i) \mid \left\{X(j) = x(j)\right\}_{j \in N(i)}\right] = \sigma^2, \quad i = 1, \ldots, n$$

where $E[\ \ldots\ |.]$ and $\mathrm{Var}[\ \ldots\ |.]$ denote conditional expectation and variance respectively, μ is a first-order parameter and τ is the spatial interaction parameter. The Markov property means observations are *conditionally* independent given the values at neighbouring sites. $\{w(i, j)\}$ denotes the neighbourhood structure of the system of areas and $w(i, j) = 1$ if i and j are neighbours ($j \in N(i)$) and $w(i, i) = 0$ for all i. $\mathbf{W}$ is the $n \times n$ matrix of $\{w(i, j)\}$ and is sometimes called the connectivity matrix. It is a requirement that τ lies between $(1/\omega_{\min})$ and $(1/\omega_{\max})$ where $\omega_{\min}$ and $\omega_{\max}$ are the smallest and largest eigenvalues of $\mathbf{W}$. For a fuller introduction to the Markov property for spatial data including how to construct higher-order spatial Markov models see, for example, Haining (2003, pp. 297–299). This approach allows the construction of a hierarchy of models of increasing complexity. As noted in Haining (2003), however, the Markov property does not have the natural appeal it has in the case of time series, because space has no natural ordering. So the neighbourhood structure can

often seem rather arbitrary especially in the case of the non-regular areal frameworks used to report Census and other social and economic data.

If the analyst of regional data does not attach importance to satisfying a Markov property another option is available called the SAR model specification. A form of this model was first introduced into statistics by Whittle (1954). Let **e** be independent normal IN($\mathbf{0}$, $\boldsymbol{\sigma}^2\mathbf{I}$) where **I** is the identity matrix and $e(i)$ is the variable associated with site i ($i = 1, \ldots, n$). Define the expression:

$$X(i) = \mu + \sum_{j \in N(i)} \rho w(i, j)[X(j) - \mu] + e(i), \quad i = 1, \ldots, n. \tag{2.7}$$

where ρ is a parameter. The bounds on ρ are set by the largest and smallest eigenvalues of **W** just as in the case of the CAR model. This is the model most often seen in the spatial analysis and regional science literature although the reason for its hegemony is far from clear and seems to be largely based on a combination of historical accident (in the sense that time series modelling preceded spatial data modelling and methods were transferred across) and subsequent 'lock-in'.

These models can be embedded into, for example, regression models either as additional covariates (as in the case of equation (2.7)) or as models for the error structure where the errors (in practice the residuals) are tested and found to show evidence of spatial autocorrelation (Anselin, 1988; Ord, 1975). It is well known that fitting regression models by ordinary least squares when errors are spatially (positively) autocorrelated gives rise to some damaging consequences. First, although we shall obtain consistent estimates of the regression parameters (there may be some small sample bias), the sampling variance of these estimates may be inflated compared with methods that take account of the spatial autocorrelation in the errors. Second, if the usual least squares formula for the sampling variances of these regression estimates is applied, the variances will be seriously underestimated. The formulae are no longer valid and conventional F and t tests of hypothesis are also not valid. We shall take a very simple example to illustrate these points, where the parameter to be estimated and tests of hypothesis relate to a constant mean μ.

Suppose n independent observations $\{x(i)\}$ are drawn from a $N(\mu, \sigma^2)$ distribution. The sample mean, $\bar{x}$, is an unbiased estimator for μ, and the variance of the sample mean is:

$$\text{Var}(\bar{x}) = \sigma^2/n. \tag{2.8}$$

If σ^2 is unknown then it is estimated by:

$$s^2 = (1/(n-1)) \sum_{i=1, \ldots, n} (x(i) - \bar{x})^2 \tag{2.9}$$

so that:

$$Var(\bar{x}) = (1/n(n-1)) \sum_{i=1, \ldots, n} (x(i) - \bar{x})^2 . \tag{2.10}$$

If the n observations are not independent then although the sample mean is still unbiased as an estimator of μ, assuming each $x(i)$ has the same variance (σ^2), the variance of the sample mean is (see, for example, Haining, 1988, p. 575):

$$Var(\bar{x}) = \sigma^2/n + \left(2/n^2\right) \times \sum_i \sum_{j(i<j)} \text{Cov}(x(i), x(j)) \tag{2.11}$$

where $\mathrm{Cov}(x(i), x(j))$ denotes the spatial autocovariance between $x(i)$ and $x(j)$. So, if there is positive spatial dependence and σ^2 is known then σ^2/n underestimates the true sampling variance of the sample mean. If σ^2 is unknown and is estimated by equation (2.9) then if there is positive spatial dependence the expected value of s^2 is (see, for example, Haining, 1988, p. 579):

$$E\left[s^2\right] = \sigma^2 - [(2/n(n-1)) \times \sum_i \sum_{j(i<j)} \mathrm{Cov}\,(x(i), x(j))] \tag{2.12}$$

so that equation (2.9) is a downward biased estimate of σ^2. This further compounds the underestimation of the sampling variance.

Modified methods to take account of spatial dependence are often based on the following argument (see, for example, Haining, 1988). Assume the data $\mathbf{x}^T = (x(1), \ldots, x(n))$, where T denotes the transpose, are drawn from a multivariate normal spatial model with mean vector given by $\mu\mathbf{1}$ and n by n variance–covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{V}$ given, say, by one of the models described above. (In the case of the CAR model (2.6), $\mathbf{V} = (\mathbf{I} - \tau\mathbf{W})^{-1}$.) The log likelihood for the data is:

$$-(n/2)\ln 2\pi\sigma^2 - (1/2)\ln|\mathbf{V}| - \left(1/2\sigma^2\right) \times (\mathbf{x} - \mu\mathbf{1})^T\ \mathbf{V}^{-1}(\mathbf{x} - \mu\mathbf{1}) \tag{2.13}$$

where $\mathbf{1}$ is a column vector of 1's and $|\mathbf{V}|$ denotes the determinant of $\mathbf{V}$. For simplicity we assume $\mathbf{V}$ is known. The maximum likelihood estimator of μ is:

$$\widetilde{\mu} = \left(\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}\right)^{-1}\left(\mathbf{1}^T\mathbf{V}^{-1}\mathbf{x}\right). \tag{2.14}$$

The estimator (2.14) is the best linear unbiased estimator (BLUE) of μ. Note that in the case of independence $\mathbf{V} = \mathbf{I}$ (the identity matrix with 1's down the diagonal and zeros elsewhere) and equation (2.14) reduces to the sample mean. In the case $\mathbf{V} \neq \mathbf{I}$ two modifications to the sample mean are occurring. First, the denominator for positive spatial dependence will be less than n. Second, the presence of $\mathbf{V}^{-1}$ in the numerator of equation (2.14) downweights the contribution of any attribute $x(i)$ which is highly correlated with other attribute values $\{x(j)\}$ – that is, where $x(i)$ is part of a cluster of observations.

The variance of $\widetilde{\mu}$ is:

$$\mathrm{Var}[\widetilde{\mu}] = \sigma^2(\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1})^{-1} \tag{2.15}$$

which reduces to σ^2/n if $\mathbf{V} = \mathbf{I}$.

Since the sample mean is an unbiased estimator of μ, one modification is to replace equation (2.8) with equation (2.15). The term $(\mathbf{1}^T\ \mathbf{V}^{-1}\ \mathbf{1})$ is proportional to Fisher's information measure (Haining, 1988, p. 586). It identifies the information about μ contained in an observation. Now equation (2.9) is not the maximum likelihood estimator for σ^2. This is given by:

$$\widetilde{\sigma}^2 = n^{-1}(\mathbf{x} - \mu\mathbf{1})^T\ \mathbf{V}^{-1}(\mathbf{x} - \mu\mathbf{1}). \tag{2.16}$$

A further refinement is to replace equation (2.9) with equation (2.16) substituting the sample mean for μ in equation (2.16) where $\mathbf{V}^{-1}$ plays a role equivalent to the second term in the right-hand side of equation (2.11).

The general results given by equations (2.11) and (2.12) are why adjustments to conventional methods are needed. The evidence suggests that it is the effect of the second term on the right-hand side of

equation (2.11) that is the more serious, at least in the usual situation of positive spatial dependence, and that one way to deal with this is to adjust n in equation (2.8) thereby increasing the sampling variance of the sample mean. The size of the adjustment to n will be sensitive to the estimates of the spatial autocorrelation in the data or, if a spatial model is fitted to the data, the choice of model. The problem is further complicated if, as is usually the case, $\mathbf{V}$ is not known and so must be estimated from the data.

Before leaving the normal model it is important to note that aggregated spatial data may violate another of the statistical assumptions of least squares regression. It was remarked in section 2.1 how rates and ratios based on areas with very different population counts will have different standard errors. It follows that the assumption of homoscedasticity (or constant error variance) is likely to be violated when developing models to explain how rates or ratios vary over a region. Data transformations or weighted least squares estimators are used to address these problems (Haining, 1990, pp. 49–50) but such adjustments may need to be implemented whilst also addressing the problems created by residual spatial autocorrelation (Haining, 1991). In addition to the problems created by failure to satisfy statistical assumptions, spatial data often create 'data-related' problems in regression modelling (Haining, 1990, pp. 332–333). For example, the fit of a trend surface model can be influenced by the configuration of the sample data points on the surface where, as a result of the particular distribution, certain values have high leverage (Unwin and Wrigley, 1987); the particular shape of the study region may also influence the trend surface model fit (Haining, 1990, p. 372). These and other issues are reviewed in Haining (1990, pp. 40–50).

We conclude this section by remarking on the implications of intra-area and inter-area spatial dependency and intra-area heterogeneity when modelling a discrete valued response variable such as the count of the number of cases of a disease across a region using the Poisson model. Spatial dependency and heterogeneity are important causes of overdispersion. For example consider a local diffusion process in which individuals are more likely to be infected if they are close to someone already infected. The result is that counts of the number of cases will reveal Poisson overdispersion because there will be areas with large counts (due to the local infection process) and areas with zero counts where the process has not yet started. These considerations require the analyst both to carry out tests for overdispersion and where necessary take appropriate action. The effects of overdispersion in generalized linear modelling are rather similar to those described for the normal model when spatial autocorrelation is detected. If overdispersion is present, ignoring it tends to have little impact on point estimates of the regression parameters (the maximum likelihood estimator is consistent, although some small sample bias might be present). However, standard error estimates for regression parameters are underestimated. Type I errors associated with the model are underestimated which is particularly problematic in relation to predictors that are close to the significance threshold. If the objective is to build a parsimonious model, the presence of overdispersion may result in an analyst constructing a model more complicated than necessary, and that overestimates the variance explained.

Ways of tackling this problem may depend on the reasons for the overdispersion. A conventional approach is through the use of a variance inflation factor (Dobson, 1999). Where the cause is inter-area spatial autocorrelation then a discrete valued 'auto-model' may be used which is analogous to equation (2.6) (see Besag, 1974). More recently attention has focused on the use

of spatial random effects models using CAR models fitted using WinBUGS (Law *et al*., 2006). These models allow for overdispersion through the random effects term. This is an area of current research in spatial modelling since the development of good modelling tools for discrete valued response variables has rather taken a back seat whilst attention for many years has focused – perhaps disproportionately – on the normal model (Law and Haining, 2004).

2.3. DRAWING INFERENCES

One of the main purposes of undertaking spatial statistical analysis is to make population inferences on the basis of the data collected. In concluding this chapter we consider some of the inference pitfalls associated with the analysis of spatial data.

What is the population about which inferences are made in an observational science? If data are point samples from a continuous surface then the population might be the surface itself. Of course the realized surface may be thought of as only one of many possible realizations (the rest not having been observed). However, with or without the concept of a 'superpopulation' of surfaces, making inferences from point samples to the (realized) surface population does represent a legitimate target. This argument is less convincing when the data represent a complete census – for example the data refer to areas and a complete (or nearly complete) enumeration has been carried out. What is the population about which inferences are being made now? A frequent answer to this is that the underlying process is stochastic (chance is an inherent part of the process) so that inferences are directed at the process (its parameters and covariates) rather than the map. The problem with this is that we have access to only one realization of the process and in order to give our inferences some broader validity other assumptions need to be invoked such as that this realization is representative of the underlying process. There may be no way to test such an assumption.

The modifiable areal units problem (MAUP) reminds us that results obtained from analyzing aggregate data are dependent on the particular scale of the partition, and, at the given scale, the particular boundaries used. In general, statistical relationships between attributes are stronger the larger the spatial aggregates because variances are reduced. Boundary shifts can influence whether or not disease clusters or crime hot spots are detected at any scale because if boundaries happen to cut through the middle of a cluster this may dilute the effect over two or more areas.

The analysis of aggregated data is particularly problematic and not just because of the MAUP. It is important to remember that conclusions drawn from aggregate data can only be transferred to the individual level under certain conditions. The ecological fallacy is the uncritical transfer of findings at the group level to the individual level. As the famous example cites, the suicide rate in Germany in the 17th century may have been larger in areas with higher percentages of Catholics but that does not mean Catholics were more prone to commit suicide than Protestants. Quite the reverse as individual level data revealed. Aggregation bias raises serious problems for epidemiological studies based on aggregate data and is one reason why it is considered the weakest of the different methodologies for assessing dose–response relationships – even though this may be the only realistic way of obtaining reasonably sound measures of exposure to an environmental risk factor. The problem is that it is not difficult to construct examples where there are complete sign reversals when going

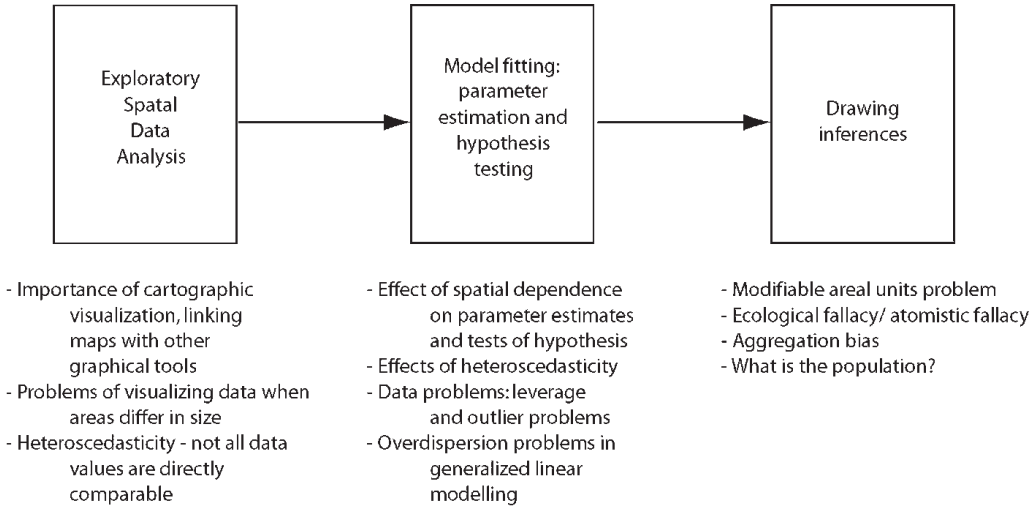


Figure 2.2 Spatial data properties and how they impact at different stages of analysis.

from the ecological to the individual level study (Richardson, 1992).

The converse of the ecological fallacy is the atomistic (or individualistic) fallacy which assumes relationships identified at the individual level apply at the group level. There may be group level or contextual effects that need to be taken into account – as for example in the study of youth offending, where the risk of becoming an offender may not depend only on personal and household level risk factors but also neighbourhood and peer group effects. This then raises the problem of defining what the 'neighbourhood' is.

Figure 2.2 provides a summary of the points raised in sections 2.2 and 2.3.

2.4. CONCLUSIONS

Spatial data possess a number of distinctive properties that derive from the fundamental nature of geographic space and the way processes unfold in geographic space, the way that spatial variation is represented for the purpose of storage in a finite digital database and the way spatial data are collected and attributes measured. Many of these properties were recognized early in geography's 'quantitative revolution' most notably the lack of independence in data values collected close together in space. Geographers then and since have made important contributions to the development of relevant statistical theory and practice.

Geographers continue to develop new methods for describing spatial variation and new methods for modelling processes that operate across geographical space. At present there are two strong traditions which provide focuses for research. On the one hand there are methodologies based on 'whole map' or global statistics that seek to capture data properties through models that are fitted to all the data. On the other hand there are methodologies based on 'local' statistics that process geographically defined subsets of the data and do not seek to impose a single statistic or model on the whole data set (Anselin, 1995, 1996; Getis and Ord, 1996; Fotheringham and Brunsdon, 2000). They represent different ways of responding to the need to develop methodologies to meet the

analytical challenges posed by the special nature of spatial data.

REFERENCES

Andrienko, G.L. and Andrienko, N.V. (1999). Interactive maps for visual data exploration. *International Journal of Geographical Information Science,* **13**: 355–374.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Dordrecht: Kluwer Academic.

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis*, **27**: 93–115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: Fischer, M., Scholten, H.J. and Unwin, D., (eds), *Spatial Analytical Perspectives on GIS,* pp. 111–125. London: Taylor & Francis.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society,* B, **36**: 192–225.

Besag, J.E. (1975). Statistical analysis of non-lattice data. *The Statistician,* **24**: 179–195.

Besag, J.E. (1978). Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute,* **47**: 77–92.

Brindley, P., Wise, S.M., Maheswaran, R. and Haining, R.P. (2005) The effect of alternative representations of population location on the areal interpolation of air pollution exposure. *Computers, Environment and Urban Systems,* **29**: 455–469.

Cerioli, A. (1997). Modified tests of independence in 2×2 tables with spatial data. *Biometrics,* **53**: 619–628.

Cliff, A.D. and Ord, J.K. (1973). *Spatial Autocorrelation.* London: Pion.

Clifford, P. and Richardson, S. (1985). Testing the association between two spatial processes. *Statistics and Decisions*, **Suppl. No. 2**: 155–160.

Cockings, S., Fisher, P.F. and Langford, M. (1997). Parametrization and visualization of the errors in areal interpolation. *Geographical Analysis,* **29**: 314–328.

Cressie, N. (1984). Towards resistant geostatistics. In: Verly, G., David, M., Journel, A.G. and Marechal, A., (eds), *Geostatistics for Natural Resources Characterization,* pp. 21–44. Dordrecht: Reidel.

Cressie, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

Dobson, A.J. (1999). *An Introduction to Generalized Linear Models.* Boca Raton: Chapman & Hall.

Dorling, D. (1992). Stretching space and splicing time: from cartographic animation to interactive visualization. *Cartography and Geographic Information Systems,* **19**: 215–227.

Dorling, D. (1994). Cartograms for visualizing human geography. Hearnshaw, H.M. and Unwin, D.J., (eds), *Visualization in Geographic Information Systems,* pp. 85–102. New York: J. Wiley & Sons.

Fisher, R. (1935). *The Design of Experiments.* Edinburgh: Oliver & Boyd.

Forster, B.C. (1980). Urban residential ground cover using LANDSAT digital data. *Photogrammetric Engineering and Remote Sensing,* **46**: 547–558.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis.* London: SAGE.

Fotheringham. A.S. and Charlton, M. (1994). GIS and exploratory spatial data analysis: an overview of some research issues. *Geographical Systems,* **1**: 315–327.

Gelman, A. and Price, P.N. (1999). All maps of parameter estimates are misleading. *Statistics in Medicine,* **18**: 3221–3234.

Getis, A. and Ord, J.K. (1996). Local spatial statistics: an overview. In: Longley, P. and Batty, M., (eds), *Spatial Analysis: Modelling in a GIS environment,* pp. 261–277. Cambridge: Geoinformation International.

Goodchild, M.F. (1989). Modelling error in objects and fields. In: Goodchild, M. and Gopal, S., (eds), *Accuracy of Spatial Databases,* pp. 107–113. London: Taylor & Francis.

Guptill, S.C. and Morrison, J.L. (1995). *Elements of Spatial Data Quality*. Oxford: Elsevier Science.

Haining, R.P. (1978). The moving average model for spatial interaction. *Transactions of the Institute for British Geographers*, **NS3**: 202–225.

Haining, R.P. (1988). Estimating spatial means with an application to remotely sensed data. *Communications in Statistics, Theory and Methods,* **17**: 573–597.

Haining, R.P. (1990). *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge: Cambridge University Press.

Haining, R.P. (1991). Estimation with heteroscedastic and correlated errors: a spatial analysis of intra-urban mortality data. *Papers in Regional Science,* **70**: 223–241.

Haining, R.P. (2003) *Spatial Data Analysis: Theory and Practice.* Cambridge: Cambridge University Press.

Haining, R.P. and Arbia, G. (1993). Error propagation through map operations. *Technometrics,* **35**: 293–305.

Haining, R.P., Wise, S.M. and Ma, J. (1998). Exploratory Spatial Data Analysis in a geographic information system environment. *The Statistician,* **47**: 457–469.

Isaaks, E.H. and Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics.* Oxford: Oxford University Press.

King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton, New Jersey: Princeton University Press.

Kulldorff, M. (1998) Statistical methods for spatial epidemiology: tests for randomness. In: Gatrell, A. and Löytönen, M., (eds) *GIS and Health*, pp. 49–62. London: Taylor & Francis.

Law, J. and Haining, R.P. (2004) A Bayesian approach to modelling binary data: the case of high intensity crime areas. *Geographical Analysis,* **36**: 197–216.

Law, J., Haining R., Maheswaran, R. and Pearson, T. (2006) Analysing the relationship between smoking and coronary heart disease at the small area level. *Geographical Analysis*, **38**: 140–159.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2001). *Geographical Information Systems and Science.* Chichester: Wiley.

Martin, D.J. (1998) Optimizing Census Geography: the separation of collection and output geographies. *International Journal of Geographical Information Science,* **12**: 673–685.

Martin, D.J. (1999). Spatial representation: the social scientists' perspective. In: Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W., (eds), *Geographical Information Systems: Volume 1. Principles and Technical Issues, 2nd edition.* pp. 71–89. New York: Wiley.

Mollie, A. (1996). Bayesian mapping of disease. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics,* pp. 359–379. London: Chapman & Hall.

Monmonier, M.S. (1989). Geographic brushing: exhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis,* **21**: 81–84.

Richardson, S. (1992). Statistical methods for geographical correlation studies. In: Elliot, P., Cuzich, J., English, D. and Stern, R., (eds), *Geographical and Environmental Epidemiology: Methods for Small Area Studies,* pp. 181–204. Oxford: Oxford University Press.

Ripley, B.D. (1981). *Spatial Statistics*. New York: Wiley.

Unwin, D.J. and Wrigley, N. (1987). Towards a general theory of control point distribution effects in trend surface models. *Computers and Geosciences,* **13**: 351–355.

Whittle, P. (1954) On stationary processes in the plane. Biometrika, **41**: 434–449.

3

The Role of GIS

David Martin

3.1. INTRODUCTION

The role of geographical information systems (GIS) in spatial analysis has for the most part been indirect, and less obvious than might at first be expected. It is probably true to say that throughout the history of GIS, researchers concerned with specific sub-fields of spatial analysis have bemoaned the fact that proprietary GIS software has been an inadequate tool for their work. Certainly Goodchild *et al*. (1992) were able to identify an extensive research and development agenda for the incorporation of spatial analytical tools within GIS, yet more recent reviews such as those by Longley and Batty (1996a, 2003a) have been equally able to identify the discrepancies between the requirements of the spatial analyst and the functionality provided by mainstream GIS software. This is not to say that there have not been many steps taken in the development of spatial analysis tools. Ungerer and Goodchild (2002) note that although progress has been made towards some level of integration between spatial analytical tools and GIS, few analytical functions are actually available as commands from within GIS. Goodchild (2000) fears that despite the many interconnections, the gap between GIS and spatial analysis may actually be widening. It is suggested that in the early years of development, GIS practitioners were more likely to possess some measure of technical expertise and be interested in spatial analytical methods, although the available tools were limited. The spatial analytical functionality of GIS has increased over time, but this has been overshadowed by the massive increase in the number and range of GIS implementations, such that the 'average' GIS user is now in command of a more powerful analytical toolkit, but has little increased ability to make use of it. In other words, the most typical GIS use has moved from a more analytical role to a more operational one, alongside a huge growth in the number of

software systems and users which comprise the GIS community. Nevertheless, GIS has contributed to the development of spatial analytical methods more indirectly through a huge growth in the data resources, structures and basic tools available. It is worth noting here that sometimes in the relevant literature it is not entirely clear whether authors are referring to GIS in the narrower sense of geographical information systems or the broader field of geographical information science (Goodchild, 1992). GIScience incorporates both GISystems and spatial analysis, and the discussion in this chapter focuses on the relationship between these two components.

The remainder of this chapter seeks to explore the complex and much-contested relationship between GIS and spatial analysis. Section 3.2 considers the definitions of each and reviews the extent to which they have become integrated. We then turn, in sections 3.3 and 3.4, to examine some different models whereby spatial analysis and GIS software tools have been connected and consider a selection of more detailed examples. The principal barriers and opportunities for closer integration between GIS and spatial analysis are presented in section 3.5 and the chapter concludes by attempting to assess the likely convergence or divergence between these families of spatial processing techniques in future. By its nature, this chapter inevitably touches on many areas that are discussed in more detail elsewhere in this volume, but the focus here is to explore the interaction between GIS and spatial analysis, and more specifically the contributory role of GIS.

3.2. GIS AND SPATIAL ANALYSIS: MADE FOR EACH OTHER?

There are very many GIS textbooks available (for example Burrough and McDonnell, 1998; Heywood *et al*., 2006; DeMers, 2002a; Longley *et al*., 2001) and it is not the purpose of this chapter to cover again the basic principles of GIS. It is, however, necessary to offer working definitions of GIS and spatial analysis so that their relationship can be effectively reviewed. What has probably become the 'classic' GIS definition is restated by Goodchild (2000), for example, as a system for creating, storing, manipulating, visualizing and analyzing geographical information. Although slightly different terms are used, the concept of GIS as a toolbox containing these core functions has become nearly universal. Whereas specialist database or visualization software may exist in isolation, the combination of these elements in an integrated software environment is generally considered necessary in order to justify the claim that a software tool is actually a GIS.

Fotheringham and Rogerson (1993) specify that spatial analysis is not just aspatial analysis applied to spatial data: it is inherent in the analytical procedures with which we are concerned here that they aim to reveal and characterize explicitly spatial patterns and processes. More subtly, there is something of a distinction between spatial data manipulation and analysis, although the exact dividing line is dependent on the commentator's view of spatial analysis itself. Techniques for spatial data manipulation, perhaps most extensively developed in the language of cartographic modelling (Berry, 1987; DeMers, 2002b), offer an extensive suite of functions for reclassification, overlay, mathematical, distance and neighbourhood operations on map layers which can be assembled into sophisticated scenarios, of which perhaps the most frequently cited example is site suitability analysis. Although it is possible for the spatial data manipulation tools within a GIS to be assembled in such a way as to carry out spatial analysis tasks, they are generally not considered to constitute spatial analysis *per se*. There is thus a sense

in which spatial analysis requires spatial data manipulation, but manipulation is not in itself analysis.

A distinction can be found between those who adopt a relatively narrow definition of spatial analysis as the extension of statistical analysis into the spatial domain, such as Bailey (1988) and those who would offer a much broader view, including visualization, cartographic modelling and computationally intensive geographical analyses. Bailey and Gatrell (1995) choose to distinguish between spatial analysis and spatial data analysis, the latter describing the situation in which methods are applied to the description and explanation of processes operating in space through the use of observational data within a conventional statistical framework. This narrower definition has strong roots in quantitative geography (see Fotheringham *et al.*, 2000), but tends to marginalize specialized analytical operations within GIS such as hydrological modelling using grid-based functions or network-based modelling for route-finding applications. These tools do not contribute to the more narrowly defined statistical spatial analysis but nevertheless make an important contribution to analytical GIS use.

A further area of development is that which has been termed geocomputation (Longley *et al.*, 1998), in which highly computationally-intensive techniques have been applied to categories of spatial analytical problems which simply could not have been tackled by conventional means. The critical reader may find few fundamentals to distinguish geocomputation from a broadly-defined spatial analysis, except for the use of new data types and computing environments. This work is also characterized by a concentration on some of the areas in which traditional analytical methods have been weak: particularly the application of high-powered computing to the study of spatio-temporal dynamics, perhaps again indicating that geocomputational approaches may serve to fill gaps in the spatial analysis toolkit rather than represent an entirely new development. In the following discussion, we adopt a broad definition, which encompasses a wide range of specialist GIS functions whose purposes are primarily analytical rather than operational. This approach is helpful in understanding the extent to which GIS has contributed to the contextual environment of a wide variety of spatial thinking and analysis tasks, but has had rather less obvious impact on the generation of tools for narrowly defined statistical spatial analysis.

The early development of GIS and spatial analysis techniques were rather separate, with GIS growing from extensive inventory applications such as the Canada GIS (CGIS) (Tomlinson *et al.*, 1976) concerned with the practical management of natural resources, while most spatial analytical techniques can trace their roots to the quantitative revolution in academic geography of the 1960s and 1970s. Typically, spatial analytical methods were developed in the context of limited spatial data and software tools, frequently being programmed in isolation by the researcher to take advantage of the research potential of specific datasets. Widespread adoption of such methods was impossible due to the absence of suitably structured data and widely available software tools. The need for a large body of transferable and well-structured spatial data, for example incorporating the topological relationships required for many types of analysis involving adjacency or contiguity, was a precondition for any broad adoption of spatial analysis methods and it is in this development of spatial data infrastructure that GIS can be seen to have played a critical role. GIS provides the essential tools for manipulation and pre-processing of spatial data that are likely to be required by the spatial analyst. There is thus great attraction to the prospect of

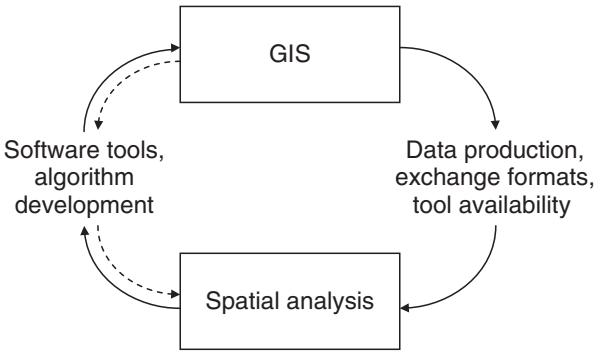


Figure 3.1 Development influences between GIS and spatial analysis.

somehow embedding a wide variety of spatial analysis tools in an existing GIS environment and thereby creating a rich integrated spatial analysis environment, although the achievement of such integration has been elusive. Figure 3.1 illustrates the principal cycle of interaction, whereby GIS use has promoted data production, standardization of formats and the availability of general purpose tools, which have in turn fostered the development of spatial analysis techniques. Relatively few of these analytical techniques have gone on to influence the functions and algorithms available in mainstream GIS. A relatively weak effect is observable in the opposite direction, whereby GIS development *per se* has led to new forms of spatial analysis. The direct impact of spatial analysis on the broader spatial data infrastructure has been very small and is not shown in the figure.

Over time, several books have addressed the theme of spatial analysis in GIS, for example, Fotheringham and Rogerson (1993), Longley and Batty (1996b), Fotheringham and Wegener (2000) and Longley and Batty (2003b). Each of these is characterized by a series of detailed chapters addressing aspects of spatial analysis that have been implemented at the edge of existing GIS technology. Interestingly, the majority of these contributions do not actually use standard GIS software in order to undertake their core spatial analysis functions, while many employ general purpose statistical packages or even develop separate spatial analysis software with various levels and types of connection to GIS. Miller and Wentz (2003) argue that in fact the use of GIS may actually be limiting the types of spatial analysis which is undertaken by many users due to the restrictions that the GIS model places on thinking about spatial relationships and interactions. Their contention is that GIS offers a much richer universe of spatial data representation strategies than are commonly adopted. Certainly, representation and analysis are closely linked. Martin (1999a) shows how different representations of a disease phenomenon can lead to quite different ways of thinking and analyzing its spatial form according to whether the disease process is seen as a point pattern, line vector, areal prevalence or continuous density surface. Miller and Wentz's (2003) particular concern is that the assumptions and alternatives of the conventional Euclidean conception of space go unquestioned by most GIS users. Marble (2000) also identifies overly simplistic representational models as one of the obstacles to demonstrating

the real applicability of spatial analysis, citing the prevalence of simple distance and the absence of direction from most spatial analytical work, despite its relevance to practical decision-making. This view that the contribution of GIS to spatial analysis is strongly tied to its provision of the underlying representational models is consistent with Goodchild's (1987) suggestion that an 'ideal' GIS would be one which incorporated a data model finely tuned to the needs of spatial analysis. At that relatively early stage in GIS take-up, he was able to observe that no contemporary commercial product met the ideal and that there would be little economic incentive for the development of a GIS incorporating such a spatial analytical model, while applications rather than abstract concepts are the drivers of proprietary software development.

Very many GIS users are not actually concerned with statistical spatial analysis, but have entirely valid requirements involving the management, query and reporting of spatially referenced data. For example, the UK census agency, the Office for National Statistics (ONS), implemented a GIS for the design of the 2001 census of population, starting with a prototype system in the mid-1990s (Martin, 1999b). The initial objective was the simple replacement of the existing labour-intensive process of creating maps for the guidance of census enumerators. A significant multi-user GIS involving sophisticated data management of multiple data sources, including a national address-level database, was established with no spatial analytical ambitions, the primary objective being to deliver printed maps and address listings for 175,000 census enumeration districts. Although aspects of this system could clearly have been developed with spatial analytical purposes in mind, it shared its principal objectives with perhaps the majority of commercial GIS implementations whose objectives are facility management and inventory applications. The ONS example is a useful one to illustrate the GIS-spatial analysis relationship because it subsequently evolved to become the basis of a spatial referencing system for census outputs that provide a rich source of socio-economic data for spatial analysis. Importantly, the contribution of the GIS application was not in the provision of analysis tools *per se* but as the means of contributing to the wider spatial data infrastructure, including user awareness and debate. In many ways this is a microcosm of the historical role of GIS in spatial analysis.

Couclelis (1998) makes some observations about the contrast between GIS and geocomputation which are also illustrative of the GIS-spatial analysis relationship. GIS has been characterized by large scale, high-visibility practical applications, resulting in great commercial and organizational interest, combined with the intuitive and visual appeal of map-based manipulation by computer. Geocomputation, and spatial analysis more generally, does not enjoy these advantages: the more sophisticated analytical methods are often lacking in immediate or obvious commercial application, are often hard to visualize and are far from intuitive to novice users. We can conclude that, although related, GIS software is not the principal driver of spatial analytical tool development. Almost always, advanced spatial analysis methods are developed separately from GIS, but in an environment in which data availability, especially in standard formats, is due to the wider adoption of GIS. Widespread use of GIS has brought about spatial data infrastructures and exchange mechanisms that make possible the practical implementation of spatial analyses that would otherwise have been quite intractable. GIS have thus come to provide the environment rather than the tools for innovative spatial

analysis, with explicit software connections between the two coming much later, if at all.

3.3. CLOSE COUPLED, LOOSE COUPLED, UNCOUPLED?

Ungerer and Goodchild (2002) provide a tabular representation of strategies for coupling GIS and spatial analysis, which is itself based on a classification by Goodchild *et al*. (1992). The coupling strategies are further illustrated in Figure 3.2 and range from isolated, through loose and close coupled to integrated: only in the case of full integration are spatial analysis functions actually performed within the GIS software itself. The extent of integration possible will to some extent be determined by the software architecture of specific GIS employed. While it is clearly possible to write stand-alone software to perform spatial analysis tasks, it is hard to identify strong advantages to this approach. Indeed the isolation from the data layers available in GIS and the obvious risk of 'reinventing the wheel' in the authoring of such tools serve to make such a strategy unattractive. At the opposite extreme, where spatial analysis functions are fully integrated within GIS software, there is a risk of promoting 'naïve' or inappropriate use of complex techniques due to a lack of specialist insight in spatial analysis. Openshaw (1996) identifies one element of this in what he terms the 'user modifiable areal unit problem' in which the well-recognized modifiable areal unit problem (Openshaw, 1984) is compounded by the availability of software that allows users extensive opportunities for creating their own spatial aggregation schemes without any necessary understanding of the impacts on spatial analysis of the resulting data. Of the intermediate positions, loose

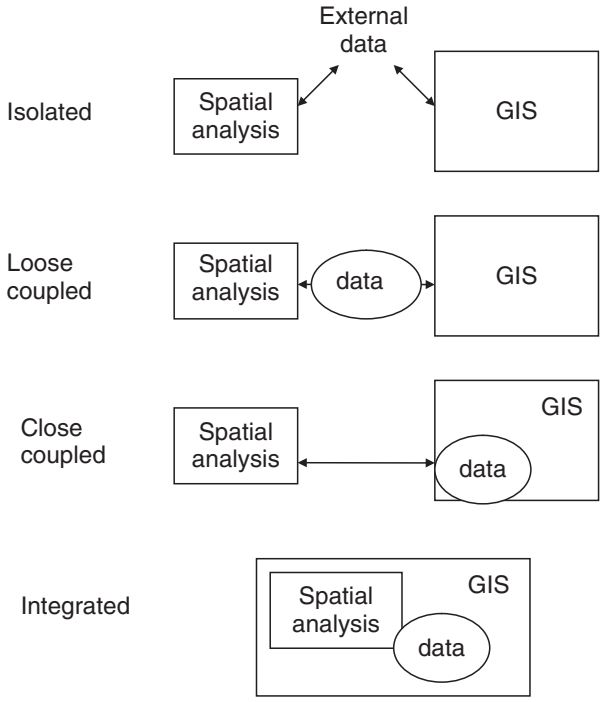


Figure 3.2 Models of relationship between spatial analysis and GIS software (after Ungerer and Goodchild, 2002 and Goodchild *et al.* 1992).

coupling generally involves file import and export at each analysis stage but little new programming, whereas close coupling seeks to overcome this necessity by investing in programming that more smoothly moves data between the two software applications, for example by developing software routines that directly access the GIS database as shown in Figure 3.2.

3.4. CASE STUDIES

In this section we briefly review a range of case studies in which spatial analysis software is more or less closely coupled to GIS. Examples are provided of each of the situations illustrated in Figure 3.2. Some of these examples will be encountered elsewhere in this book, but the objective in considering them here is not to provide an overview of the analysis methods, but to review the role of GIS in the implementation of these spatial analysis tools.

3.4.1. Isolated

Some isolated spatial analysis tools have very specific and limited applications while others are well-developed spatial analysis toolkits. These programs rarely justify the term of GIS in their own right, as one or more of the basic GIS operations (often in the data creation, storage and manipulation domains) are entirely missing or very elementary.

GWR, the software produced by Fotheringham *et al*. (2002) for geographically weighted regression (http://ncg.nuim.ie/GWR), serves as an example of an explicitly spatial analysis method which has been implemented entirely separately from GIS software. In this case, although the input data are conventional spatial coordinates with associated attributes, generic input and output file formats are used and the software operates independently of any GIS. An editing tool has been developed in Microsoft Visual Basic (VB) which provides a user interface to the developers' Fortran regression program and produces outputs which are intended for further analysis in other software, including GIS. It is in the very nature of geographically weighted regression that the results are themselves spatial data, comprising parameter estimates and other statistics relating either to every sample location or every point on a regular spatial grid. Interpretation of these results requires cartographic visualization, but it is expected that the user will undertake this using other software, for which purpose two GIS output file formats are offered. Code is also available for running GWR within the statistical package R, although this provides no direct data management or mapping functions.

A second example is GeoDa (http://www.csiss.org/clearinghouse/GeoDa/) which incorporates limited data manipulation, but has a range of spatial analysis functions and visualization tools and works with the less sophisticated GIS data structures such as Shapefiles. Anselin (1999) explains how this type of exploratory spatial data analysis can bridge the gap between cartographic visualization and statistical analysis. GeoDa is a tool for exploratory spatial data analysis (ESDA), and allows the user to work with linked plots and interactive visualizations, a distinctive characteristic of ESDA tools (Brunsdon and Charlton, 1996). The spatial analysis methods present in GeoDa focus on measures of spatial association, particularly the calculation of local indicators and weights. Spatial data manipulation functions are limited, but do allow for point and polygon data through tools for the creation of centroids and Thiessen polygons. The software can thus be used to provide

additional spatial analysis functions to the GIS user through file export, or to provide stand-alone analysis of suitably structured point or polygon data (Anselin, 2005).

Accession (http://www.accessiongis.com/) provides another interesting example, whereby a software tool has been produced specifically for the calculation of geographical accessibility. Higgs (2005) provides an extensive review of health accessibility modelling in GIS, but notes that attempts to incorporate public transport accessibility are underdeveloped. This tool has been designed to undertake precisely that task, and thereby illustrates an approach to the concerns of Miller and Wentz (2003) by combining conventional spatial network analysis with the very unconventional spaces of public transport timetables. The software offers a wider range of conventional GIS functions than GWR or GeoDa but is still not a fully developed general-purpose GIS, its unique functionality being the spatial analysis of accessibility using a combination of timetable and network data.

3.4.2. Loose coupled

AZM (http://www2.geog.soton.ac.uk/users/martindj/davehome/software.htm) is a loose-coupled tool because it does not undertake any data management or display itself, but requires data import and export from a GIS. In this case, the software is intended for automated zone design and best-matching of incompatible zonal systems (Martin, 2003a) and is dependent on an external GIS to provide the topological data structure which is a central requirement of zone design. More recently, the software has been re-engineered, again to take direct advantage of widely-used Shapefiles, with the additional topological structuring being undertaken within the software. This is an interesting example because its purpose is not to be used as a stand-alone tool but to supply a spatial analysis function to the GIS user that is not otherwise available within the GIS software environment. In this sense it provides additional external functionality to the GIS user, who must manually export and transfer the necessary data.

The history of AZM demonstrates something of the separate origins of GIS and spatial analysis tools noted above. Openshaw (1977) describes an automated zoning procedure (AZP) initially developed to run on an exemplar dataset comprising a limited set of regular cells, which could be aggregated into clusters according to a variety of objective functions. Although the method was of demonstrable practical utility, the absence of widely available topologically structured census or administrative area boundaries and the small problem size that could be handled by available computing power meant that the method was hardly applied until Openshaw and Rao (1995) returned to the problem, using 1991 census data and mid-1990s computing to demonstrate its practical large-scale application. Effectively, the practical application of the method had to wait until GIS development had fostered the general availability of the necessary data in a suitable topological structure. AZM is based around Openshaw's AZP and is closely related to the system used to create output areas for the 2001 census of population in England and Wales, itself a loose-coupled configuration with zone design software processing topologically structured data exported from an ArcInfo GIS application.

3.4.3. Close coupled

SAGE (Spatial Analysis in a GIS Environment) is another example of a system developed as a spatial analysis toolkit (Haining *et al.*, 2001) but this time calling software

routines within the ArcInfo GIS. Although SAGE consisted of external custom-written code, data were held within the GIS, whose functionality was also called for specific data manipulation functions and cartographic visualization. The software was developed specifically to overcome perceived analytical shortcomings in the GIS, yet with a desire not to reinvent those important functions which were already well provided for. Specifically, SAGE attempted to enhance the GIS functionality in the areas of visualization and statistical techniques. Although cartographic visualization is one of the central functional elements of GIS, scientific visualization, particularly that involving real-time interaction with datasets, is generally absent from GIS software. SAGE incorporated exploratory analyses through the use of linked windows common to many ESDA applications. The specific motivation for creation of SAGE was the analysis of health events. Haining *et al.* (2001) explain the rationale for creating a spatial analysis software suite integrated with a proprietary GIS, citing the inconvenience of having to transfer data between two software tools, but also the unnecessary duplication of effort when external tools need to provide their own basic mapping and spatial manipulation functions which are already well-provided for by GIS. At the core of the spatial analysis tool were two separate programs, one providing the spatial analysis and the other a linkage tool, both running in client/server mode with the GIS. SAGE provided a range of classification and regionalization functions in addition to spatial statistical analyses.

The fate of systems such as SAGE is typical of many such attempts in that although a great deal was achieved, the lack of true integration between the two software systems and the academically driven motivation for the analysis program resulted in divergence. Subsequent releases of the ArcInfo and ArcGIS software have moved to different operating systems and hardware architectures, and eventually the adoption of different scripting languages, making SAGE unusable with more modern versions. External, non-commercial tools such as SAGE cannot realistically hope to track the relatively rapid software redesign cycle of leading GIS software. The analytical functions embedded in SAGE were not absorbed into the GIS software, so there has actually been a decrease in the range of tools available to the spatial analyst. Isolated and loose-coupled tools, relying only on generic spatial data transfer formats, will probably survive several GIS software versions without the need for significant reprogramming. Similarly, fully integrated tools have the potential to evolve with the GIS itself if they are actually adopted as part of the core product. Close-coupling however is perhaps the most problematic software architecture, carrying a high risk of being left behind by developments in the GIS and the greatest maintenance burden for the spatial analysis programmer if they are to ensure the continued utility of their tool.

Ungerer and Goodchild (2002) describe a close-coupled component object model (COM) approach to linking GIS and spatial analysis software. Their tool is an extension written for ArcInfo which undertakes spatial interpolation by creating an instance of a statistics package, using it to run an analysis on the GIS data and then placing the results within the GIS. This is just one step short of writing spatial analysis functions that are fully integrated with the host GIS. Their implementation uses Microsoft Visual Basic for Applications (VBA) which has become common as a macro language across multiple software packages, overcoming some of the restrictions of software-specific macro programming languages found, for example, in earlier GIS versions. Clearly a programming language of this type could be used to

develop entirely integrated spatial analysis tools but this example demonstrates its power as a means for finding a 'common language' for close-coupling GIS with external statistical software.

3.4.4. Integrated

In addition to those analytical functions which are actually included as part of the core software, examples of customized spatial analysis operations fully integrated within GIS may be found at all periods in GIS development. These are generally the result of spatial analysts being able to directly access macro programming functions. Early instances involved languages such as ArcInfo's Arc Macro Language (AML), while more recent examples are likely to use Microsoft VBA, perhaps interfacing directly with components of the GIS software such as ArcObjects.

Ding and Fotheringham (1991) describe an application called STACAS (SpaTial AutoCorrelation and ASsociation analysis) that was completely embedded within the GIS software, being assembled from ArcInfo functions and custom-written programs. As with GeoDa described above, analysis of spatial association requires knowledge of the spatial relationships between GIS objects, for example the adjacencies between polygons and distances between points or polygon centroids. It is also necessary to link attribute values with these locations and of value to display the resulting measures of association in cartographic form. For all of these reasons there is a considerable attraction to embedding the analytical functions within the GIS environment where the spatial relationships and support functions are already available. Ding and Fotheringham's solution was to construct their analysis routines using ArcInfo's own macro programming language, AML. Calculations that could not be readily assembled using AML were programmed in C and called from within the AML routines so that the resulting analysis functions were presented to the user as additional commands within the GIS. Embedding of this type is generally robust against incremental updating of GIS software but becomes obsolete when major changes to software architecture take place, affecting the spatial database and macro programming language on which it is based.

Evans and Steadman (2003) describe a more modern application, interfacing a land use transport model known as TRANUS with a desktop GIS. The objectives are to quickly visualize the results of the transport model and to provide a means of exporting data for further analysis in additional software environments. The TRANUS GIS module has been built using ArcObjects technology from ESRI's ArcGIS which effectively allows Microsoft VB to be used to customize interfaces and develop further software. Automated procedures handle the transfer of results between the transport modelling and GIS tools. In this instance visualization in the GIS is not the final objective, with model outputs being passed on from the GIS to other external analysis tools. Effectively the GIS provides the visualization and post-processing of specialized model results. The GIS environment is additionally relevant as the context for the creation and manipulation of many of the data layers that contribute to the original transport modelling. Interestingly, the authors note that a question mark hangs over the demand for such integrated or closely coupled solutions.

3.5. BARRIERS AND OPPORTUNITIES

Brown (2000) argues strongly that after so many years of discussion, not enough

progress has been made towards the genuine integration of spatial analysis and GIS, especially when considered from the perspective of the substantive researcher who has practical analysis requirements but is not able to engage in the development of software tools. He notes that the growth of GIS has been propelled by the spread of less sophisticated GIS (such as ArcView) that are less readily turned to spatial analytical applications. The result is that while there is widespread use of GIS, this is often naïve or at least goes little further than cartographic visualization. It follows from this reasoning that it is the spatial analytical tools embedded within the simplest GIS software, not the most sophisticated, that will actually determine the future uptake and development of spatial analysis methods. Given the enormous contextual influence of GIS on the practical use of spatial analysis, prevalent standards of GIS training can be seen to have a significant impact on the overall level of spatial analytical methods demanded and employed.

Public awareness of spatial data continues to increase massively through the popularity of web-based mapping tools, of which Multimap (http://www.multimap.com/), the Neighbourhood Statistics Service (http://www.neighbourhood.statistics.gov.uk/), Windows Live Local (http://local.live.com/) and Google Earth (http://earth.google.com/) provide just a few examples. These developments bring spatial data and concepts onto the desktops of millions who will remain unaware that there has even been a debate about the role of GIS in spatial analysis. Such tools embody various simple GIS analysis functions such as route-finding (Multimap, Windows Live Local), tagging and grouping of spatial objects (Google Earth) and interactive choropleth mapping (Neighbourhood Statistics). While it seems unlikely that these 'populist' tools will develop a need for much more sophisticated spatial statistical functions, there is every possibility that they find increasing use in the presentation of results and visualizations from complex analyses run externally, for example of climate change, environmental sensitivity or neighbourhood property prices. The increasing pool of low-level users remains at the same time one of the greatest opportunities for spatial analytical development, yet a barrier to the emergence of a well-skilled user base.

Goodchild (2000) sees four tensions in the popularization of spatial analysis through incorporation of tools within GIS software: (a) populism and elitism, (b) visual and numeric, (c) open and closed, and (d) local and global. The first of these, populism and elitism, is very much concerned with the difficulty noted above: although GIS use is becoming massively more widespread, this does not directly increase the ability of users to appropriately engage with sophisticated spatial analysis methods. There is in reality no organization with the authority to either 'restrict' or 'educate' GIS users in this respect, so the spatial analysis community must address itself to the challenge of awareness-raising among an ever-multiplying community of low-level GIS users. The incorporation of visualization functions in spatial analysis tools, for example in GeoDa described above, goes some way towards the enhanced communication of spatial analysis concepts to more advanced GIS users who might otherwise be unlikely to engage with purely statistical aspects. An increasing tendency towards open-source software development may eventually assist in exposing underlying algorithms but it is inevitably the case that only a small proportion of users will concern themselves with such a level of technical detail. The fourth tension between local and global analysis represents a continuum, with a need for analytical techniques appropriate for each scale of analysis.

Two of the outstanding technical barriers facing spatial analysis and GIS are the (related) development of methods and techniques that begin to seriously tackle both space and time, a dimension whose importance is considered critical to the future integration of GIS and spatial analysis by Marble (2000), and the availability of large computational models to the ordinary user. Batty (2003) suggests that the inability to adequately handle temporal dynamics has 'long been the Achilles heel of geography and GIS' (p. 83) and when considering many of the current 'grand challenges' of spatial analysis this certainly remains the case. Real world problems frequently demand answers with temporal dimensions, for example 'how might this neighbourhood change over time?', 'what will happen here in an extreme flood event?' or 'how will the best route change as congestion increases?' The GIS industry has never developed a consensus model for temporal representation (Langran, 1992; Peuquet, 2002) and many spatio-temporal analyses based on GIS technology have for the most part continued to use inadequate data models. While this may have been sufficient when data originated from pre-digital sources such as land surveys and population censuses, it is inadequate to high-frequency satellite imagery or real time monitoring of traffic flows or mobile telephony operations. These data volumes not only offer the potential for genuinely temporal analysis, requiring new data architectures and analytical techniques, but also demand massively greater computational power. Spatio-temporal dynamics was an important element of the geocomputational techniques noted above (Longley *et al*., 1998), and contemporary developments in pervasive and grid-enabled computing (Martin, 2003b) seem set to offer the data access and computational power required for a new generation of spatio-temporal analysis.

3.6. CONCLUSION: CONVERGENCE OR DIVERGENCE?

Marble (2000) sees it as essential that developments in both GIS and spatial analysis achieve closer integration. In this context, he specifically cites the role of both spatial and temporal aspects. His argument is that researchers in both domains must more seriously get to grips with modern computational approaches. An obstacle to this is seen as the conservative (actually 'myopic' Marble, 2000, p. 32) definition of spatial analysis as only that which is strictly statistical spatial analysis (a distinction certainly made by Bailey (1998) and endorsed by Ungerer and Goodchild (2002)) resulting in the exclusion of some of the more modelling-oriented approaches described above. The lack of integration is primarily due to the characterization of user demand in determining what functionality is incorporated into commercial GIS software. Key to further integration is therefore the unambiguous demonstration of the utility of spatial analytical approaches which have the capability of stirring up user demand to see different types of tools within their GIS environments. Operationally, Marble sees the adoption of object-oriented data models as one of the keys to advancing integration. Moves in this direction in the data architectures of major software such as the most recent versions of ArcGIS certainly provide far richer environments for the customization of the GIS and the writing of new spatial analysis tools using languages such as VBA, as used by Ungerer and Goodchild (2002). Longley and Batty (2003a) trace the historical development of GIS and its extension to incorporate contemporary spatial analysis, specifically drawing out the three themes of temporal and spatial representation, agent and institutional communications and geographical networks. These are major areas in which both GIS and

stand-alone spatial analysis software have a long way to go. Again, it is the need for much more sophisticated handling of space and time and the incorporation of different types of spatial computation that are the underlying themes.

In the preceding sections we have reviewed various examples of the relationships between GIS and spatial analysis which, despite differences of detail, display remarkably little change over the last two decades. It seems improbable that GIS software intended for an increasingly wide user base will ever incorporate a high level of spatial analytical functionality as the use of complex and advanced methods will never be a concern of the ordinary GIS user. Although the absolute levels of spatial analytical functionality in GIS continues to increase, the gap between populist software and research-oriented analytical tools cannot be closed in relative terms. More realistically, a groundswell of open software standards and, potentially, grid-based computing applications may make practical communication between GIS software and the more sophisticated analysis tools much easier. There is thus no greater prospect of true convergence between GIS and spatial analysis than at any previous time, yet the two fields will continue to grow and feed off one another. What we still need are more realistic expectations of what drives the design of commercial software and a concerted effort on more sustainable ways of embedding spatial analytical tools within the broader GIS landscape.

REFERENCES

Anselin, L. (1999). Interactive techniques and exploratory spatial data analysis. In: Longley, P., Goodchild, M., Maguire, D. and Rhind, D. (eds), *Geographical Information Systems: Principles, Techniques, Applications and Management*, Second Edition, pp. 253–266. Chichester: Wiley.

Anselin, L. (2005). *Exploring Spatial Data with GeoDa: A Workbook*. Urbana-Champaign: University of Illinois.

Bailey, T.C. (1998). Review of statistical spatial analysis in GIS. In: Fotheringham, A.S. and Rogerson, P. (eds), *Spatial Analysis and GIS*, pp. 13–45. Philadelphia: Taylor and Francis.

Bailey, T.C. and Gatrell, A.C. (1995). *Interactive Spatial Data Analysis*. Harlow: Longman.

Batty, J.M. (2003). Agent-based pedestrian modelling. In: Longley, P.A. and Batty J.M. (eds), *Advanced Spatial Analysis: The CASA Book of GIS*. pp. 81–106. Redlands, CA: ESRI Press.

Berry, J.K. (1987). Fundamental operations in computer-assisted map analysis. *International Journal of Geographical Information Systems*, **1**(2): 119–136.

Brown, L.A. (2000). The GIS/SA interface for substantive research(ers): a critical need. *Geographical Systems*, **2**: 43–47.

Brunsdon, C. and Charlton, M. (1996). Developing an exploratory spatial analysis system in XLisp-Stat. In: Parker, D. (ed.), *Innovatons in GIS 3* pp. 135–146. London: Taylor and Francis.

Burrough, P.A. and McDonnell, R.A. (1998). *Principles of Geographical Information Systems*. Oxford: Oxford University Press.

Couclelis, H. (1998). Geocomputation in context. In: Longley, P.A., Brooks, S.M., McDonnell, R.A. and Macmillan, B. (eds), *Geocomputation: A Primer* pp. 17–30. Chichester: Wiley.

DeMers, M.N. (2002a). *Fundamentals of Geographic Information Systems*, Second Edition Update. New York: Wiley.

DeMers, M.N. (2002b). *GIS Modelling in Raster*. New York: Wiley.

Ding, Y. and Fotheringham, S. (1991). *The Integration of Spatial Analysis and GIS: the Development of the STACAS Module for ArcInfo*. Technical Paper 91–5, National Center for Geographic Information and Analysis, Buffalo, NY: NCGIA.

Evans, S. and Steadman, P.J. (2003). Interfacing land-use transport models with GIS: the Inverness model. In: Longley, P.A. and Batty, J.M. (eds), *Advanced Spatial Analysis: The CASA Book of GIS*, pp. 289–308. Redlands, CA: ESRI Press.

Fotheringham, A.S., Brundson, C. and Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis*. London: Sage.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression*. Chichester: Wiley.

Fotheringham, A.S. and Rogerson, P.A. (1993). GIS and spatial analytical problems. *International Journal of Geographical Information Systems*, **7**(1): 3–19.

Fotheringham, A.S. and Wegener, M. (2000). *Spatial Models and GIS: New Potential and New Models.* London: Taylor and Francis.

Goodchild, M.F. (1987). A spatial analytical perspective on geographical information systems. *International Journal of Geographical Information Systems*, **1**(4): 327–34.

Goodchild, M.F. (1992). Geographical information science. *International Journal of Geographical Information Systems*. **6**(1): 31–45.

Goodchild, M.F. (2000). The current status of GIS and spatial analysis. *Geographical Systems*, **2**: 5–10.

Goodchild, M.F., Haining, R., Wise, S. and 12 others (1992). Integrating GIS and spatial data analysis: problems and possibilities. *International Journal of Geographical Information Systems*, **6**(5): 407–23.

Haining, R., Wise, S. and Ma, J. (2001). Providing spatial statistical data analysis functionality for the GIS user: the SAGE project. *International Journal of Geographical Information Science*, **15**(3): 239–254.

Higgs, G. (2005). A literature review of the use of GIS-based measures of access to health care services. *Health Services and Outcomes Research Methodology*, **5**(2): 119–39.

Heywood, I., Cornelius, S. and Carver, S. (2006). *An Introduction to Geographical Information Systems*, Third Edition. London: Pearson.

Langran, G. (1992). *Time in Geographic Information Systems*. London: Taylor and Francis.

Longley, P.A. and Batty J.M. (1996a). Analysis, modelling, forecasting, and GIS technology. In: Longley, P.A. and Batty J.M. (eds), *Spatial Analysis: Modelling in a GIS Environment*. pp. 1–16. Cambridge: GeoInformation International.

Longley, P.A. and Batty, J.M. (eds) (1996b). *Spatial Analysis: Modelling in a GIS Environment.* Cambridge: GeoInformation International.

Longley, P.A. and Batty J.M. (2003a). Advanced spatial analysis: extending GIS. In: Longley, P.A. and Batty, J.M. (eds), *Advanced Spatial Analysis: The CASA Book of GIS.* pp. 1–18. Redlands, CA: ESRI Press.

Longley, P.A. and Batty J.M. (eds) (2003b). *Advanced Spatial Analysis: The CASA Book of GIS.* Redlands, CA: ESRI Press.

Longley, P.A., Brooks, S.M., McDonnell, R.A. and Macmillan, B. (eds) (1998). *Geocomputation: A Primer.* Chichester: Wiley.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2001). *Geographic Information Systems and Science*. Chichester: Wiley.

Marble, D. (2000). Some thoughts on the integration of spatial analysis and Geographic Information Systems. *Geographical Systems*, **2**: 31–35.

Martin, D. (1999a). Spatial representation: the social scientist's perspective. In: Longley, P., Goodchild, M., Maguire, D. and Rhind, D. (eds). *Geographical Information Systems: Principles, Techniques, Applications and Management*, Second Edition, pp. 71–80. Chichester: Wiley.

Martin, D. (1999b). The use of GIS in census planning. In: Stillwell, J., Geertman, S. and Openshaw, S. (eds), *Geographical Information and Planning*, Berlin: Springer. pp. 283–298.

Martin, D. (2003a). Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science*, **17**(2): 181–196.

Martin, D. (2003b). Reconstructing social GIS. *Transactions in GIS*, **7**(3): 305–307.

Miller, H.Z. and Wentz, E.A. (2003). Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers*, **93** (3): 574–594.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers*, **NS 2**(4): 459–472.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography 38*. Norwich: Geo Books.

Openshaw, S. (1996). Developing GIS-relevant zone-based spatial analysis methods. In: Longley, P.A. and Batty, J.M. (eds) (1996). *Spatial Analysis: Modelling in a GIS Environment*, pp. 55–73. Cambridge: GeoInformation International.

Openshaw, S. and Rao, L. (1995). Algorithms for reengineering 1991 Census geography. *Environment and Planning A*, **27**(3): 425–446.

Peuquet, D. (2002). *Representations of Space and Time*. New York: Guilford.

Tomlinson, R.F., Calkins, H.W. and Marble, D.F. (1976). *Computer Handling of Geographical Data*. Paris: UNESCO Press.

Ungerer, M.J. and Goodchild, M.F. (2002). Integrating spatial data analysis and GIS: a new implementation using the component object model (COM). *International Journal of Geographical Information Science*, **16**(1): 41–53.

4

Geovisualization and Geovisual Analytics

Urška Demšar

Geographic information science has encountered a new challenge in the recent explosion of availability of spatial data sets. Current spatial data sets tend to be very large – examples are the terabytes of data generated by Earth Observation Satellites, census databases and large databases of climate and environmental data. The data are recorded via sensors and monitoring systems that capture many parameters, which makes the data highly multidimensional. Another trend which follows current developments in spatial data interoperability and management is that data from different and until recently incompatible sources are nowadays commonly integrated into larger and even more multidimensional collections (Miller and Han, 2001).

These data are regarded as a source of potentially valuable knowledge, which exists in the form of patterns and relationships in the data. Uncovering such knowledge is sometimes a difficult task, which current computational methods are not always able to perform, especially in large and highly multidimensional data sets. This is where visual exploratory data analysis becomes useful, since multivariate visualization is one way in which humans can deal with complex data. Its purpose is to reveal knowledge in the data which is not detectable by current computational methods, but which can easily be identified by the human visual system. The value of visualization is in the fact that it can force us to notice something in the data that we never expected to see. As Plaisant (2004) puts it, 'Information visualization is sometimes described as a way to answer questions you didn't know you had'.

This chapter discusses visualization as means for exploring spatial data with the aim to create new knowledge and provide

new scientific insight. Visual data exploration implies generation of new ideas through creation, inspection and interpretation of visual representations and can be considered a part of Exploratory Data Analysis (EDA) (Tukey, 1977). When looking at spatial data, we are talking about Exploratory Spatial Data Analysis (ESDA) (Unwin and Unwin, 1998). Visual exploration is essential as the first step of data analysis and serves to uncover any indications of what there actually is in the data, to prompt ideas and generate hypotheses. It is usually followed by confirmatory data analysis and as the last step by visual communication where results are presented and disseminated in visual form (DiBiase, 1990). This last step is the focus of traditional cartography, which is beyond the scope of this chapter.

The rest of this chapter is structured as follows: the following section introduces the general visualization terminology, describes what role visualization plays in data exploration, presents one of the many possible classifications of visualization methods and lists some examples of general (not necessarily spatial) visualization methods. The rest of the chapter focuses on geospatial data, presents the state-of-the-art in geovisualization research, lists a brief selection of geovisualization software and shows several examples. Finally, a new emerging discipline of Geovisual Analytics is introduced together with some of the future challenges in geovisualization research.

4.1. INFORMATION VISUALIZATION AND VISUAL DATA EXPLORATION

Visualization is the graphical (as opposed to textual or verbal) presentation of data. It translates complex data into visual displays where a human can look for structure, patterns, trends and relationships that make it easier to quickly perceive the significant aspects and characteristics of the data. The main purpose of visualization is to provide insight into data, which is usually done by displaying them with reduced complexity, while at the same time preserving the interesting structure characteristics and minimizing the loss of information. 'Scientific visualization' was first defined 20 years ago (McCormick *et al*., 1987) as the use of computing technology to create visual displays with the goal to facilitate thinking and problem solving. The term 'data visualization' sometimes stands as a synonym for scientific visualization and is usually defined as visualization of data that have a natural geometric structure. A more general term 'information visualization' refers to graphical representations of any type of data, including abstract structures, such as trees, networks or graphs. Even though borders between these different terms are sometimes blurred, in all cases the emphasis is on supporting knowledge construction from visual displays of data (Card *et al*., 1999; Fayyad *et al*., 2002).

Knowledge construction from data is the process of actively manipulating data in order to discover patterns, relationships or other abstract knowledge representations that facilitate the understanding of the phenomenon under investigation. All knowledge construction is therefore a form of pattern recognition. The most formidable pattern recognition apparatus known to the human race is the human brain, which can analyze complex events in a short time interval, recognize important patterns and make decisions much more effectively than any computer can do. The question is how to enable this formidable apparatus to work in the knowledge construction process. Given that vision is the predominant sense and that computers have been created to communicate

with humans visually, computerized data visualization provides an efficient connection between data and mind to support the data exploration process (Keim and Ward, 2003).

The main goal of visual data exploration is to get an idea of what the data contain, or what the data look like. This process does not provide a complete understanding of the phenomenon behind the data – that is not the point. Visual data exploration is intended to provide ideas about the general characteristics of the data which are to serve as a basis for new hypotheses. These can then be further tested using confirmatory data analysis methods (for example, statistics or other mathematical methods). The observations can also be used to choose an appropriate method for further scientific in-depth analysis (Keim and Ward, 2003).

Visual data exploration is usually performed in three steps according to the Visual Information Seeking Mantra (Shneiderman, 1996): overview first, zoom and filter, then details-on-demand. One of the fundamental concepts in this process is interaction. The user can typically interact with the visualization in a number of different ways, such as browsing, selecting, querying and manipulating the graphical parameters or displaying other available information about the data – all with the goal to discover interesting patterns which are valid, novel, useful and comprehensible. A valid pattern is general enough to apply to new data. Novel means that the pattern is non-trivial and unexpected. Usefulness refers to the property that the pattern can be used for either decision-making or further scientific investigation. Comprehensibility means that the pattern is simple enough to be interpretable by humans, which is important because the analyst's trust in the exploration result depends on how comprehensible it is to him/her (Miller and Han, 2001).

There are many ways to represent data graphically. There are also many ways of grouping these displays according to some orderly fashion, such as for example if their focus is geometric or symbolic, if the display is static or dynamic, according to the amount of structure the visualization method requires, etc. One of the more comprehensive classifications is presented by Keim and Ward (2003), who construct a three-dimensional space of visualizations by classifying the methods according to three orthogonal criteria: the data type, the type of the visualization method and the interaction method (Figure 4.1). Table 4.1 names some examples of each type of visualization methods according to Keim and Ward's classification – to give the reader some idea what kind of methods we are talking about. A more comprehensive coverage of information visualization techniques can be found for example in Card *et al*., (1999), Ware (2000) or other recent books on information visualization.

4.2. GEOVISUALIZATION AND SPATIAL DATA EXPLORATION

Geovisualization or visualization of geospatial data (any data with a given geographic location) is defined as the use of visual representations in order to employ the vision to solve spatial problems (MacEachren *et al*., 1999). It can be considered as a perceptual-cognitive process of interpreting and understanding georeferenced visual displays and 'provides theory, methods and tools for visual exploration, analysis, synthesis and presentation of geospatial data' (MacEachren and Kraak, 2001). While its roots lie in cartography and geographic techniques for representing spatial data, geovisualization integrates these traditions with scientific and information visualization principles and

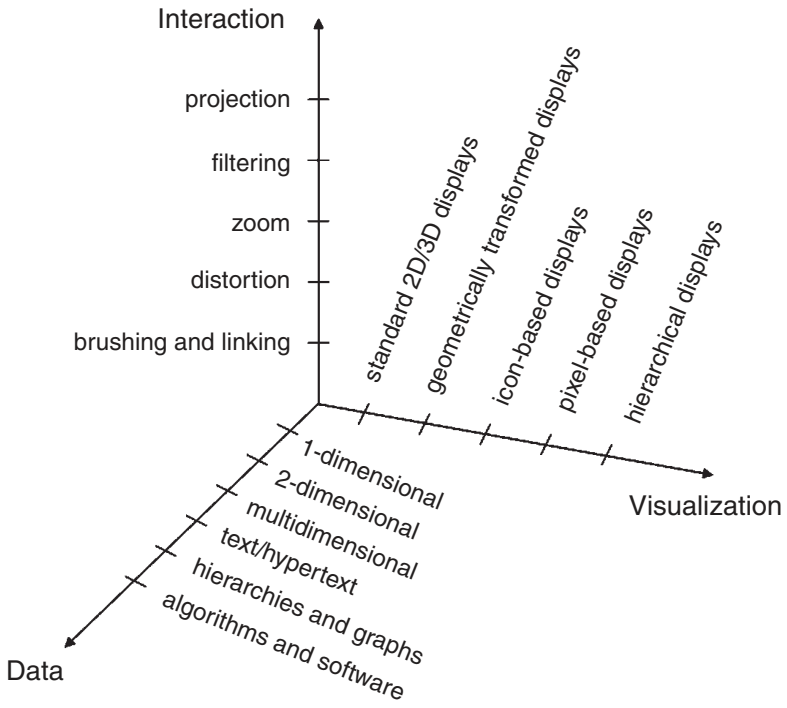


Figure 4.1 The three-dimensional space of visualizations (redrawn after Keim and Ward (2003)).

developments in exploratory data analysis. Research in this discipline dates several decades back, starting with Bertin's work on cartographic design (Bertin, 1967/1983) and followed by the establishment of the Commission on GeoVisualization of the International Cartographic Association (ICA) in 1995 (ICA, 2008), which has ever since played a major role in the development of the discipline. For those interested, a more detailed description of the history of geovisualization can be found in MacEachren *et al.*, (2004). Let it suffice to say that the area has advanced from the first attempts of analyzing how maps facilitate scientific thinking into a broad multidisciplinary research discipline that converges theory, methods and ideas from information visualization, cartography, graphic design, image analysis, perception and cognition, computer science, statistics, exploratory data analysis (EDA), knowledge discovery in databases (KDD) and geographic information science.

In addition to the challenges of ordinary information visualization, namely the volume and multidimensionality of data, geovisualization faces a task of preserving the richness and particular characteristics of geospatial data (such as for example spatial dependency and spatial heterogeneity). With the display possibilities restricted to the usual two or three dimensions plus perhaps the additional dimension given by time and animation, geovisualization provides a clear linkage to the geographic space, so that the user can relate the observed patterns to a particular geographic location (Fotheringham *et al.*, 2000; Miller and Han, 2001). Most current geovisualization systems attempt to solve this problem by displaying data in a number of linked displays – sometimes called multiple linked views. These displays typically include geographic visualizations,

Table 4.1 Examples of visualization methods, classified according to Keim and Ward (2003)

Visualization type	*Examples of visualization methods*
Standard 2D/3D displays	Line graphs and surfaces (Figure 4.3)
	A histogram
	A kernel plot
	A box-and-whiskers plot
	A scatterplot
	A contour plot
	A pie chart
Geometrically transformed visualizations	Scatterplot matrix
	Multiform bivariate matrix (Figure 4.4) (MacEachren *et al.*, 2003)
	Parallel coordinates plot (Figure 4.4) (Inselberg, 2002)
Icon-based display methods	Star icons (Fayyad *et al.*, 2002)
	Chernoff faces (Chernoff, 1973)
Dense pixel visualizations	Recursive pattern visualization (Keim, 2002)
	Circle segment view (Keim, 2002)
	Spacefills (Figure 4.4) (Gahegan *et al.*, 2002)
Hierarchical displays	Dendrogram as a top-down rooted tree (Müller-Hannemann, 2001), combined with a scatterplot (Seo and Shneiderman, 2002) or mapped on a sphere – The Magic Eye View (Kreuseler and Schumann, 2002)
	A treemap (Bederson *et al.*, 2002)
	A sunburst (Stasko and Zhang, 2000)

such as maps or cartograms, as well as other multivariate visualizations, for example any of the visualization methods described in the previous section or even constructs consisting of several of those visualizations, such as bivariate matrices or similar multi-displays (see, for example, systems in Gahegan *et al.*, (2002), Takatsuka and Gahegan (2002), G. Andrienko *et al.*, (2003a), Dykes and Mountain (2003), etc.). Roberts (2005) has a more comprehensive list of examples. All these displays are usually interactively connected by the concept of brushing and linking, which means that data elements which are in some way interactively selected in one display (usually either through mouse-over operation, direct selection or by some other interaction method) are simultaneously highlighted or selected everywhere. This provides a better visual impression and facilitates pattern recognition across multiple displays. The key word here is interaction: high levels of interaction are necessary for any kind of data exploration task (Dykes, 2005).

Sometimes the sheer volume and complexity of the geospatial data makes it impossible to rely solely on human vision for knowledge discovery. Successful knowledge construction is therefore more likely if the advantages of visual exploration are combined with computational exploration methods. The goal then becomes to construct visually enabled knowledge discovery systems that can facilitate the automatic process of pattern and relationship recognition in complex data and the subsequent interpretation of the discovered patterns and relationships. The data could, for example, first be visually explored with direct manipulation of the visual displays and then when something interesting appears, computational tools could be applied. Alternatively computational data mining can be used as a first pass and the results can then be examined visually only to reiterate the process with another pass of computational mining and/or visual exploration if required. By merging automatic and visual exploration the flexibility, creativity and knowledge of a person are combined with the storage capacity and computational power

of the computer which results in a faster and more effective knowledge discovery. In practice, however, how to enable such synergy is not yet fully understood and the problem of integrating combined and visual exploration tools in the best manner is not trivial to solve (MacEachren *et al*., 1999; Shneiderman, 2001; MacEachren and Kraak, 2001).

Visual data exploration of spatial data has several advantages: it is intuitive and does not require understanding of complex mathematical and computational methodology. It is also effective when little is known about the data, when the exploration goals are vague or when the data are noisy and/or heterogeneous (Keim, 2002). On the other hand, during visual exploration the analyst typically looks at data from various perspectives, at various scales and combines use of multiple techniques and approaches. No single visualization is capable of providing all the required views of the data, from the general overview to indicating various anomalies and patterns. It is therefore often necessary for the analyst to simultaneously use several techniques for various purposes. Different exploration tasks might also require different visualizations. The fundamental questions to address prior to any exploration is what is the current task, what way of thinking does it require and which tools best support the task and way of thinking at hand (Gahegan, 2005). Additionally, it is of course also important to find out which visualization methods are available and what type of data and phenomena they are suitable for. This is not the only complexity issue: during the actual exploration, the analyst is required to decompose the exploration problem into smaller subproblems in a proper and efficient manner which might be different for each exploration task. In the last exploration step, the fragmentary knowledge resulting from each of the subproblem explorations needs to be merged into a consistent interpretation for the entire data set in order to obtain proper understanding of the underlying phenomenon and form appropriate hypotheses. Visual exploration is therefore a complex process which requires training and expertise to be performed properly (G. Andrienko *et al*., 2006).

An important issue to consider when developing new geovisualization tools is therefore how users use these tools and how the tools support particular exploration tasks. These questions can be answered by investigating the usability properties of the tools. Usability is defined as 'the extent to which a computer system supports users to achieve specified goals and does so effectively, efficiently, and in a satisfactory way' (Nielsen, 1993). The idea behind usability is that information systems designed with their users' psychology and physiology in mind are easier to learn and more efficient and satisfying to use. The principle of usability originates from user-centred design in Human–Computer Interaction (HCI), which is a discipline that explores the quality of interaction between the users and information systems. One of the basic requirements for developing a usable and useful information system is knowledge about users and how they use the system. This is the basic principle of the user-centred design, which is a philosophy where the needs, wants, and limitations of the users of an information system are given attention at every stage of the design process (Preece *et al*., 2002).

Design of exploratory geovisualization tools has been technology driven for many years. Tools and systems were developed from a purely technical point of view, where knowledge about users did not play a major role. In recent years, however, the approach has shifted towards user-centred design with the aim of providing useful and usable geovisualization tools which support analytical reasoning (Fuhrmann *et al*., 2005). While the importance of geovisualization

tools for exploration of spatial data has been generally recognized, the issues of usability testing for geovisualization are not exactly the same as those in human–computer interaction and how the visual tools support human analytical reasoning is still not fully explained. Traditional usability methods borrowed from human–computer interaction therefore need to be adapted accordingly. The key issue in visual data exploration is the intuitive search process in a visualized environment. It is therefore necessary to incorporate physiological and psychological findings about the process of human vision as well as knowledge of the relation between geospatial objects and their representation in the process of system engineering (Fuhrmann *et al*., 2005; N. Andrienko and G. Andrienko 2006a). The potentials and limitations of information visualization tools have been explored in numerous recent experiments focusing on some aspect of the usability of geovisualization tools (for example, N. Andrienko *et al*., 2002; Suchan, 2002; Tobón, 2002; Edsall, 2003; Haklay and Tobón, 2003; Slocum *et al*., 2003; Griffin, 2004; van Elzakker, 2004; Ahonen-Rainio, 2005; Koua, 2005; Robinson *et al*., 2005; Tobón, 2005; G. Andrienko *et al*., 2006; Demšar, 2006, 2007a), but much still remains to be investigated.

4.3. GEOGRAPHIC INFORMATION SYSTEMS AND GEOVISUALIZATION SOFTWARE

Today's geovisualization is much more than just map design, even though it is firmly rooted in cartographic traditions of map design and display. Most of the contemporary commercial Geographic Information Systems (GIS) provide a set of mapping tools, with appropriate symbology, graphical representation, classification and so on; nevertheless they differ from the information visualization systems in several ways. For example, data representation in GIS packages is limited to predefined object- (point, line, area) or field-based representations, while information visualization software does not usually have this assumption and treats all data types as equal, regardless if this makes sense geographically or not. This can be beneficial to reveal patterns that would otherwise remain obscured in traditional geographic representations. Most of the GIS also offer only limited support for dynamics, animation, interactivity between a number of different visualizations and any integrated computational methods (although there are some attempts to implement data mining methods in the context of GIS, see for example, Lacayo and Skupin, 2007).

On the other side of the story, there exist numerous information visualization environments that support development of visual exploration systems for multivariate data. Examples of well-known information visualization environments are XGobi, R and SPSS, but for this chapter, those that focus on spatial data are more relevant. Three that deserve a description here are GeoVISTA Studio, CommonGIS and GeoDa, but this selection is far from exhaustive and new tools and environments are developed continuously.

GeoVISTA Studio is a java-based collection of various geographic and other visualizations and computational data mining methods (MacEachren *et al*., 1999; Gahegan *et al*., 2000; Takatsuka, 2001; Dai and Hardisty, 2002; Gahegan *et al*., 2002; Gahegan and Brodaric, 2002; Takatsuka and Gahegan, 2002; Guo, 2003; MacEachren *et al*., 2003; Edsall, 2003; Guo *et al*., 2004; Guo *et al*., 2005; Robinson *et al*., 2005). Its components are implemented as Java Beans, which are self-contained software components that can be easily connected into a customized data exploration system

by visual programming. Furthermore, using Java Beans technology makes it possible to integrate external methods and bespoke components in the system. Visualizations include a parallel coordinates plot, various bivariate visualizations (scatterplots, spacefills, bivariate maps, etc.) that can be either independent or elements in different types of multiform matrices, as well as time series visualizations and visual classifiers. Computational methods include a statistics package and several types of classification methods (*k*-means, ISODATA, maximum likelihood and a Self-Organising Map) with respective visualizations. Studio is free and can be downloaded from its website http://www.geovistastudio.psu.edu. Two of the figures in this chapter were produced using GeoVISTA-based exploration systems (Figures 4.4 and 4.5). A selection of GeoVISTA components has recently been assembled in an application version called the GeoVizToolkit, which is freely available at http://www.geovista.psu.edu/geoviztoolkit/. This application has a user-friendly interface and represents a good starting point for learning how to explore multidimensional spatial data. It is a lot easier to learn and use than original GeoVISTA Studio, but it does not provide the full functionality and all computational capabilities of the Studio.

CommonGIS consists of various methods for cartographic visualization, non-spatial graphs, tools for querying, search and classification and computation-enhanced visual techniques for exploration of spatio-temporal data. Main features are interactive thematic mapping techniques, statistical computations and displays, animated maps, dynamic queries, table lenses, parallel coordinate plots and time-aware geovisualization techniques. All the tools have a high level of interaction and are dynamically linked via highlighting, selection and brushing. The system has been gradually developed over a number of years and was used by the authors for exploration problems in such various disciplines as social geography, forestry, meteorology, seismology, crime and environment (G. Andrienko *et al.*, 2003a, 2003b; N. Andrienko *et al.*, 2003; N. Andrienko and G. Andrienko, 2006b). More information is available from the authors' homepage, http://www.ais.fraunhofer.de/and.

GeoDa or Geodata analysis software (Anselin *et al.*, 2004) is an interactive environment that provides a user-friendly and graphical introduction to spatial analysis for non-experts, from simple mapping to more advanced exploratory data analysis. The functionality ranges from spatial data manipulation, data transformation, mapping – including choropleth maps, cartograms and map animation, to statistical graphic tools for exploratory data analysis and the visualization of various spatial statistical characteristics, such as spatial autocorrelation and spatial regression. GeoDa is also free and is downloadable from http://www.geoda.uiuc.edu.

4.4. SOME GEOVISUALIZATION EXAMPLES

This section attempts to introduce the reader to several examples of geovisualization. Due to the constraints of this publishing medium (a printed book) we are unfortunately limited to present examples of these visualizations as black and white images. Colour, animation and interactivity, which are all integral parts of geovisualization, are not supported. The reader is therefore encouraged to follow up the references in this section for a more realistic illustration of what geovisualization can do.

An alternative visualization to traditional maps are cartograms, which distort the display space according to a specific attribute (Tobler, 2004). The objective of the distortion

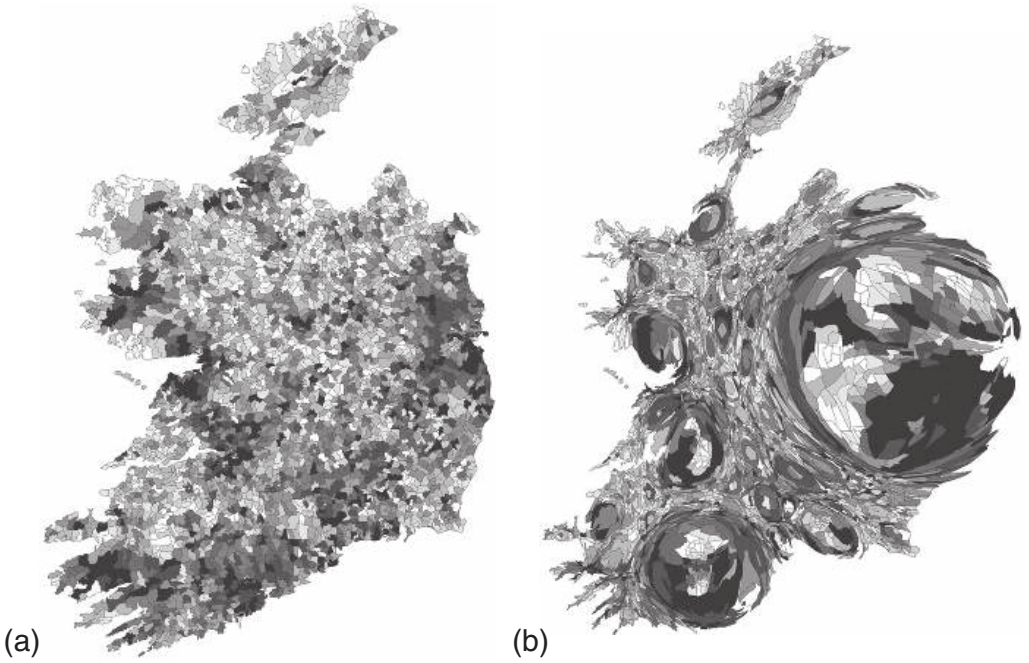


Figure 4.2 An example of (a) a choropleth map of the proportion of residents in Social Class 1 in the Electoral Divisions (EDs) of the Republic of Ireland and (b) an area cartogram of the same phenomenon where the areas of EDs are scaled according to the population size. Dark colour indicates a high proportion and light colour a low proportion of residents of Social Class 1 (i.e., 'rich' residents) in a particular ED. On the cartogram in (b), the pattern in Dublin can be clearly seen: the South side has the largest proportion of rich people, and there are three areas in the north-east, north-west and south-west of the city where the proportion of the rich is the lowest. This pattern can be barely recognized in the choropleth map in (a), but the cartogram distortion makes it very eye-catching.

is to reveal patterns that are not apparent in the conventional map. Typical examples are linear cartograms, where the space (usually represented as a spatial network) is distorted according to some distance other than the geometric one, for example travel time. Such cartograms are commonly used to represent public transit systems in larger cities – any subway map or a map of commuter rail services is typically a linear cartogram. Another principle is to stretch the space continuously according to the distribution of values of some attribute, but to preserve the general shape and adjacency of polygons to produce an area cartogram (Tobler, 2004). Figure 4.2 shows an example of a choropleth map (Figure 4.2a) versus the area cartogram of the same phenomenon (Figure 4.2b). The figure shows two displays of the spatial variation in the proportion of residents in Social Class 1 in the Electoral Divisions (EDs) of the Republic of Ireland in 2002. Residents in Social Class 1 are the most affluent. The map on the left (Figure 4.2a) is drawn using the Irish National Grid projection in which the polygons are scaled in proportion to their land area. It is difficult to see what spatial variations there are in the main urban centres, and the boundaries are visually intrusive. The areas in the cartogram on the right (Figure 4.2b) have been redrawn so that their areas are in proportion to their population – this is an area cartogram or a density equalized projection.

The urban centres (starting from Dublin as the largest distorted area located on the east coast, followed by Waterford, Cork, Limerick, Galway and Sligo in clockwise-order along the coastline) become dominant in the display, and we can easily see the spatial variation in the proportions of affluent residents across the country – a spatial pattern which was not obvious in the traditional choropleth map (Figure 4.2a). The cartogram in this figure was produced using the algorithm and software by Gastner and Newman (2004).

Another example of a fairly common geovisualization are 3D displays. These project the three locational dimensions onto a 2D display using a set of perceptual depth cues to reinforce this projection, such as perspective, occlusion and parallax motion (Ware and Plumlee, 2005). Here we present some examples of 3D geovisualizations, but only in the context of visual knowledge discovery from spatial data. The reader can explore other issues, such as the use of 3D georepresentations in Virtual Reality and Virtual Environments, elsewhere (two starting points for that would be Fisher and Unwin (2002) and Bodum (2005)).

One of the most common methods of representing multivariate geospatial data in three-dimensions for knowledge discovery are surfaces, which are sometimes also referred to as 2.5D representations when displayed on the screen, as they are not literally three dimensional. A general approach to produce a surface is to map the two basic geographic dimensions, longitude and latitude, to the x and y-axis respectively and show the variable of interest on the z-axis. Over this surface some other type of geographic information can be draped to provide texture: a thematic map or a satellite image. Traditionally the attribute mapped to the z-axis represents the third dimension in the real world, such as the elevation above the sea level or the depth of the sea bottom (Kreuseler, 2000). In some cases, the attribute mapped to the z-axis represents time and instead of the surfaces, trajectories of movements of objects are projected through the display space. This type of geovisualization is very common in time-geography (Kraak and Koussoulakou, 2004) and in transportation studies (Kwan, 2000, 2004). In the third type of the surfaces the z-axis attribute represents neither a real geographic dimension nor time, but some other variable of interest, such as the population density, the temperature, the density of human activity or travel (Kwan, 2000), or in geosciences the magnetic variation or the kriging variance (Carr, 2002). Figure 4.3 shows a surface where the z-axis represents the concentration of radon in the groundwater. The surface is covered with two maps of the area, one showing the bedrock and another one showing locations of fractures (Demšar and Skeppström, 2005). Visual exploration of this representation clearly indicates that high values of radon in this area (the highest peaks of the surface) occur only on a particular type of bedrock, which is shown with medium grey shade.

Figure 4.4 shows a screenshot of a visual exploratory system built using GeoVISTA Studio. The system consists of a multiform bivariate matrix, a geoMap and a parallel coordinates plot (PCP), which all share the same colour scheme (except the spaceFills in the matrix). This principle of colouring the graphical entities belonging to the same data element with the same colour in all visualizations is called visual brushing. All visualizations are also connected by interactive selection and brushing through mouse-over operation – interaction, which unfortunately cannot be adequately presented through a simple screenshot image, but is essential for successful data exploration.

The parallel coordinates plot (PCP) maps the n dimensional space onto the two display

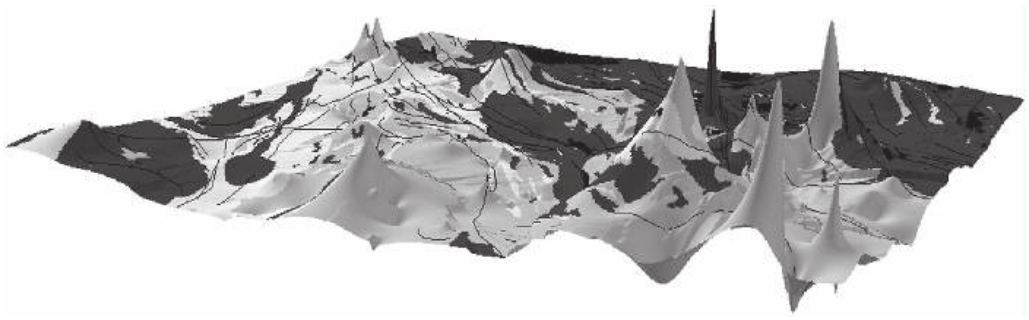


Figure 4.3 A bedrock-fractures-radon visualization as a 2.5D surface. The height of the surface represents the concentration of radon in the groundwater. Most of the peaks which indicate high radon values are located on a certain type of bedrock, shown in medium grey shade on the geological map, which is draped over the radon surface.

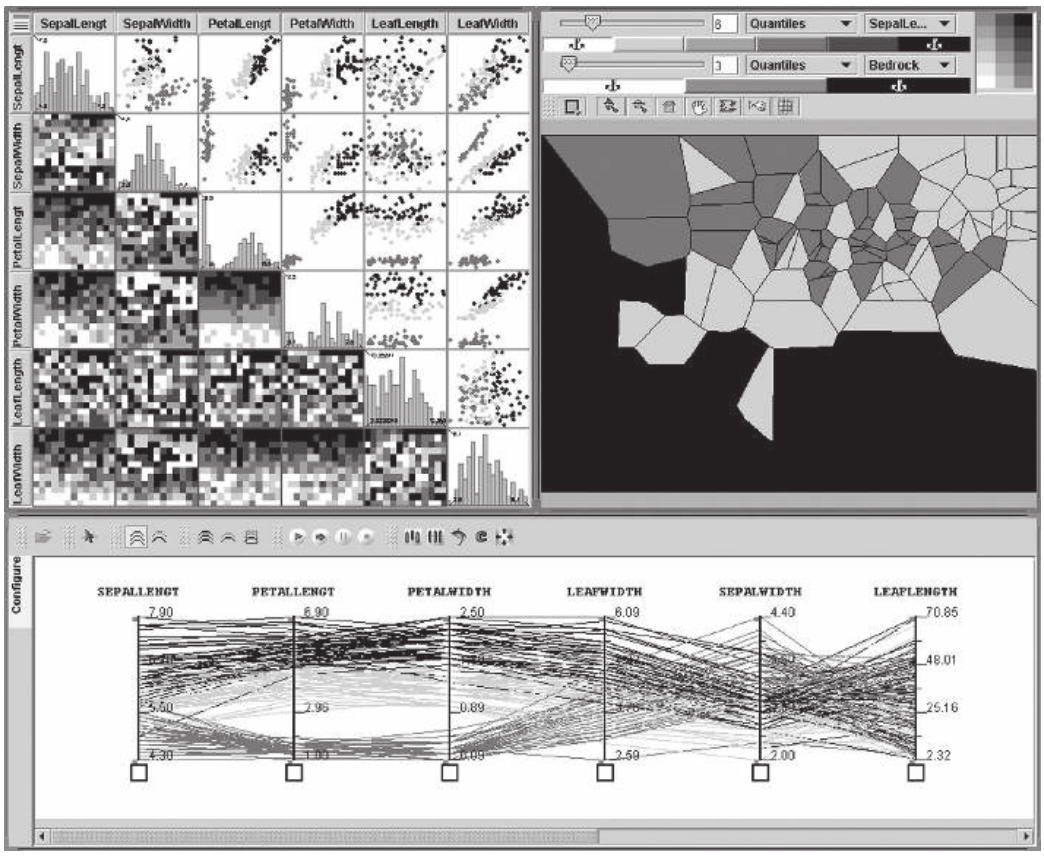


Figure 4.4 A GeoVISTA-based system displaying a synthetic spatial dataset (Demšar, 2006) based on the famous iris data (Fisher, 1936).

dimensions by using n equidistant parallel vertical axes produced (Inselberg, 2002). The axes correspond to the dimensions and are linearly scaled from the minimum to the maximum value of the corresponding dimension. Each data item is then drawn as a polygonal line intersecting each of the axes at the point which corresponds to the data value. Figure 4.4 shows an example.

A multiform bivariate matrix in Figure 4.4 is a generalization of a scatterplot matrix and consists of univariate visualizations – histograms on the diagonal and bivariate visualizations at other positions in the matrix (MacEachren *et al*., 2003). In the matrix in Figure 4.4, scatterplots of each corresponding pair of variables are located above the diagonal and spacefills below the diagonal. Spacefills are dense pixel bivariate visualizations, where each data element is represented by a grid square. The first of the two variables defines the colour of each square, ranging from light for low values to dark for high values of the variable. The second variable defines the order of the squares inside the rectangular display: the cell with the lowest value of this variable is situated in the bottom-left corner, from where the cells then proceed along a scan line towards the cell with the highest value in the top-right corner. If the attribute defining the colour of the cells is correlated with the attribute defining the order of the cells, there is a relatively smooth transition from the lightest to the darkest colour from bottom to top (or from top to bottom). If the correlation is weaker, the pattern appears more scattered and if there's no correlation, all that can be seen is a random distribution of the cells in the display (Gahegan *et al*., 2002).

The geographic visualization in Figure 4.4 is the GeoVISTA's geoMap, which is a bivariate choropleth map, whose colour scheme is defined by a cross-tabulation of the two display attributes (Gahegan *et al*., 2002) or can alternatively be inherited from other visualizations through visual brushing as mentioned above.

Notice the linear separability of the three clusters in several scatterplots in the matrix: the dark grey cluster can be linearly separated from the light grey and the black one, which are mixed in several displays. The same clusters are clearly separated in the petal length and petal width variables in the PCP, but not in other dimensions. There is also a distinct spatial pattern of the three clusters in the map, where the black cluster is separated from the other two. The spacefills in the matrix indicate correlations between several pairs of variables – the strongest one seems to be between petal length and petal width, where the pattern in the relevant spacefill proceeds from white to black in a relatively smooth manner. Other spacefills display a completely random distribution of cells, such as for example the one in the row belonging to the leaf length variable and the column belonging to the sepal width variable – this indicates that there is probably no correlation between the corresponding two variables (which can be confirmed by looking at the appropriate scatterplot).

The visualization in Figure 4.5 shows another type of GeoVISTA matrix – a so-called fixed-row matrix (MacEachren *et al*., 2003), where a selected row variable (in this case sepal length) is mapped against each of the column variables using a different bivariate visualization for each row. In this particular example the first row contains bivariate choropleth maps of sepal length vs. all other five variables, while the second row shows scatterplots of the same pairs of variables. Such matrices can either be used separately or form a part of a larger exploratory system as one of the multiple linked views.

A popular recent approach to combine visual and computational data exploration

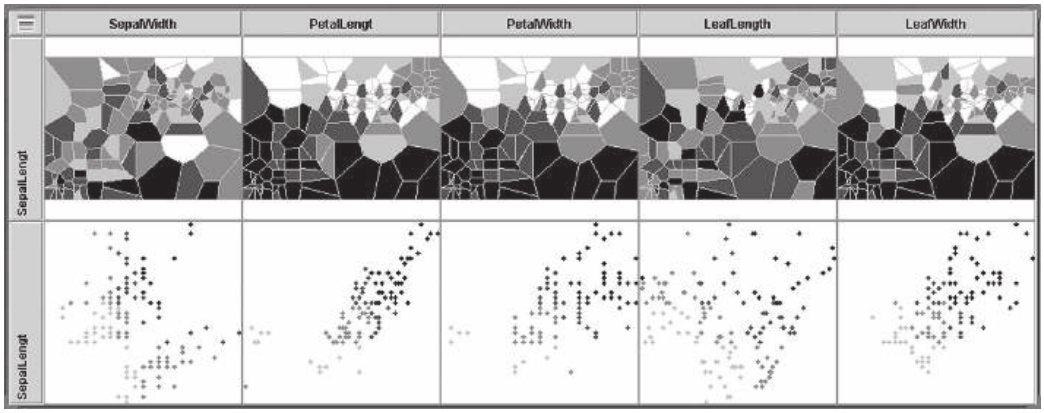


Figure 4.5 A fixed row matrix of bivariate visualizations, again a component from GeoVISTA Studio and displaying the same data as in the previous figure.

is to use a Self-Organizing Map (SOM) as the computational method together with other spatial or non-spatial visualizations. The SOM is an unsupervised neural network which projects multidimensional data onto a two-dimensional lattice of cells while preserving the topology and the probability density of the input data space. This means that similar data vectors are mapped to the same neuron cell or to the neighbour cells in the two-dimensional output map, which makes it useful as a knowledge discovery tool (Kohonen, 1997; Silipo, 2003). The SOM has been recently used for knowledge discovery in a number of spatial and spatio-temporal applications (Takatsuka, 2001; Gahegan *et al*., 2002; Jiang and Harrie, 2004; Koua and Kraak, 2004; Guo *et al*., 2005; Skupin and Hagelman, 2005; Demšar, 2007b; Lacayo and Skupin, 2007; Špatenková *et al*., 2007).

One reason for its popularity for integration into a visual system is that SOM produces a very visualizable result due to its two-dimensionality. Vesanto (1999) lists a number of possible visualizations. An example that we present here are the component planes (Figure 4.6), where each plane is a SOM lattice. In the D-matrix the grey shade represents how similar each cell is to its neighbours. Dark areas in the D-matrix consist of very similar cells and therefore represent clusters. Light areas in contrast indicate borders between clusters. In each other plane (except in the D-matrix) the grey shade of each cell indicates the average value of a particular attribute calculated from values of all data elements assigned to that cell. Relationships between attributes are discovered by comparing the values in the same area of the SOM lattice in different planes.

While geographic location is the core concept of geographic information science, the visual models and methods that geographers and cartographers have been using for a long time can also be applied to the representation of objects, phenomena or processes with spatial characteristics and behaviour in abstract spaces. The use of geographic and cartographic concepts to represent data which are not inherently spatial is called spatialization (Skupin and Fabrikant, 2003). The aim is to systematically transform highly multidimensional data into spatial representations in lower-dimensional

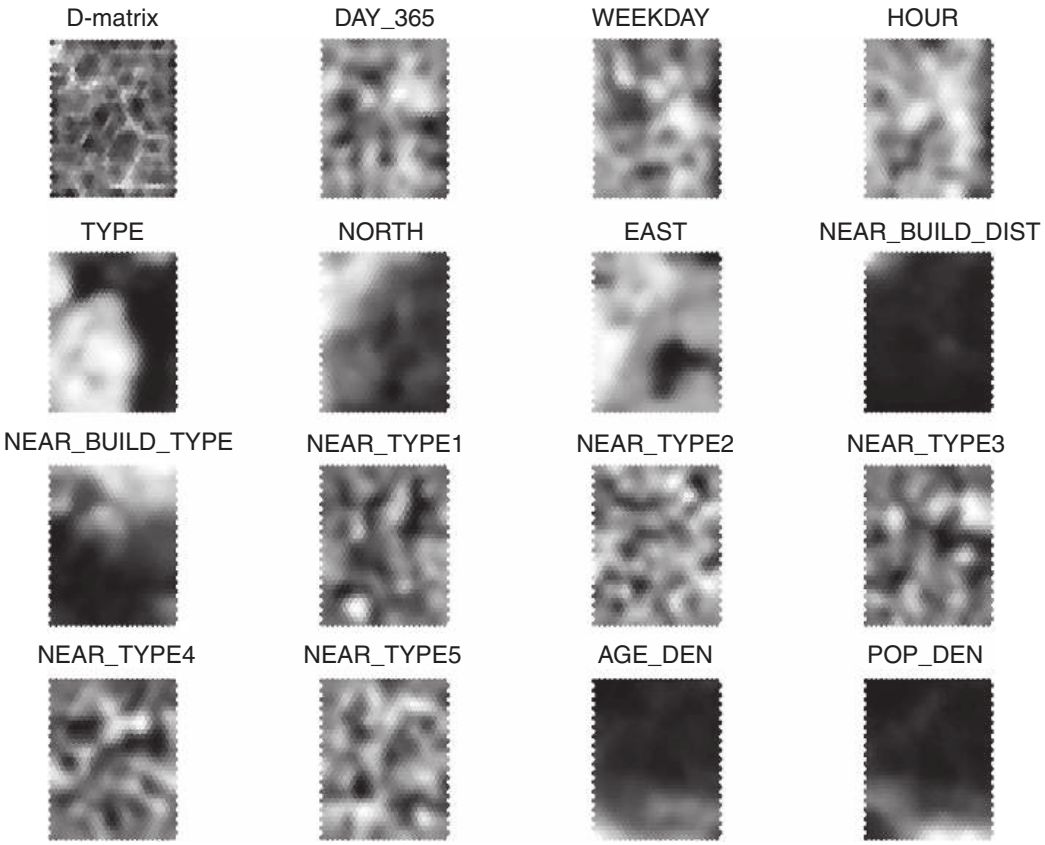


Figure 4.6 Visually discovering relationships between the spatio-temporal attributes from the SOM component planes visualization. The image was produced using a spatio-temporal data set of emergency response data (Špatenková *et al.*, 2007) and the SOM toolbox for Matlab.

abstract spaces with the goal to facilitate data exploration and knowledge construction (Fabrikant *et al.*, 2002). Examples of spatializations are spatial representations of scientific co-authorship networks (Newman, 2004), protein–receptor interaction networks in medicine, genealogies and citation networks (Batagelj and Mrvar, 2003).

Figure 4.7 shows an example of a spatialization: two different visualization of a citation network of the GeoVISTA Studio related papers from the reference list of this chapter. In Figure 4.7(a), arrows indicate citations, i.e., the paper that the arrow points from cites the paper which the arrow points to. The size of the vertices in Figure 4.7(a) shows the in-degree of each publication, i.e., how many other publications cite it. The direction of the arrows is ignored in Figure 4.7(b) and the size of vertices in this picture represents the betweenness centrality (from social network analysis (Freeman, 1979)) which measures the importance of each vertex. Note that vertex representing the paper by MacEachren *et al.*, (1999) has a high in-degree as well as high betweenness because many other papers cite it, while the relatively large betweenness of Guo *et al.*, (2005) is a result of the fact that this paper cites many other papers (even though it's never cited itself and has a low in-degree – compare with Figure 4.7(a).

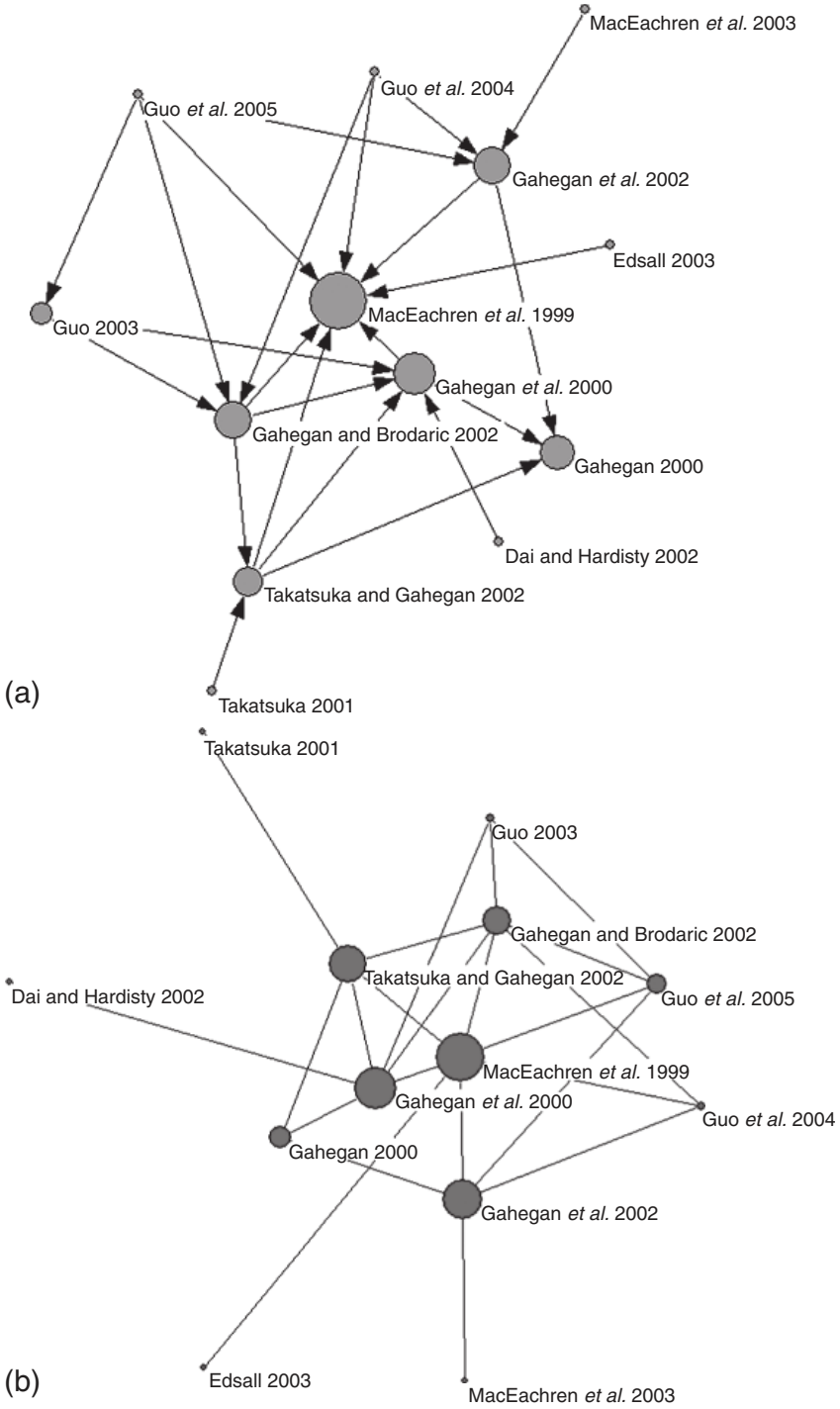


Figure 4.7 A spatialization of a non-spatial phenomenon: the citation network of GeoVISTA publications from the reference list of this chapter as (a) a directed and (b) an undirected graph. The size of the vertices in (a) indicates the in-degree of each vertex and in (b) the relative importance of the vertex in the network, measured by the betweenness centrality (Freeman 1979). Spatial positions of the vertices were calculated according to their in-degree in (a) and betweenness in (b). First the vertex with the highest in-degree/ betweenness was placed in a central position. Then the vertices with every lower in-degree/betweenness value were iteratively placed in the nearest proximity while at the same time minimizing the energy of the edges according to an energy-preserving graph-drawing algorithm. The spatialization was produced using the Pajek software for analysis and visualization of large networks (Batagelj and Mrvar, 2007).

4.5. THE FUTURE: FROM GEOVISUALIZATION TO GEOVISUAL ANALYTICS

A recent new research area that has emerged in information visualization is Visual Analytics, which is defined as 'the science of analytical reasoning supported by highly interactive visual interfaces' (NVAC, 2005). Visual Analytics tools are used for synthesizing information into knowledge, derive insight from massive, dynamic and conflicting data, discover the unexpected, and provide and communicate timely and understandable assessments. The recent research agenda (NVAC, 2005) for Visual Analytics identifies four major research areas:

- *the science of analytical reasoning*, which provides the reasoning framework to serve as the basis for visual technologies for data analysis;
- *visual representations and interaction techniques*, which provide the mechanism to see and understand large volumes of data;
- *data representations and transformations* appropriate to the analytical task that correctly convey the important content of large, complex and dynamic data sets; and
- *production, presentation and dissemination of results*, where the goal is to reduce the time to present the results to the audience in a more effective communication manner.

While the primary goal of developing the Visual Analytics research agenda was for US security purposes, the challenges listed above will have impact on any field of scientific research where understanding complex and dynamic data is important. Geovisualization is no exception. In this context, a subfield of Visual Analytics relevant for geovisualization is Geovisual Analytics, which integrates perspectives from Visual Analytics and geographic information science for analysis of spatio-temporal data. Geovisual Analytics is defined as 'the science of analytical reasoning and decision-making with geospatial information, facilitated by interactive visual interfaces, computational methods, and knowledge construction, representation, and management strategies' (G. Andrienko et al., 2007). The four research areas defined in the research agenda of Visual Analytics are also applicable for Geovisual Analytics and the tools developed in the Geovisual Analytics context are starting to be of increasing importance for various applications fields, such as crisis management (Tomaszewski *et al*., 2007) and spatial decision support (G. Andrienko *et al*., 2007).

Apart from the above-mentioned research areas which are common to Geovisual and Visual Analytics, there are many questions in geovisualization which are related to the particularities of geospatial data and phenomena and thereby inherently characteristic to our discipline. Here we present a short selection of topics, although the list is far from exhaustive – the reader is encouraged to turn to the following sources for a more comprehensive review and a research agenda (MacEachren and Kraak, 2001; Dykes *et al*., 2005; Andrienko *et al*., 2007; ICA, 2008).

Support for collaborative group work and distributed geovisualization: with the advent of ubiquitous computing, many potential application areas for Geovisual Analytics will require actions distributed over geographic space and time. Such tasks will include exploration of various spatio-temporal distributions of complex data and events, and will be physically performed in more than one location. Geovisual Analytics tools of the future should be able to support such exploration. This is particularly important in situations where the users do not have time to consider all possible solutions to their problems or cannot afford to search for an

optimal one, such as in, for example, crisis management. In order to deal efficiently with time pressure and stress in such situations, Geovisual Analytics tools need to provide support for a shared collaborative work during a process where key parameters change quickly, such as, for example, for spatial decision support in emergencies. Open issues here range from developing distributed system architectures to intelligent solutions that support fast knowledge capture, rational reasoning and time-critical spatial decision making.

A related topic is *mobile geovisualization and location-based visual exploration*. Present technological advances in mobile communications and the ubiquity of various mobile devices (mobile phones, PDAs, BlackBerries, etc.) are likely to change the way people use information systems – and this includes tools for geovisualization and Geovisual Analytics. The emerging location-based personalization raises not only technical questions such as how to perform on-the-fly location-based computation or how to display as much information as possible without losing the clarity on a small display of most of today's mobile devices, but also conceptual issues, for example the use of individually personalized dynamic egocentric maps (Meng, 2004; Meng, 2005) instead of the traditional geocentric visualizations that remain static for a longer period and aim to communicate geographic information to a variety of users.

Finally, one of the recurrent topics in visualization research are *cognitive and perceptual questions and evaluation of the tools*. Not only are visualization tools difficult to evaluate objectively, the results of such evaluations might not be replicable nor generalizable and are in general difficult to interpret (Plaisant, 2004). Additionally, there exist some evidence that there may be fundamental differences between information visualization and geovisualization (Tobón, 2005), which implies that the cognitive processes that must be supported in geovisualization are different and possibly more complex than when non-spatial data are investigated. Experiments have also shown that there exist significant interpersonal differences in the way people visually explore spatial data, how they interpret what they see and what exploration strategies they form (G. Andrienko *et al.*, 2006; Demšar, 2006, 2007a). All this suggests that visual data exploration is inherently complex. What can be done to alleviate the complexity? How is the ability to use the tools related to users' background and experience? These are just some questions to be considered. In order to resolve them, work on technological advances should be combined with work on human spatial cognition to fully reveal the potential of visual representations to support spatial analytical reasoning, spatial problem solving and spatial decision making.

ACKNOWLEDGEMENTS

The author would like to thank Mark Gahegan from The University of Auckland for kindly consenting to read the first draft of this chapter and providing helpful comments and suggestions. Thanks goes also to Olga Špatenková from Helsinki University of Technology and Martin Charlton from the National Centre of Geocomputation, National University of Ireland, Maynooth, who prepared the illustrations showing the SOM component planes and the cartograms respectively. Finally, research presented in this paper was supported by a grant to the National Centre for Geocomputation by Science Foundation Ireland (03/RP1/1382) and by a Strategic Research Cluster grant (07/SRC1/1168) from Science Foundation Ireland under the national Development Plan. The author gratefully acknowledges this support.

REFERENCES

Ahonen-Rainio, P. (2005). *Visualization of geospatial metadata for selecting geographic data sets*. PhD thesis. Helsinki University of Technology, Espoo, Finland.

Andrienko, N., Andrienko, G., Voss, H., Bernardo, F., Hipolito, J. and Kretchmer, U. (2002). Testing the usability of interactive maps in CommonGIS. *Cartography and Geographic Information Science*, **29**(4): 325–342.

Andrienko, G., Andrienko, N. and Voss, H. (2003a). GIS for everyone: the CommonGIS project and beyond. In: Peterson, M. (ed.), *Maps and the Internet*, pp. 131–146. Elsevier Science.

Andrienko, G., Andrienko, N. and Gitis, V. (2003b). Interactive maps for visual exploration of grid and vector geodata. *ISPRS Journal of Photogrammetry and Remote Sensing*, **57**: 380–389.

Andrienko, G., Andrienko, N., Fischer, R., Mues, V. and Schuck, A. (2006). Reactions to geovisualization: an experience from a European project. *International Journal of Geographic Information Science*, **20**(10): 1149–1171.

Andrienko, G., Andrienko, N., Jankowski, P., Keim, D., Kraak, M.-J., MacEachren, A.M. and Wrobel, S. (2007). Geovisual analytics for spatial decision support: setting the research agenda. *International Journal of Geographic Information Science*, **21**(8): 839–857.

Andrienko, N., Andrienko, G. and Gatalsky, P. (2003). Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages and Computing*, **14**: 503–541.

Andrienko, N. and Andrienko, G. (2006a). The complexity challenge to creating useful and usable geovisualization tools. In: *Proceedings of GIScience 2006*, Münster, Germany.

Andrienko, N. and Andrienko, G. (2006b). *Exploratory Analysis of Spatial and Temporal Data*. Berlin–Heidelberg: Springer Verlag.

Anselin, L., Syabri, I. and Youngihn, K. (2004). GeoDa: An introduction to spatial data analysis. *Geographical Analysis*, **38**: 5–22.

Batagelj, V. and Mrvar, A. (2003). Pajek – Analysis and visualization of large networks. In: Jünger, M. and Mutzel, P. (eds), *Graph Drawing Software*, pp. 77–103. Berlin–Heidelberg: Springer Verlag.

Batagelj, V. and Mrvar, A. (2007). *Pajek – Program for Analysis and Visualization of Large Networks*, version 1.20., 25 June 2007, available for download at: http://vlado.fmf.uni-lj.si/pub/networks/pajek/ (last accessed 10 July 2007).

Bederson, B.B., Shneiderman, B. and Wattenberg, M. (2002). Ordered and quantum treemaps: making effective use of 2D space to display hierarchies. *ACM Transactions on Graphics*, **21**(4): 833–854.

Bertin, J. (1983). *The Semiology of Graphics*, University of Wisconsin Press, Madison, Winsconsin. Translation of: Bertin, J. (1967). *Semiologie Graphique*, Paris: Mouton.

Bodum, L. (2005). Modelling virtual environments for geovisualization: a focus on representation. In: Dykes, J., MacEachren, A.M. and Kraak, M.-J. (eds), *Exploring Geovisualization*, pp. 389–402. Amsterdam: Elsevier.

Card, S.K., Mackinlay, J. and Shneiderman, B. (eds) (1999). *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufmann Publishers.

Carr, J.R. (2002). *Data Visualization in the Geological Sciences*. Upper Saddle River: Prentice Hall.

Chernoff, H. (1973). The use of faces to represent points in *k*-dimensional space graphically. *Journal of the American Statistical Association*, **68**(342): 361–368.

Dai, X. and Hardisty, F. (2002). Conditioned and manipulable matrix for visual exploration. In: *Proceedings of the National Conference for Digital Government Research 2002.*

Demšar, U. and Skeppström, K. (2005). Use of GIS and 3D visualisation to investigate radon problem in groundwater. In: *Proceedings of the 10th Scandinavian Research Conference on Geographic Information Science, ScanGIS2005*, Stockholm, Sweden.

Demšar, U. (2006). Investigating visual exploration of geospatial data: an exploratory usability experiment for visual data mining. *Computers, Environment and Urban Systems* (accepted for a special issue). Short version presented at the *1st ICA Workshop on Geospatial Analysis and Modelling*, Vienna, July, 2006.

Demšar, U. (2007a). Combining formal and exploratory methods for evaluation of an exploratory geovisualization application in a low-cost usability experiment. *Cartography and Geographic Information Science*, **34**(1): 29–45.

Demšar, U. (2007b). Knowledge discovery in environmental sciences: visual and automatic data mining for radon problems in groundwater. *Transactions in GIS*, **11**(2): 255–281.

DiBiase, D. (1990). Visualization in the Earth Sciences. *Earth and Mineral Sciences*, **59**(2): 13–18.

Dykes, J.A. and Mountain, D.M. (2003). Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and applications. *Computational Statistics and Data Analysis*, **43**(4): 581–603.

Dykes, J.A. (2005). Facilitating interaction for geovisualization. In: Dykes, J., MacEachren, A.M. and Kraak, M.-J. (eds), *Exploring Geovisualization*, pp. 265–292. Amsterdam: Elsevier.

Dykes, J.A., MacEachren, A.M. and Kraak, M.-J. (2005). Advancing geovisualization. In: Dykes, J., MacEachren, A.M. and Kraak, M.-J. (eds), *Exploring Geovisualization*, pp. 693–704. Amsterdam: Elsevier.

Edsall, R.M. (2003). The parallel coordinate plot in action: design and use for geographic visualization. *Computational Statistics and Data Analysis*, **43**: 605–619.

van Elzakker, C.P.J.M. (2004). *The use of maps in the exploration of geographic data*. PhD thesis. Utrecht University, Utrecht, The Netherlands.

Fabrikant, S.A., Skupin, A. and Couclelis, H. (2002). *Spatialization: Spatial Metahphors and Methods for Handling Non-Spatial Data*. Web document (last accessed 12 July 2007), http://www.geog.ucsb.edu/~sara/html/research/ucgis/spatialization_ucsb.pdf

Fayyad, U., Grinstein, G.G. and Wierse, A. (eds) (2002). *Information Visualization in Data Mining and Knowledge Discovery*. San Francisco: Morgan Kaufmann Publishers.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2): 179–188. In: Fisher, R.A. (1950). *Contributions to Mathematical Statistics*. New York: John Wiley & Sons.

Fisher, P.F. and Unwin, D.J. (eds) (2002). *Virtual Reality in Geography*. London: Taylor and Francis.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2000). Exploring Spatial Data Visually, Chapter 4 in *Quantitative Geography – Perspectives on Spatial Data Analysis*, 65–92. Sage Publications. London, UK.

Freeman, L.C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, **1**: 215–239.

Fuhrmann, S., Ahonen-Rainio, P., Edsall, R.M., Fabrikant, S.I., Koua, E.L., Tobón, C., Ware, C. and Wilson, S. (2005). Making useful and useable geovisualization: design and evaluation issues. In: Dykes, J., MacEachren, A.M. and Kraak, M.-J. (eds). *Exploring Geovisualization*, pp. 553–566. Amsterdam: Elsevier.

Gahegan, M. (2000). On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, **32**(2): 113–139.

Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F. (2000). GeoVISTA Studio: a geocomputational workbench. In: *Proceedings of Geocomputation 2000.* Univeristy of Greenwich, UK.

Gahegan, M. and Brodaric, B. (2002). Computational and visual support for geographical knowledge construction: filling in the gaps between exploration and explanation. In: *Proceedings of the Spatial Data Handling 2002.* Ottawa, Canada.

Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F. (2002). Introducing Geo-VISTA Studio: an integrated suite of visualization and computational methods for exploration and knowledge construction in geography. *Computers, Environment and Urban Systems*, **26**: 267–292.

Gahegan, M. (2005). Beyond tools: visual support for the entire process of GIScience. In: Dykes, J., MacEachren, A.M. and Kraak, M.-J. (eds), *Exploring Geovisualization*, pp. 83–99. Amsterdam: Elsevier.

Gastner, M.T. and Newman, M.E.J. (2004). Diffusion-based method for producing density-equalizing maps. In: *Proceedings of the National Academy of Sciences*, **101**(20): 7499–7504.

Griffin, A. (2004). *Understanding how scientists use datadisplay devices for interactive visual computing with geographical models*. PhD thesis. The Pennsylvania State University, Pennsylvania, USA.

Guo, D. (2003). Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, **2003**(2): 232–246.

Guo, D., Gahegan, M. and MacEachren, A.M. (2004). An Integrated Environment for High-dimensional Geographic Data Mining. In: *Proceedings of GIScience 2004.* University of Maryland, USA.

Guo, D., Gahegan, M., MacEachren, A.M. and Zhou, B. (2005). Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Cartography and Geographic Information Science*, **32**(2): 113–132.

Haklay, M. and Tobón, C. (2003). Usability evaluation and PPGIS: Towards a user-centred design approach. *International Journal of Geographical Information Science*, **17**(6): 577–592.

International Cartographic Association (ICA) (2008). *ICA Commission on GeoVisualization*, website of the commission, http://geoanalytics.net/ica (last accessed: 11 September 2008).

Inselberg, A. (2002). Visualization and data mining of high-dimensional data. *Chemometrics and Intelligent Laboratory Systems*, **60**: 147–159.

Jiang, B. and Harrie, L. (2004). Selection of streets from a network using self-organizing maps. *Transactions in GIS*, **8**: 335–350.

Keim, D.A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, **7**(1): 100–107.

Keim, D.A. and Ward, M. (2003). Visualization. In: Berthold, M. and Hand, D.J. (eds), *Intelligent Data Analysis*, 2nd edn, pp. 403–428. Berlin–Heidelberg: Springer Verlag.

Kohonen, T. (1997). *Self-Organizing Maps*, 2nd edn. Berlin–Heidelberg: Springer Verlag.

Koua, E.L. and Kraak, M.-J. (2004). Alternative visualization of large geospatial datasets. *The Cartographic Journal*, **41**: 217–228.

Koua, E.L. (2005). *Computational and visual support for exploratory geovisualization and knowledge construction*. PhD thesis. Utrecht University, Utrecht, The Netherlands.

Kraak, M.-J. and Koussoulakou, A. (2004). A visualization environment for the space–time cube. In: Fisher, P.F. (ed.) *Developments in Spatial Data Handling, 11th International Symposium on Spatial Data Handling*, pp. 189–200. Berlin–Heidelberg: Springer Verlag.

Kreuseler, M. (2000). Visualization of geographically related multidimensional data in virtual 3D scenes. *Computers & Geosciences*, **26**: 101–108.

Kreuseler, M. and Schumann, H. (2002). A flexible approach for visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, **8**(1): 39–51.

Kwan, M.P. (2000). Interactive geovisualization of activity-travel patterns using three dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research Part C*, **8**: 185–203.

Kwan, M.P. (2004). GIS methods in time-geographic research: geocomputation and geovisualization of human activity patterns. *Geografiska Annaler B*, **86**: 267–280.

Lacayo, M. and Skupin, A. (2007). A GIS-based module for training and visualization of self-organizing maps. Working paper, accepted to the *Workshop of the ICA Commission on Visualization and Virtual Environments, 'From Geovisualization to Geovisual Analytics'*, Helsinki, August 2007.

MacEachren, A.M., Wachowitz, M., Edsall, R., Haug, D. and Masters, R. (1999). Constructing knowledge from multivariate spatio-temporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographic Information Science*, **13**(4): 311–334.

MacEachren, A.M. and Kraak, M.-J. (2001). Research challenges in geovisualization. *Cartography and Geographic Information Science*, **28**(1): 3–12.

MacEachren, A., Dai, X., Hardisty, F., Guo, D. and Lengerich, G. (2003). Exploring High-D Spaces with Multiform Matrices and Small Multiples. In: *Proceedings of the International Symposium on Information Visualization 2003.* Seattle, Washington, USA.

MacEachren, A.M., Gahegan, M., Pike, W., Brewer, I., Cai, G. and Lengerich, E. (2004). Geovisualization for knowledge construction and decision support. *IEEE Computer Graphics and Applications*, **24**(1): 13–17.

McCormick, B.H., DeFanti, T.A. and Brown, M.D. (1987). Visualization in Scientific Computing – A Synopsis. *IEEE Computer Graphics and Applications*, **7**(7): 61–70.

Meng, L. (2004). About egocentric geovisualization. In: *Proceedings of the 12th International Conference on Geoinformatics*, Gävle, Sweden, June 2004.

Meng, L. (2005). Egocentric design of map-based mobile services. *The Cartographic Journal*, **42**(1): 5–13.

Miller, H.J. and Han, J. (eds) (2001). *Geographic Data Mining and Knowledge Discovery*. London and New York: Taylor & Francis.

Müller-Hannemann, M. (2001). Drawing trees, series–parallel digraphs and lattices. In: Kaufmann, M. and

Wagner, D. (eds), *Drawing Graphs – Methods and Models*. Lecture Notes in Computer Science, 2025: 46–70. Berlin–Heidelberg: Springer Verlag.

National Visualization and Analytics Center (NVAC) (2005). *Illuminating the Path: Creating the R&D Agenda for Visual Analytics*. Available at: http://nvac.pnl.gov/agenda.stm (last accessed 17 July 2007).

Newman, M.E.J. (2004). Who is the best connected scientist? A study of scientic coauthorship networks. In: Ben-Naim, E., Frauenfelder, H. and Toroczkai, Z. (eds), *Complex Networks*, pp. 337–370. Berlin–Heidelberg: Springer Verlag.

Nielsen, J. (1993). *Usability Engineering*. San Francisco: Morgan Kaufmann Publishers.

Plaisant, C. (2004). The challenge of information visualization evaluation. In: *Proceedings of the IEEE Conference on Advanced Visual Interfaces AVI'04*, Gallipoli, Italy.

Preece, J. Rogers, Y. and Sharp, H. (2002). *Interaction Design: Beyond Human–Computer Interaction*. New York: John Wiley and Sons.

Roberts, J.C. (2005). Exploratory visualization with multiple linked views. In: Dykes, J., MacEachren, A.M. and Kraak, M.-J. (eds), *Exploring Geovisualization*, pp. 159–180. Amsterdam: Elsevier.

Robinson, A.C., Chen, J., Lengerich, E.J., Meyer, H.G. and MacEachren, A.M. (2005). Combining usability techniques to design geovisualization tools for epidemiology. In: *Proceedings of Auto-Carto 2005*, Las Vegas, USA.

Seo, J. and Shneiderman, B. (2002). Interactively Exploring Hierarchical Clustering Results. *IEEE Computer*, **35**(7): 80–86.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualization. *IEEE Proceedings of Visual Languages*. Boulder, Colorado, USA.

Shneiderman, B. (2001). Inventing discovery tools: combining information visualization with data mining. In: *Proceedings of the 12th International Conference on Algorithmic Learning Theory*. Lecture Notes in Computer Science, 2226: 17–28. Berlin–Heidelberg: Springer Verlag.

Silipo, R. (2003). Neural networks. In: Berthold, M. and Hand, D.J. (eds), *Intelligent Data Analysis*, 2nd edn, pp. 269–32. Berlin–Heidelberg: Springer Verlag.

Skupin, A. and Fabrikant, S.A. (2003). Spatialization methods: a cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science*, **30**(2): 99–119.

Skupin, A. and Hagelman, R. (2005). Visualizing demographic trajectories with self-organizing maps. *Geoinformatica*, **9**(2): 159–179.

Slocum, T.A., Cliburn, D.C., Feddema, J.J. and Miller, J.R. (2003). Evaluating the usability of a tool for visualizing the uncertainty of the future global water balance. *Cartography and Geographic Information Science*, **30**(4): 299–317.

Špatenková, O., Demšar, U. and Krisp, J.M. (2007). Self-organising maps for exploration of spatio-temporal emergency response data. In: *Proceedings of Geocomputation 2007*. Maynooth, Ireland.

Stasko, J. and Zhang, E. (2000). Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In: *Proceedings of InfoVis2000, IEEE Symposium on Information Visualization*, pp. 57–68. Salt Lake City, Utah, USA.

Suchan, T.A. (2002). Usability studies of geovisualization software in the workplace. In: *Proceedings of the National Conference for Digital Government Research*, Los Angeles, USA.

Takatsuka, M. (2001). An application of the self-organizing map and interactive 3-D visualization to geospatial data. In: *Proceedings of the Sixth International Conference on Geocomputation*, Brisbane, Australia.

Takatsuka, M. and Gahegan, M. (2002). GeoVISTA Studio: a codeless visual programming environment for geoscientific data analysis and visualization. *Computers & Geosciences*, **28**: 1131–1144.

Tobler, W. (2004). Thirty Five Years of Computer Cartograms. *Annals of the Association of American Geographers*, **94**(1): 58–73.

Tobón, C. (2002). *Usability Testing for Improving Interactive Geovisualization Techniques*. Working paper, Centre for Advanced Spatial Analysis, University College London, available at: http://www.casa.ucl.ac.uk/working_papers/Paper45.pdf (last accessed 13 July 2007).

Tobón, C. (2005). Evaluating geographic visualization tools and methods: an approach and experiment based upon user tasks. In: Dykes, J. MacEachren, A.M. and Kraak, M.-J. (eds), *Exploring Geovisualization*, pp.645–666. Amsterdam: Elsevier.

Tomaszewski, B., Robinson, A.C., Weaver, C., Stryker, M. and MacEachren, A.M. (2007). Geovisual analytics and crisis management. In: *Proceedings of the 4th International ISCRAM Conference*, Delft, The Netherlands.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

Unwin, A. and Unwin, D. (1998). Exploratory spatial data analysis with local statistics. *The Statistician*, **47**(3): 415–421.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, **3**: 111–126.

Ware, C. (2000). *Information Visualization: Perception for Design*. San Francisco, USA: Morgan Kaufmann Publishers.

Ware, C. and Plumlee, M. (2005). *3D geovisualization and the structure of visual space*. In: Dykes, J. MacEachren, A.M. and Kraak, M.-J. (eds), Exploring Geovisualization, pp. 567–576. Amsterdam: Elsevier.

5

Availability of Spatial Data Mining Techniques

Shashi Shekhar, Vijay Gandhi, Pusheng Zhang and Ranga Raju Vatsavai

5.1. INTRODUCTION

The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Spatial data mining (Roddick and Spiliopoulou, 1999; Shekhar and Chawla, 2003) is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limit the usefulness of conventional data mining techniques for extracting spatial patterns. Efficient tools for extracting information from geo-spatial data are crucial to organizations which make decisions based on large spatial datasets, including NASA, the National Imagery and Mapping Agency, the National Cancer Institute, and the United States Department of Transportation. These organizations are spread across many application domains including ecology and environmental management, public safety, transportation, Earth science, epidemiology, and climatology.

Extracting interesting and useful patterns from spatial datasets is more difficult than extracting corresponding patterns from traditional numeric and categorical data. Specific features of spatial data that preclude the use of general purpose data mining algorithms are: (a) rich data types (e.g., extended spatial objects) (b) implicit spatial relationships among the variables, (c) observations that are not independent, and (d) spatial autocorrelation among the features.

This chapter is organized as follows. In section 5.2, we provide an overview of spatial data. Section 5.3 presents important statistical concepts used in spatial data mining. Spatial Data Mining techniques, the main focus of this chapter, are explained in section 5.4. Specifically, we present major accomplishments in mining output patterns known as predictive models, semi-supervised approaches, outliers, co-location rules, and clustering. In section 5.5, we briefly review the computational processes for spatial data mining techniques. Finally, in section 5.6, we identify areas of spatial data mining where further research is needed. This chapter does not discuss spatial statistics or algorithm-level computational processes in depth as these topics are beyond the scope of this chapter.

5.2. DATA INPUT

The data inputs of spatial data mining are more complex than the inputs of classical data mining because they include extended objects such as points, lines, and polygons. The data inputs of spatial data mining have two distinct types of attributes: non-spatial attribute and spatial attribute. Non-spatial attributes are used to characterize non-spatial features of objects, such as name, population, and unemployment rate for a city. They are the same as the attributes used in the data inputs of classical data mining. Spatial attributes are used to define the spatial location and extent of spatial objects (Bolstad, 2002). The spatial attributes of a spatial object most often include information related to spatial locations, e.g., longitude, latitude and elevation, as well as shape.

Relationships among non-spatial objects are explicit in data inputs, e.g., arithmetic relation, ordering, is instance of, subclass of, and membership of. In contrast, relationships among spatial objects, such as overlap, intersect, and behind are often implicit. Table 5.1 lists non-spatial relationships and their corresponding spatial relationship. One possible way to deal with implicit spatial relationships is to materialize the relationships into traditional data input columns and then apply classical data mining techniques (Agrawal and Srikant, 1994; Jain and Dubes, 1988; Quinlan, 1993). However, the materialization can result in loss of information. Another way to capture implicit spatial relationships is to develop models or techniques to incorporate spatial information into the spatial data mining process. We discuss a few case studies of such techniques in section 5.4.

The representation of spatial data and use of spatial operators has been standardized by the Open GIS (OGIS) consortium for interoperability of spatial applications, such as Geographic Information Systems. OGIS defines standard spatial data types which can be used in combination to represent a spatial object. Some examples of OGIS data types include *Point*, *Curve*, *Surface*, and *Geometry Collection*. In addition to specifying data types, the OGIS standard also includes three categories of spatial operations: (a) basic spatial operations which can to applied to all

Table 5.1 Relationships among non-spatial data and spatial data

Non-spatial relationship (explicit)	*Spatial relationship (often implicit)*
Arithmetic	Set-oriented: union, intersection, membership, ···
Ordering	Topological: meet, within, overlap, ···
Is instance of	Directional: North, NE, left, above, behind, ···
Subclass of	Metric: e.g., distance, area, perimeter, ···
Part of	Dynamic: update, create, destroy, ···
Membership of	Shape-based and visibility

geometry datatypes, e.g., to find the boundary of a spatial object, (b) operations to test for topological relationship between objects, e.g., to find if two spatial objects *overlap*, and (c) operations to perform spatial analysis, e.g., to calculate the shortest distance path between two spatial objects.

A recent topic of research is the representation of spatial data which have an associated temporal aspect. A location based service is an example in which a service is offered based on the location and time of an entity. Current OGIS standards do not yet support such systems.

5.3. STATISTICAL FOUNDATION

Readers of this handbook will be exposed to more statistical foundations in later chapters. Here we address only the basic concepts needed to follow the rest of this chapter.

Statistical models (Cressie, 1993) are often used to represent observations in terms of random variables. These models can then be used for estimation, description, and prediction based on probability theory. Spatial data can be thought of as resulting from observations on the stochastic process $Z(s) : s \in D$, where s is a spatial location and D is possibly a random set in a spatial framework. Here we present three spatial statistical problems one might encounter: point process, lattice, and geostatistics.

Point process

A point process is a model for the spatial distribution of the points in a point pattern. Several natural processes can be modeled as spatial point patterns, e.g., positions of trees in a forest and locations of bird habitats in a wetland. Spatial point patterns can be broadly grouped into random or non-random processes. Real point patterns are often compared with a random pattern (generated by a Poisson process) using the average distance between a point and its nearest neighbor. For a random pattern, this average distance is expected to be $1/(2 \times \sqrt{density})$, where density is the average number of points per unit area. If for a real process, the computed distance falls within a certain limit, then we conclude that the pattern is generated by a random process; otherwise it is a non-random process.

Lattice

A lattice is a model for a gridded space in a spatial framework. Here the lattice refers to a countable collection of regular or irregular spatial sites related to each other via a neighborhood relationship. Several spatial statistical analyses, e.g., the spatial autoregressive model and Markov random fields can be applied on lattice data.

Geostatistics

Geostatistics deals with the analysis of spatial continuity and weak stationarity (Cressie, 1993), which is an inherent characteristic of spatial datasets. Geostatistics provides a set of statistics tools, such as kriging (Cressie, 1993), to the interpolation of attributes at unsampled locations.

One of the fundamental assumptions of statistical analysis is that the data samples are independently generated: like successive tosses of coin, or the rolling of a die. However, in the analysis of spatial data, the assumption about the independence of samples is generally false. In fact, spatial data tends to be highly self correlated. For example, people with similar characteristics, occupation and background tend to cluster together in the same neighborhoods. The economies of a region tend to be similar. Changes in natural resources, wildlife, and temperature vary gradually over space. The property of like things to cluster in space is so

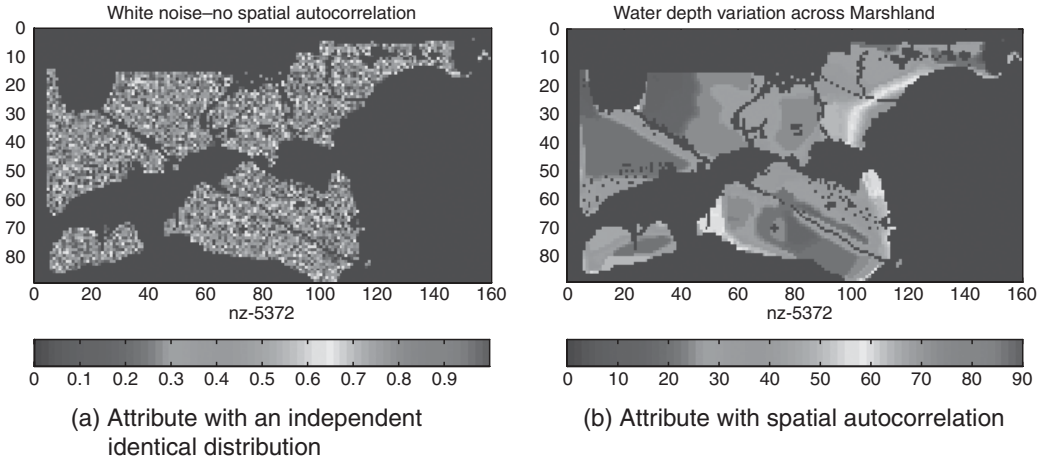


Figure 5.1 Attribute values in space with independent identical distribution and spatial autocorrelation.

fundamental that geographers have elevated it to the status of the first law of geography: *Everything is related to everything else but nearby things are more related than distant things* (Tobler, 1979). In spatial statistics, an area within statistics devoted to the analysis of spatial data, this property is called spatial autocorrelation (Shekhar and Chawla, 2003). For example, Figure 5.1 shows the value distributions of an attribute in a spatial framework for an independent identical distribution (Figure 5.1(a)) and a distribution with spatial autocorrelation (Figure 5.1(b)).

Knowledge discovery techniques which ignore spatial autocorrelation typically perform poorly in the presence of spatial data. Often the spatial dependencies arise due to the inherent characteristics of the phenomena under study, but in particular they arise due to the fact that the spatial resolution of imaging sensors are finer than the size of the object being observed. For example, remote sensing satellites have resolutions ranging from 30 m (e.g., the Enhanced Thematic Mapper of the Landsat 7 satellite of NASA) to 1 m (e.g., the IKONOS satellite from SpaceImaging), while the objects under study (e.g., urban, forest, water) are often much larger than 30 m. As a result, per-pixel-based classifiers, which do not take spatial context into account, often produce classified images with *salt* and *pepper* noise. These classifiers also suffer in terms of classification accuracy.

The spatial relationship among locations in a spatial framework is often modeled via a contiguity matrix. A simple contiguity matrix may represent a neighborhood relationship defined using adjacency, Euclidean distance, etc. Example definitions of neighborhood using adjacency include a four-neighborhood and an eight-neighborhood. Given a uniform gridded spatial framework, a four-neighborhood assumes that a pair of locations influence each other if they share an edge. An eight-neighborhood assumes that a pair of locations influence each other if they share either an edge or a vertex.

Figure 5.2(a) shows a gridded spatial framework with four locations, A, B, C, and D. A binary matrix representation of a four-neighborhood relationship is shown in Figure 5.2(b). The row-normalized representation of this matrix is called a contiguity matrix, as shown in Figure 5.2(c).

(a) Spatial framework (b) Neighbor relationship (c) Contiguity matrix

Figure 5.2 A spatial framework and its four-neighborhood contiguity matrix.

Other contiguity matrices can be designed to model neighborhood relationships based on distance. The essential idea is to specify the pairs of locations that influence each other along with the relative intensity of interaction. More general models of spatial relationships using cliques and hypergraphs are available in the literature (Warrender and Augusteijn, 1999). In spatial statistics, spatial autocorrelation is quantified using measures such as Ripley's K-function and Moran's I (Cressie, 1993).

A topic of recent interest in the field of spatial statistics is multiscale modeling. Since most physical and human processes vary with spatial scale, multi-scale representation is very important. Much of the current work related to multi-scale modeling lacks a formal statistical framework. However, Kolaczyk *et al*. (2005) use statistical models called *mixlets* which allow representation of spatial information at multiple scales.

5.4. OUTPUT PATTERNS

In this section, we present spatial data mining techniques for different output patterns: predictive models, spatial clustering, semi-supervised learning, spatial outliers, and spatial co-location rules.

5.4.1. Predictive models

The prediction of events occurring at particular geographic locations is very important in several application domains. Examples of problems which require location prediction include crime analysis, cellular networking, and natural disasters such as fires, floods, droughts, vegetation diseases, and earthquakes. Here we describe one such problem domain and provide two spatial data mining techniques for predicting locations, namely the Spatial Autoregression Model (SAR) and Markov Random Fields (MRF).

An application domain

We begin by introducing an example to illustrate the different concepts related to location prediction in spatial data mining. We are given data about two wetlands, named Darr and Stubble, on the shores of Lake Erie in Ohio, USA in order to *predict* the spatial distribution of a marsh-breeding bird, the red-winged blackbird (*Agelaius phoeniceus*). The data was collected from April to June in two successive years, 1995 and 1996.

A uniform grid was imposed on the two wetlands and different types of measurements were recorded at each cell or pixel. In total, the values of seven attributes were recorded at each cell. Domain knowledge is crucial in deciding which attributes are important

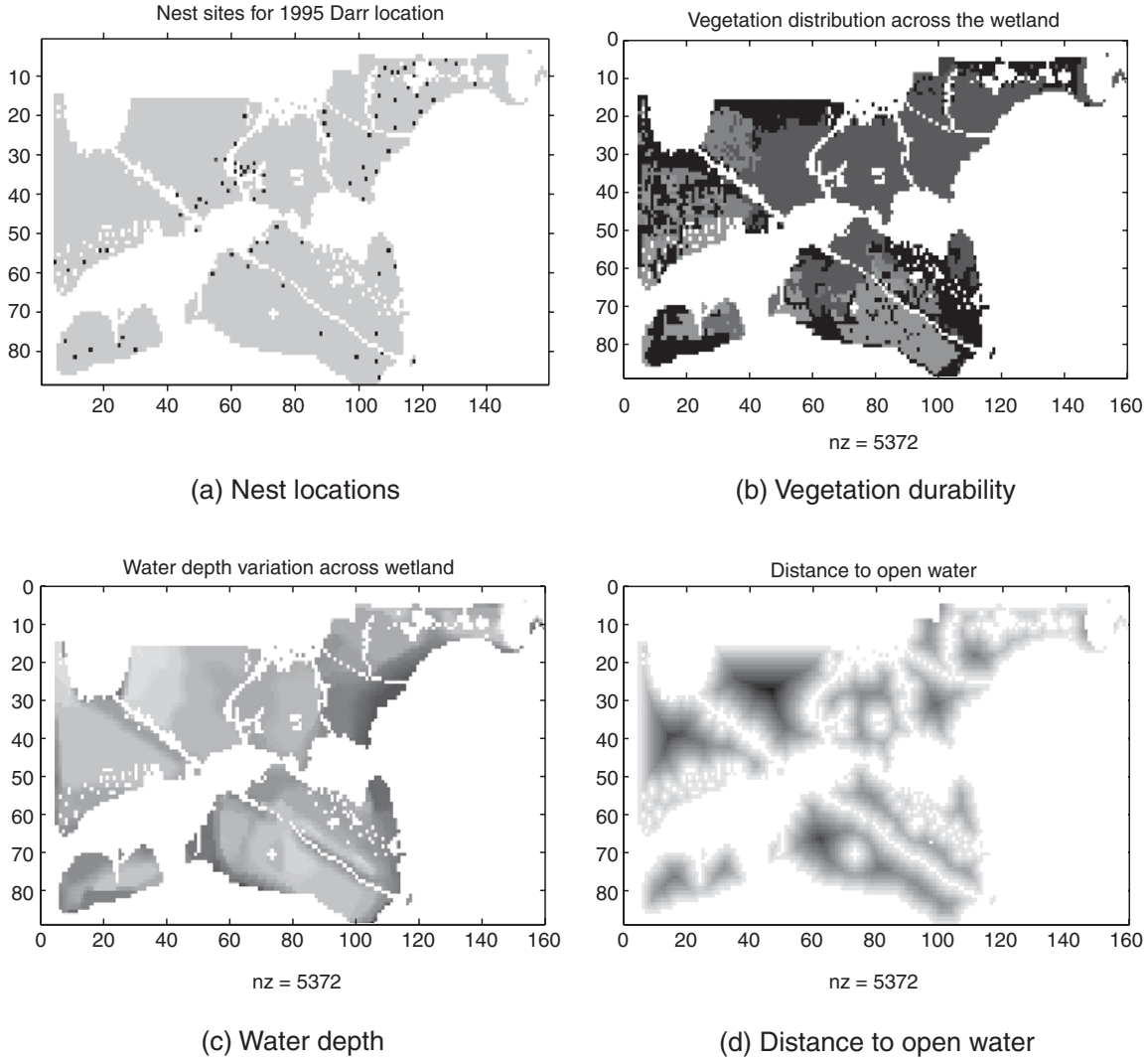


Figure 5.3 (a) Learning dataset: the geometry of the Darr wetland and the locations of the nests, (b) the spatial distribution of *vegetation durability* over the marshland, (c) the spatial distribution of *water depth*, and (d) the spatial distribution of *distance to open water*.

and which are not. For example, *vegetation durability* was chosen over *vegetation species* because specialized knowledge about the bird-nesting habits of the red-winged blackbird suggested that the choice of nest location is more dependent on plant structure, plant resistance to wind, and wave action than on the plant species.

An important goal is to build a model for predicting the location of bird nests in the wetlands. Typically, the model is built using a portion of the data, called the learning or training data, and then tested on the remainder of the data, called the testing data. In this study a model was built using the 1995 Darr wetland data and then tested using the 1995 Stubble wetland data. In the learning data, all the attributes are used to build the model and in the training data, one value is hidden, in this case the location of the nests. Using knowledge gained from the 1995 Darr data and the value of the independent attributes in the test data, the goal is to predict the location of the nests in the 1995 Stubble data.

Modeling spatial dependencies using the SAR and MRF models

Several previous studies (Jhung and Swain, 1996; Solberg *et al*., 1996) have shown that the modeling of spatial dependency (often called context) during the classification process improves overall classification accuracy. Spatial context can be defined by the relationships between spatially adjacent pixels in a small neighborhood. In this section, we present two approaches to modeling spatial dependency: the SAR and MRF-based Bayesian classifiers.

Spatial autoregression model

The spatial autoregressive model decomposes a classifier $\hat{f_C}$ into two parts, namely spatial autoregression and logistic transformation. We first show how spatial dependencies are modeled using the framework of logistic regression analysis. In the spatial autoregression model, the spatial dependencies of the error term, or, the dependent variable, are directly modeled in the regression equation (Anselin, 1988). If the dependent values y_i are related to each other, then the regression equation can be modified as

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\beta + \varepsilon \qquad (5.1)$$

Here $\mathbf{W}$ is the neighborhood relationship contiguity matrix and ρ is a parameter that reflects the strength of the spatial dependencies between the elements of the dependent variable. After the correction term $\rho\mathbf{W}\mathbf{y}$ is introduced, the components of the residual error vector ε are then assumed to be generated from independent and identical standard normal distributions. As in the case of classical regression, the SAR equation has to be transformed via the logistic function for binary dependent variables.

We refer to this equation as the Spatial Autoregressive Model (SAR). Notice that when $\rho = 0$, this equation collapses to the classical regression model. The benefits of modeling spatial autocorrelation are many: the residual error will have much lower spatial autocorrelation (i.e., systematic variation). With the proper choice of W, the residual error should, at least theoretically, have no systematic variation. If the spatial autocorrelation coefficient is statistically significant, then SAR will quantify the presence of spatial autocorrelation. It will indicate the extent to which variations in the dependent variable (y) are explained by the average of neighboring observation values. Finally, the model will have a better fit, (i.e., a higher R-squared statistic).

Markov random field-based Bayesian classifiers

Markov random field-based Bayesian classifiers estimate the classification model $\hat{f_C}$ using MRF and Bayes' rule. A set of random variables whose interdependency relationship is represented by an undirected graph (i.e., a symmetric neighborhood matrix) is called a Markov random field (Li, 1995). The Markov property specifies that a variable depends only on its neighbors and is independent of all other variables. The location prediction problem can be modeled in this framework by assuming that the class label, $l_i = f_C(s_i)$, of different locations, s_i, constitutes an MRF. In other words, random variable l_i is independent of l_j if $W(s_i, s_j) = 0$.

The Bayesian rule can be used to predict l_i from feature value vector X and neighborhood class label vector L_i as follows:

$$Pr\left(l_i \mid X, L_i\right) = \frac{Pr\left(X \mid l_i, L_i\right) Pr\left(l_i \mid L_i\right)}{Pr\left(X\right)}. \qquad (5.2)$$

The solution procedure can estimate $Pr(l_i \mid L_i)$ from the training data, where L_i denotes a set of labels in the neighborhood of s_i, excluding the label at s_i, by examining the ratios of the frequencies of class labels to the total number of locations in the spatial framework. $Pr(l_i \mid L_i)$ can be estimated using kernel functions from the observed values in the training dataset. For reliable estimates, even larger training datasets are needed relative to those needed for the Bayesian classifiers without spatial context, since we are estimating a more complex distribution. An assumption on $Pr(l_i \mid L_i)$ may be useful if the training dataset available is not large enough. A common assumption is the uniformity of influence from all neighbors of a location. For computational efficiency it can be assumed that only local explanatory data $X(s_i)$ and neighborhood label L_i are relevant in predicting class label $l_i = f_C(s_i)$. It is common to assume that all interaction between neighbors is captured via the interaction in the class label variable. Many domains also use specific parametric probability distribution forms, leading to simpler solution procedures. In addition, it is frequently easier to work with a Gibbs distribution specialized by the locally defined MRF through the Hammersley–Clifford theorem (Besag, 1974).

A more detailed theoretical and experimental comparison of these methods can be found in Shekhar *et al.* (2002). Although MRF and SAR classification have different formulations, they share a common goal, estimating the posterior probability distribution: $p(l_i \mid X)$. However, the posterior for the two models is computed differently with different assumptions. For MRF the posterior is computed using Bayes' rule. On the other hand, in logistic regression, the posterior distribution is directly fit to the data. One important difference between logistic regression and MRF is that logistic regression assumes no dependence on neighboring classes. Logistic regression and logistic SAR models belong to a more general exponential family. The exponential family is given by $Pr(u \mid v) = \exp(A(\theta_v) + B(u,\ p) + \theta_v^T u)u$ where u, and v are location and label respectively. This exponential family includes many of the common distributions such as Gaussian, Binomial, Bernoulli, and Poisson as special cases.

Experiments were carried out on the Darr and Stubble wetlands to compare classical regression, SAR, and the MRF-based Bayesian classifiers. The results showed that the MRF models yield better spatial and classification accuracies over SAR in the prediction of the locations of bird nests. It was also observed that SAR predictions are extremely localized, missing actual nests over a large part of the marshlands (Shekhar *et al.*, 2002).

5.4.2. Spatial clustering

Spatial clustering is a process of grouping a set of spatial objects into clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters. For example, clustering is used to determine the 'hot spots' in crime analysis and disease tracking. Hot spot analysis is the process of finding unusually dense event clusters across time and space. Many criminal justice agencies are exploring the benefits provided by computer technologies to identify crime hot spots in order to take preventive strategies such as deploying saturation patrols in hot spot areas.

Spatial clustering can be applied to group similar spatial objects together; the implicit assumption is that patterns in space tend to be grouped rather than randomly located. However, the statistical significance of spatial clusters should be measured by testing the assumption in the data. The test is critical

before proceeding with any serious clustering analyses.

Complete spatial randomness, cluster, and decluster

In spatial statistics, the standard against which spatial patterns are often compared is a completely spatially random point process, and departures indicate that the pattern is not distributed randomly in space. *Complete spatial randomness (CSR)* (Cressie, 1993) is synonymous with a homogeneous Poisson process. The patterns of the process are independently and uniformly distributed over space, i.e., the patterns are equally likely to occur anywhere and do not interact with each other. However, patterns generated by a non-random process can be either cluster patterns (aggregated patterns) or decluster patterns (uniformly spaced patterns).

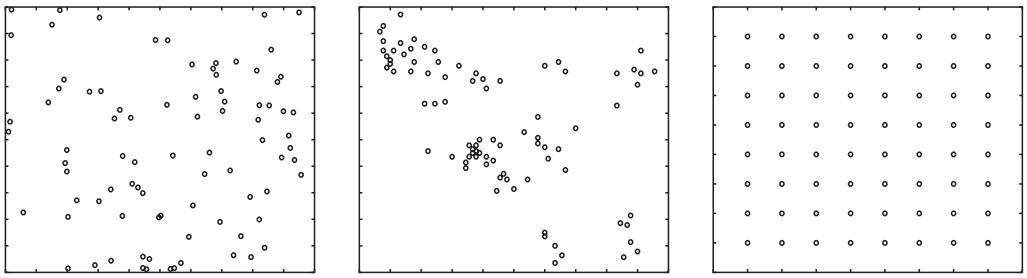
To illustrate, Figure 5.4 shows realizations from a completely spatially random process, a spatial cluster process, and a spatial decluster process (each conditioned to have 80 points) in a square. Notice in Figure 5.4(a) that the complete spatial randomness pattern seems to exhibit some clustering. This is not an unrepresentive realization, but illustrates a well-known property of homogeneous Poisson processes: event-to-nearest-event distances are proportional to χ^2 random variables, whose densities have a substantial amount of probability near zero (Cressie, 1993). Spatial clustering is more statistically significant when the data exhibit a cluster pattern rather than a CSR pattern or decluster pattern.

Several statistical methods can be applied to quantify deviations of patterns from a complete spatial randomness point pattern (Cressie, 1993). One type of descriptive statistic is based on quadrats (i.e., well defined area, often rectangular in shape). Usually quadrats of random location and orientations in the quadrats are counted, and statistics derived from the counters are computed. Another type of statistic is based on distances between patterns; one such type is Ripley's K-function (Cressie, 1993).

After the verification of the statistical significance of the spatial clustering, classical clustering algorithms (Han *et al*., 2001) can be used to discover interesting clusters.

Clustering point process

As discussed in section 5.3, a point process is a model for the spatial distribution of



(a) Complete spatial randomness (CSR) pattern

(b) Cluster pattern

(c) Decluster pattern

Figure 5.4 Illustration of CSR, cluster, and decluster patterns.

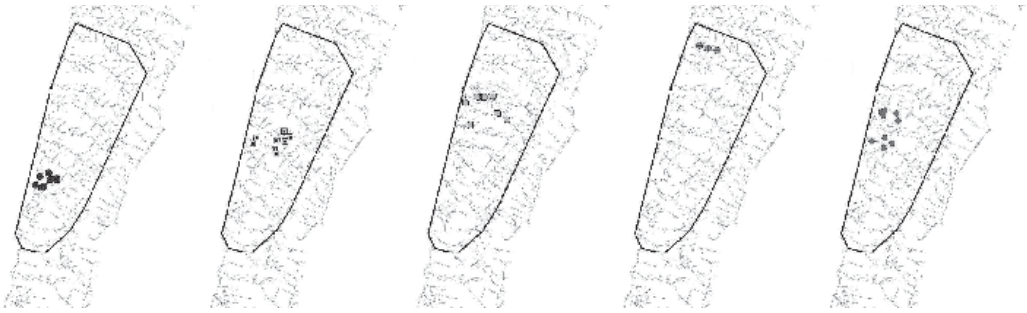


Figure 5.5 Marked spatial point process. Spatial locations for different female chimpanzees at the Gombe National Park, Tanzania.

the spatial points in a point pattern. A point process in which each of the spatial locations is marked with a unique label is called a marked spatial point process. Clustering of marked spatial point processes is an interesting problem in many application domains. For example, in behavioral ecology, ecologists are interested in finding clusters of individual chimpanzees based on their space usage, which usually consists of several spatial points for each individual. An example of marked spatial point processes is shown in Figure 5.5.

The problem of clustering marked spatial point processes is a generalization of the problem of clustering spatial points, where instead of a single spatial location for each category, we have multiple spatial locations for each category. Each category is a spatial point process. Classical clustering approaches handle homogeneous spatial points and hence cannot cluster marked spatial point processes. A very limited amount of research has been done in the area of clustering marked spatial point processes. (Han *et al*., 2001).

A data mining technique for clustering marked spatial point processes is proposed by (Shekhar *et al*., 2006). This algorithm is based on the intuition that the intra-cluster similarity must be significantly higher than the inter-cluster similarity. During clustering, Besag's *L*-function (Besag, 1977), which is a modified version of Ripley's *K*-function (Cressie, 1993), is used to quantify the second-order interaction between point processes. This measure provides the correlation between the observed and expected pairs of points at a certain distance from each other. Based on the value of this measure, marked point processes can be clustered hierarchically, to produce a dendrogram or a block diagonal matrix, which can be analyzed by domain experts to find a threshold level to identify proper clusters.

5.4.3. Semi-supervised learning

The methods described in the previous section are examples of supervised learning algorithms. In supervised methods, the model is built using a training dataset. For example, in a remote sensing image, training data will be a collection of labeled pixels. Practically, it is very difficult to collect labels for all training data. Hence an approach which does not require many labeled samples is needed. Such an approach which uses less labeled samples and a large number of unlabeled samples is called semi-supervised learning (Vatsavai and Shekhar, 2005). Based on the Expectation–Maximization (EM) algorithm,

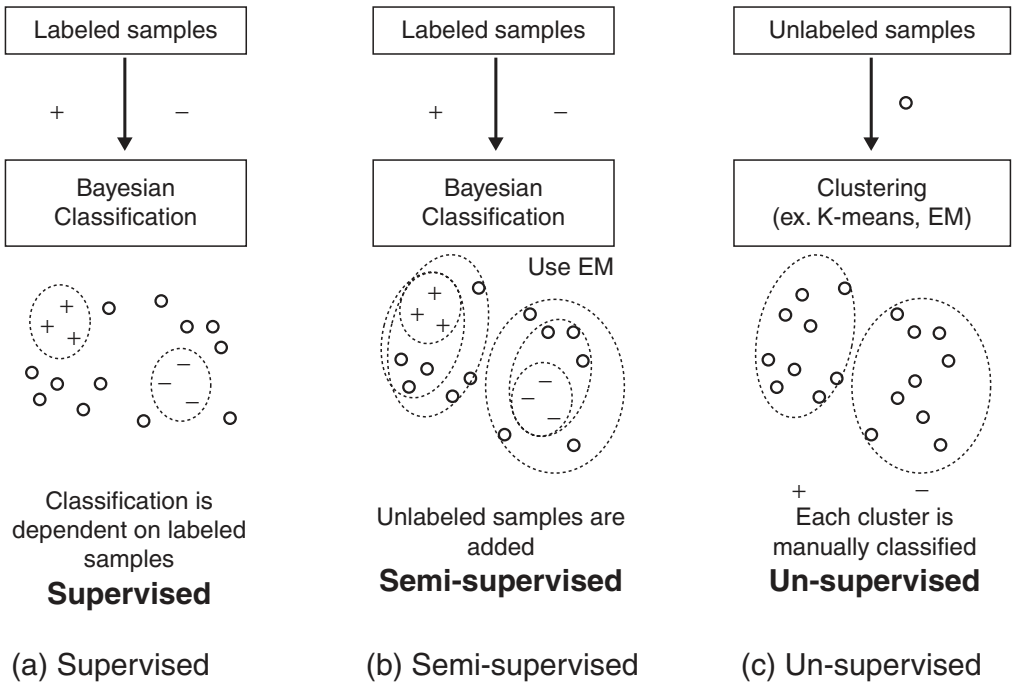


Figure 5.6 Illustration of different approaches in Classification. '+' and '−' indicates labeled data, 'o' indicates unlabeled data.

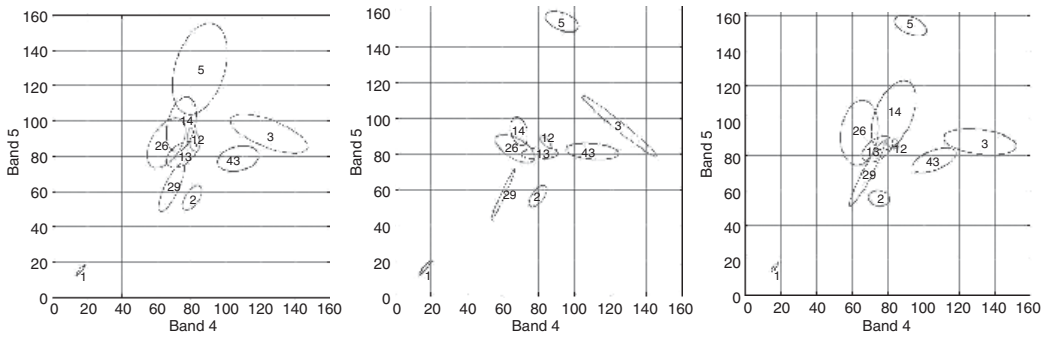
maximum likelihood, and maximum *a posteriori* classifiers, the semi-supervised method utilizes a small set of labeled and a large number of unlabeled training samples to build a model.

Figure 5.6 illustrates the difference between different approaches used in classification. The supervised approach shown in Figure 5.6(a) requires many labeled data, in this case '+' and '−' to build a model. An unsupervised approach does not require any training dataset to build a model. The semi-supervised approach shown in Figure 5.6(c) uses a small number of labeled and a large number of unlabeled datasets to build a model.

A semi-supervised approach is better than using a supervised approach with a smaller number of labeled samples. Figure 5.7 shows an example which proves that including an unlabeled dataset and using a semi-supervised approach improves the classification model. Figure 5.7 shows classification of satellite imagery into different classes. Figure 5.7(a) is obtained by using 100 labeled data points in the training dataset. The model obtained using only 20 labeled data points is shown in Figure 5.7(b). As it can be seen the model with a lesser number of labeled data points is poorer as compared to the model with a greater number of data points. However, the model with a lesser number of labeled data points can be improved by including unlabeled data points and using a semi-supervised technique. The resulting model is shown in Figure 5.7(c).

5.4.4. Spatial outliers

Outliers have been informally defined as observations in a dataset which appear to be inconsistent with the remainder of that set of data (Barnett and Lewis, 1994), or which



(a) Results with 100 labeled samples, supervised

(b) Results with 20 labeled samples, supervised

(c) Results with 20 labeled and 80 unlabeled samples, semi-supervised

Figure 5.7 Illustration of supervised and semi-supervised approach.

deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism (Hawkins, 1980). The identification of global outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as credit card fraud, athlete performance analysis, voting irregularity, and severe weather prediction. This section focuses on spatial outliers, i.e., observations which appear to be inconsistent with their neighborhoods. Detecting spatial outliers is useful in many applications of geographic information systems and spatial databases, including transportation, ecology, public safety, public health, climatology, and location-based services.

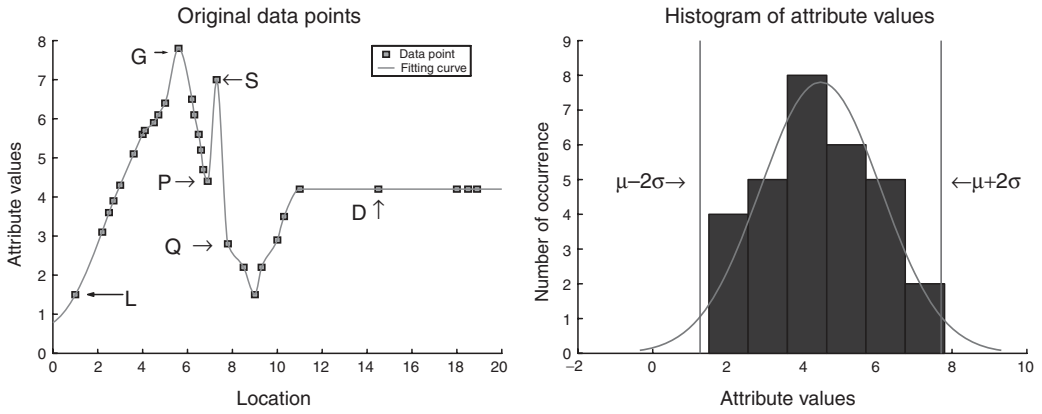
A spatial outlier is a spatially referenced object whose non-spatial attribute values differ significantly from those of other spatially referenced objects in its spatial neighborhood. Informally, a spatial outlier is a local instability (in values of non-spatial attributes) or a spatially referenced object whose non-spatial attributes are extreme relative to its neighbors, even though the attributes may not be significantly different from the entire population. For example, a new house in an old neighborhood of a growing metropolitan area is a spatial outlier based on the non-spatial attribute house age.

Illustrative examples

We use an example to illustrate the differences among global and spatial outlier detection methods. In Figure 5.8(a), the X-axis is the location of data points in one-dimensional space; the Y-axis is the attribute value for each data point. Global outlier detection methods ignore the spatial location of each data point and fit the distribution model to the values of the non-spatial attribute. The outlier detected using this approach is the data point G, which has an extremely high attribute value 7.9, exceeding the threshold of $\mu + 2\sigma = 4.49 + 2 \times 1.61 = 7.71$, as shown in Figure 5.8(b). This test assumes a normal distribution for attribute values. On the other hand, S is a spatial outlier whose observed value is significantly different than its neighbors P and Q.

Tests for detecting spatial outliers

Tests to detect spatial outliers separate spatial attributes from non-spatial attributes.



(a) An example dataset (b) Histogram

Figure 5.8 A dataset for outlier detection.

Spatial attributes are used to characterize location, neighborhood, and distance. Non-spatial attribute dimensions are used to compare a spatially referenced object to its neighbors. The spatial statistics literature provides two kinds of bi-partite multidimensional tests, namely graphical tests and quantitative tests. Graphical tests, which are based on the visualization of spatial data, highlight spatial outliers. Example methods include variogram clouds and Moran scatterplots. Quantitative methods provide a precise test to distinguish spatial outliers from the remainder of data. Scatterplots (Anselin, 1994) are a representative technique from the quantitative family.

A variogram-cloud (Cressie, 1993) displays data points related by neighborhood relationships. For each pair of locations, the square-root of the absolute difference between attribute values at the locations versus the Euclidean distance between the locations are plotted. In datasets exhibiting strong spatial dependence, the variance in the attribute differences will increase with increasing distance between locations. Locations that are near to one another, but with large attribute differences, might indicate a spatial outlier, even though the values at both locations may appear to be reasonable when examining the dataset non-spatially. Figure 5.9(a) shows a variogram cloud for the example dataset shown in Figure 5.8(a). This plot shows that two pairs (P, S) and (Q, S) on the left-hand side lie above the main group of pairs, and are possibly related to spatial outliers. The point S may be identified as a spatial outlier since it occurs in both pairs (Q, S) and (P, S). However, graphical tests of spatial outlier detection are limited by the lack of precise criteria to distinguish spatial outliers. In addition, a variogram cloud requires non-trivial post-processing of highlighted pairs to separate spatial outliers from their neighbors, particularly when multiple outliers are present, or density varies greatly.

A Moran scatterplot (Anselin, 1995) is a plot of a normalized attribute value $(Z[f(i)] = (f(i) - \mu_f)/\theta_f)$ against the neighborhood average of normalized attribute values $(W \bullet Z)$, where W is the row-normalized (i.e., $\sum_j W_{ij} = 1$) neighborhood matrix, (i.e., $W_{ij} > 0$ iff neighbor (i, j)). The upper left and lower right quadrants of Figure 5.9(b) indicate a spatial association of dissimilar values: low values surrounded by high value neighbors (e.g., points P and Q),

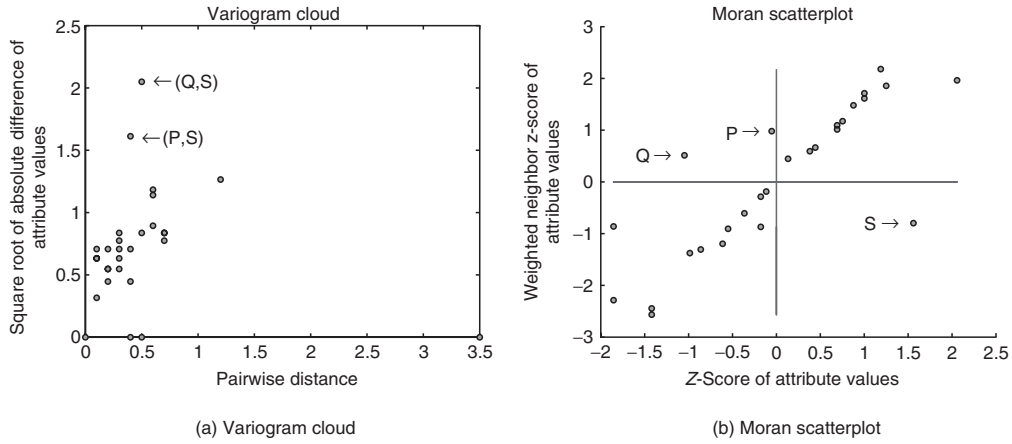


Figure 5.9 Variogram cloud and Moran scatterplot to detect spatial outliers.

and high values surrounded by low values (e.g., point S). Thus we can identify points (nodes) that are surrounded by unusually high or low value neighbors. These points can be treated as spatial outliers.

A scatterplot (Anselin, 1994) shows attribute values on the X-axis and the average of the attribute values in the neighborhood on the Y-axis. A least square regression line is used to identify spatial outliers. A scatter sloping upward to the right indicates a positive spatial autocorrelation (adjacent values tend to be similar); a scatter sloping upward to the left indicates a negative spatial autocorrelation. The residual is defined as the vertical distance (Y-axis) between a point P with location (X_p, Y_p) to the regression line $Y = mX + b$, that is, residual $\varepsilon = Y_p - (mX_p + b)$. Cases with standardized residuals, $\varepsilon_{standard} = (\varepsilon - \mu_\varepsilon)/\sigma_\varepsilon$, greater than 3.0 or less than -3.0 are flagged as possible spatial outliers, where μ_ε and σ_ε are the mean and standard deviation of the distribution of the error term ε. In Figure 5.10(a), a scatterplot shows the attribute values plotted against the average of the attribute values in neighboring areas for the dataset in Figure 5.8(a). The point S turns out to be the farthest from the regression line and may be identified as a spatial outlier.

A location (sensor) is compared to its neighborhood using the function $S(x) = [f(x) - E_{y\in N(x)}(f(y))]$, where $f(x)$ is the attribute value for a location x, $N(x)$ is the set of neighbors of x, and $E_{y\in N(x)}(f(y))$ is the average attribute value for the neighbors of x. The statistic function $S(x)$ denotes the difference of the attribute value of a sensor located at x and the average attribute value of x's neighbors.

Spatial statistic $S(x)$ is normally distributed if the attribute value $f(x)$ is normally distributed. A popular test for detecting spatial outliers for normally distributed $f(x)$ can be described as follows: spatial statistic $Z_{s(x)} = |(S(x) - \mu_s)/\sigma_s| > \theta$. For each location x with an attribute value $f(x)$, the $S(x)$ is the difference between the attribute value at location x and the average attribute value of x's neighbors, μ_s is the mean value of $S(x)$, and σ_s is the value of the standard deviation of $S(x)$ over all stations. The choice of θ depends on a specified confidence level. For example, a confidence level of 95 percent will lead to $\theta \approx 2$.

Figure 5.10(b) shows the visualization of the spatial statistic method described above. The X-axis is the location of data points in one-dimensional space; the Y-axis is the value of spatial statistic $Z_{s(x)}$ for each

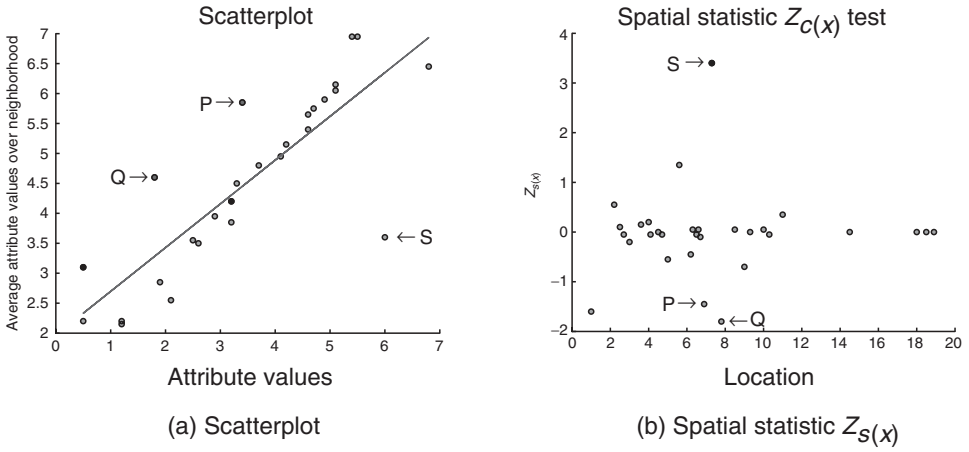


Figure 5.10 Scatterplot and Spatial Statistic $Z_{S(X)}$ to Detect Spatial Outliers.

data point. We can easily observe that point S has a $Z_{s(x)}$ value exceeding 3, and will be detected as a spatial outlier. Note that the two neighboring points P and Q of S have $Z_{s(x)}$ values close to -2 due to the presence of spatial outliers in their neighborhoods.

Designing computationaly efficient techniques to find spatial outliers is important. One efficient method is to compute the global statistical parameters using a spatial join (Shekhar *et al*., 2003). In this method, the algorithm computes the algebraic aggregate functions in a single scan of a spatial self-join from a spatial dataset using a neighborhood relationship. The computed values from the algebraic aggregate functions can be used to validate the outlier measure of a dataset.

A drawback in most of the techniques to detect multiple spatial outliers is that some of the data points are misclassified, i.e., either some of the true spatial outliers are ignored or some points are wrongly identified as spatial outliers. This misclassification occurs because most algorithms tend not to take into account the effect of an outlier in the neighborhood of another outlier. To overcome this problem, iterative algorithms and a median-based non-iterative algorithm can be used. In the iterative algorithms (Kou *et al*., 2003), only one outlier is detected in each iteration, and then its attribute value is modified in subsequent iterations so that it does not have a negative impact in detecting a new outlier. The median-based algorithm (Kou *et al*., 2003) reduces the impact of the presence of data points with extreme high or low attribute values.

5.4.5. Spatial co-location rules

Boolean spatial features are geographic object types which are either present or absent at different locations in a two-dimensional or three-dimensional metric space, e.g., the surface of the Earth. Examples of Boolean spatial features include plant species, animal species, road types, cancers, crime, and business types. Co-location patterns represent the subsets of the Boolean spatial features whose instances are often located in close geographic proximity. Examples include symbiotic species, e.g., Nile crocodile and Egyptian plover in ecology, and frontage roads and highways in metropolitan road maps.

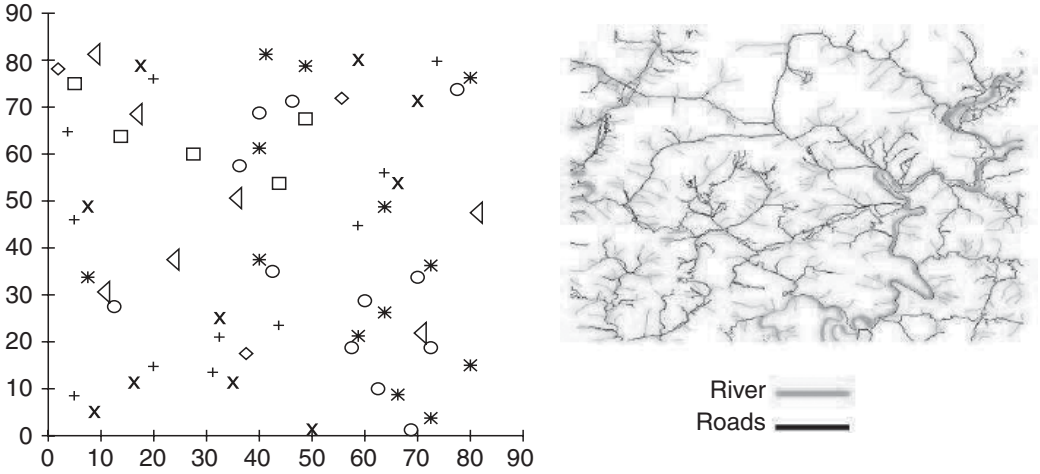


Figure 5.11 (a) Illustration of point spatial co-location patterns. Shapes represent different spatial feature types. Spatial features in sets {'+', '×',} and ('o', '∗') tend to be located together. (b) Co-location between roads and rivers. (Courtesy: Architecture Technology Corporation).

Co-location rules are models to infer the presence of Boolean spatial features in the neighborhood of instances of other Boolean spatial features. For example, 'Nile Crocodiles Egyptian Plover' predicts the presence of Egyptian Plover → birds in areas with Nile Crocodiles. Figure 5.11(a) shows a dataset consisting of instances of several Boolean spatial features, each represented by a distinct shape. A careful review reveals two co-location patterns, i.e., {'+', '×'} and {'○', '∗'}

Co-location rule discovery is a process to identify co-location patterns from large spatial datasets with a large number of Boolean features. The spatial co-location rule discovery problem looks similar to, but, in fact, is very different from the association rule mining problem (Agrawal and Srikant, 1994) because of the lack of transactions. In market basket datasets, transactions represent sets of item types bought together by customers. The support of an association is defined to be the fraction of transactions containing the association. Association rules are derived from all the associations with support values larger than a user given threshold. The purpose of mining association rules is to identify frequent item sets for planning store layouts or marketing campaigns. In the spatial co-location rule mining problem, transactions are often not explicit. The transactions in market basket analysis are independent of each other. Transactions are disjoint in the sense of not sharing instances of item types. In contrast, the instances of Boolean spatial features are embedded in a continuous space and share a variety of spatial relationships (e.g., neighbor) with each other.

Co-location rule approaches

Approaches to discovering co-location rules can be categorized into two classes, namely spatial statistics, and data mining approaches. Spatial statistics-based approaches use measures of spatial correlation to characterize the relationship between different types of spatial features. Measures of spatial correlation include the cross K-function with Monte Carlo simulation (Cressie, 1993),

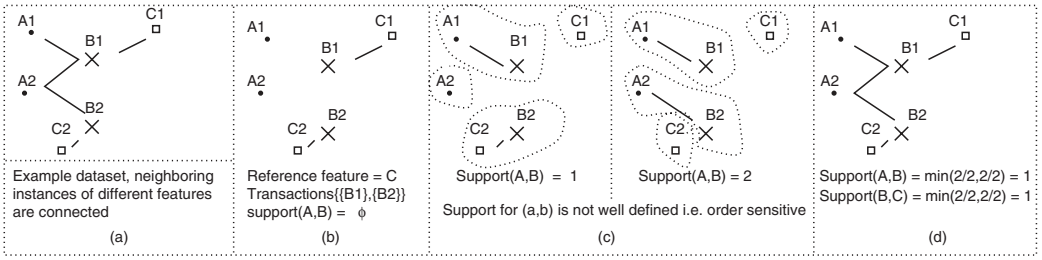


Figure 5.12 Example to illustrate different approaches to discovering co-location patterns (a) Example dataset. (b) Transaction based approach. Support measure is ill-defined and order sensitive. (c) A distance-based approach with *k*-neighbouring class sets. (d) A distance-based approach with event-centric model.

mean nearest-neighbor distance, and spatial regression models. Computing spatial correlation measures for all possible co-location patterns can be computationally expensive due to the exponential number of candidate subsets given a large collection of spatial Boolean features.

Data mining approaches can be further divided into a clustering-based map overlay approach and association rule-based approaches. A clustering-based map overlay approach treats every spatial attribute as a map layer. The spatial clusters (regions) of point-data in each layer are candidates for mining associations. Given X and Y as sets of layers, a clustered spatial association rule is defined as $X \Rightarrow Y(CS, CC\%)$, for $X \cap Y = \phi$, where CS is the clustered support, defined as the ratio of the area of the cluster (region) that satisfies both X and Y to the total area of the study region S. $CC\%$ is the clustered confidence, which can be interpreted as the percentage of area of clusters (regions) of X that intersect with the area of clusters (regions) of Y.

Association rule-based approaches can be divided into transaction- and distance-based approaches. Transaction-based approaches focus on defining transactions over space so that an *a priori*-like algorithm can be used. Transactions over space can be defined by a reference-feature centric model. Under this model, transactions are created around instances of one user-specified spatial feature. The association rules are derived using the *a priori* (Agarwal *et al*., 1993) algorithm. The rules formed are related to the reference feature. For example, consider the spatial dataset in Figure 5.12(a) with three feature types, A, B and C, each of which has two instances. The neighborhood relationships between instances are shown as edges. Co-locations (A, B) and (B, C) may be considered to be frequent in this example. Figure 5.12(b) shows transactions created by choosing C as the reference feature. Co-location (A, B) will not be found since it does not involve the reference feature. Generalizing the paradigm of forming rules related to a reference feature to the case where no reference feature is specified is non-trivial. Also, defining transactions around locations of instances of all features may yield duplicate counts for many candidate associations.

A distance-based approach was proposed concurrently by Morimoto (2001) and Shekhar and Huang (2001). Morimoto defined distance-based patterns called k-neighboring class sets, in which instances of objects are grouped together based on

their Euclidean distance from each other. In Morimoto's work, the number of instances for each pattern is used as the prevalence measure, which does not possess an anti-monotone property by nature. Since anti-monotonicity is required for such algorithms, Morimoto used a non-overlapping constraint to get the anti-monotone property for this measure. Also, it is possible that the instances of a k-neighboring class set are different depending on the order the class is added into the class set. This in turn yields different values of support of a given colocation. Figure 5.12(c) shows two possible partitions for the dataset of Figure 5.12(a), along with the supports for co-location (A, B).

The distance-based approach by Shekhar and Huang (2001) eliminates the non-overlapping-instance constraint. Their event-centric model finds subsets of spatial features likely to occur in a neighborhood around instances of given subsets of event types. For example, let us determine the probability of finding at least one instance of feature type B in the neighborhood of an instance of feature type A in Figure 5.12(a). There are two instances of type A and both have some instance(s) of type B in their neighborhoods. The conditional probability for the co-location rule is: *spatial feature A at location l* $\rightarrow$ *spatial feature type B in neighborhood is 100%*. This yields a well-defined prevalence measure (i.e., support) without the need for transactions. Figure 5.12(d) illustrates that the event-centric model will identify both (A, B) and (B, C) as frequent patterns.

Prevalence measures and conditional probability measures, called interest measures, are defined differently in different models, as summarized in Table 5.2. The transaction-based and distance-based k-neighboring class sets 'materialize' transactions and thus can use traditional support and confidence measures. The event-centric approach defined new transaction free measures, e.g., the participation index (see Shekhar and Huang (2001) for details).

To find co-locations, much of the time is spent in computing joins to identify instances of candidate co-location patterns. To decrease this computation time, a partial-join based approach (Yoo, 2004) or a join-less approach (Yoo, 2006) can be used. In the partial-join based approach, the number of instance joins for identifying candidate co-locations are minimized by transactionizing a spatial

Table 5.2 Interest measures for different co-location approaches

Model	*Items*	*Transactions defined by*	*Interest measures for* $C1 \rightarrow C2$		*Algorithm*
			Prevalence	*Conditional probability*	
Transaction based	Predicates on reference and relevant features	Instances of reference feature $C1$ and $C2$ involved with	Fraction of instance of reference feature with $C1 \cup C2$	$Pr(C2$ is true for an instance of reference features given $C1$ is true for that instance of reference feature)	*A priori*
Distance-based k-neighboring class sets	Boolean feature types	A partitioning of spatial dataset	Fraction of partitions with $C1 \cap C2$	$Pr(C2$ in a partition given $C1$ in that partition)	Partition-based
Distance-based event-centric	Boolean feature types	Neighborhoods of instances of feature types	Participation index of $C1 \cup C2$	$Pr(C2$ in a neighborhood of $C1)$	Join-based and Join-less

dataset under a neighbor relationship and tracing only residual neighborhood instances cut apart via the transactions. The key component is identifying instances of co-locations split across explicit transactions.

The join-less approach uses an instance-look-up scheme instead of an expensive spatial join for identifying co-location instances. Without any loss of co-location instances, this approach is more efficient and scalable for dense data than the join-based method.

5.5. COMPUTATIONAL PROCESS

Many generic algorithmic strategies have been generalized to apply to spatial data mining. For example, as shown in Table 5.3, algorithmic strategies, such as divide-and-conquer, filter-and-refine, ordering, hierarchical structure, and parameter estimation, have been used in spatial data mining.

In spatial data mining, spatial autocorrelation and low dimensionality in space (e.g., 2–3) provide more opportunities to improve computational efficiency than classical data mining. NASA Earth observation systems currently generate a large sequence of global snapshots of the Earth, including various atmospheric, land, and ocean measurements such as sea surface temperature, pressure, precipitation, and net primary production. Each climate attribute in a location has a sequence of observations at different time slots, e.g., a collection of monthly temperatures from 1951–2000 in Minneapolis. Finding locations where climate attributes are highly correlated is frequently used to retrieve interesting relationships among spatial objects of Earth science data. For example, such queries are used to identify the land locations whose climate is severely affected by El Niño. However, such correlation-based queries are computationally expensive due to the large number of spatial points, e.g., more than 250k spatial cells on the Earth at a 0.5 degree by 0.5 degree resolution, and the high dimensionality of sequences, e.g., 600 for the 1951–2000 monthly temperature data.

Table 5.3 Algorithmic strategies in spatial data mining.

Generic	*Spatial data mining*
Divide-and-conquer	Space partitioning
Filter-and-refine	Minimum-bounding-rectangle (MBR)
Ordering	Plane sweeping, Space-filling curves
Hierarchical structures	Spatial index, tree matching
Parameter estimation	Parameter estimation with spatial autocorrelation

A spatial indexing approach proposed by Zhang *et al*. (2003) exploits spatial autocorrelation to facilitate correlation-based queries. The approach groups similar time series together based on spatial proximity and constructs a search tree. The queries are processed using the search tree in a filter-and-refine style at the group level instead of at the time series level. Algebraic analyses using cost models and experimental evaluations showed that the proposed approach saves a large portion of computational cost, ranging from 40% to 98% (see Zhang *et al*. (2003) for details).

An important task in most of the spatial data mining techniques is to estimate the values of parameters. Inclusion of an autocorrelation parameter for spatial data mining increases the computation cost. For example, in the SAR model, we have to estimate the value of the spatial autocorrelation parameter, ρ and value of the regression coefficient β. Estimation of such parameters is done using maximum likelihood (ML) and Bayesian based techniques. A number of algorithms used for parameter estimation in a SAR model are provided in Table 5.4.

Maximum likelihood based SAR solution involves computation of two terms; a cheaper sum-of-squares error (SSE) term and a

Table 5.4 Classification of algorithms for spatial autoregression model (Celik *et al.*, 2006)

Method used	*Exact*	*Approximate*
Maximum likelihood	Applying direct sparese matrix algorithms, eigenvalue based 1-D surface Partitioning	ML based matrix exponential specification, graph theory approach, Taylor series approximation, Chebyshev polynomial approximation method, semiparametric estimates, characteristic polynomial approach, double bounded likelihood estimator, upper and lower bounds via divide and conquer, spatial autoregression local estimation
Bayesian	None	Bayesian matrix exponential specification, Markov Chain Monte Carlo (MCMC)

computationally expensive term, called the likelihood function, which involves a number of computations of determinant of a large matrix. ML-based solutions can be divided into exact and approximate solutions based on how they compute the computationally intensive term. Exact solutions suffer from high computational complexities and memory requirements. Approximate solutions are computationally feasible but many of these formulations still suffer from large memory requirements. One way to reduce the computational complexity of exact SAR solutions is to reduce the number of computation of a large martix in its likelihood function. This is done by setting an upper bound on the spatial autocorrelation parameter (Celik *et al.*, 2006).

5.6. RESEARCH NEEDS

In this section, we discuss some areas where further research is needed in spatial data mining.

5.6.1. Comparison of classical data mining techniques with spatial data mining techniques

As discussed in section 5.2, relationships among spatial objects are often implicit. It is possible to materialize the implicit relationships into traditional data input columns and then apply classical data mining techniques (Quinlan, 1993; Barnett and Lewis, 1994; Agrawal and Srikant, 1994; Jain and Dubes, 1988). Another way to deal with implicit relationships is to use specialized spatial data mining techniques, e.g., spatial autoregression and co-location mining. However, the existing literature does not provide guidance regarding the choice between classical data mining techniques and spatial data mining techniques to mine spatial data. New research is needed to compare the two sets of approaches in effectiveness and computational efficiency.

5.6.2. Spatial interest measures

The interest measures of patterns in spatial data mining are different from those in classical data mining, especially regarding the four important output patterns shown in Table 5.5.

For a two-class problem, the standard way to measure classification accuracy is to calculate the percentage of correctly classified objects. However, this measure may not be the most suitable in a spatial context. Spatial accuracy – how far the predictions are from the actuals – is equally important in this application domain due to the effects of the

Table 5.5 Interest measures of patterns for classical data mining and spatial data mining

	Classical data mining	*Spatial data mining*
Predictive model	Classification accuracy	Spatial accuracy
Cluster	Low coupling and high cohesion in feature space	Spatial continuity, unusual density, boundary
Outlier	Different from population or neighbors in feature space	Significant attribute discontinuity in geographic space
Association	Subset prevalence, $Pr[B \in T \mid A \in T, T : a\ transaction]$ Correlation	Spatial pattern prevalence $Pr[B \in N(A) \mid N :\ neighborhood]$ cross K-Function

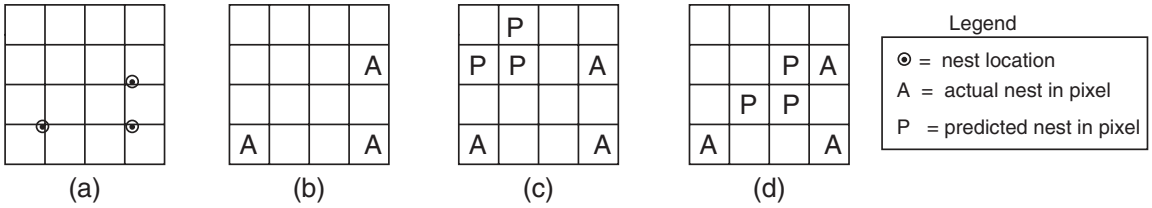


Figure 5.13 (a) The actual locations of nests. (b) Pixels with actual nests, (c) Location predicted by a model. (d) Location predicted by another model. Prediction (d) is spatially more accurate than (c).

discretizations of a continuous wetland into discrete pixels, as shown in Figure 5.13. Figure 5.13(a) shows the actual locations of nests and Figure 5.13(b) shows the pixels with actual nests. Note the loss of information during the discretization of continuous space into pixels. Many nest locations barely fall within the pixels labeled 'A' and are quite close to other blank pixels, which represent 'no-nest'. Now consider the two predictions shown in Figure 5.13(c) and (d). Domain scientists prefer prediction (d) over (c), since the predicted nest locations are closer on average to some actual nest locations. However, the classification accuracy measure cannot distinguish between Figure 5.13(c) and (d) since spatial accuracy is not incorporated in the classification accuracy measure. Hence, there is a need to investigate proper measures for location prediction to improve spatial accuracy.

5.6.3. Spatio-temporal data mining

Spatio-temporal data mining is done to extract patterns which have both spatial and temporal dimensions. Two examples where spatio-temporal data mining could be useful are – in a transportation network, to detect patterns of vehicle movement; and in a location based service, where a service can be offered to a customer by predicting his future location.

Consider a location-based service. It relies on tracking the positions of a mobile object. Since the positions change continuously, large volumes of updates are required on the database side. Mining the frequently used paths of a mobile object will reduce the number of updates required on the database. The main challenge here is to reduce the communication between the mobile object and the system (Civilis and Jensen, 2005).

Finding spatio-temporal sequential patterns is another research area in spatio-temporal data mining. The traditional sequential pattern mining algorithms are not applicable for spatio-temporal data. A recent algorithm to find such patterns is given in Cao and Cheung (2005).

Another challenge in spatio-temporal data mining is to find co-evolving spatial patterns. A spatially co-located pattern represents a pattern in which the instances are often located in close geographic proximity. Co-evolving spatial patterns are co-located spatial patterns whose temporal occurences are correlated with a special time series. For example, droughts and fires in Australia show similar variation as El Niño index values over the last 50 years (Taylor, 1998). Finding co-evolving spatial patterns is computationally expensive. Most of the methods in the literature do not work well for co-evolving spatial patterns because they do not consider the temporal domain of the co-location pattern. An efficient algorithm which takes temporal domain into account is proposed by Rogers *et al*. (2006).

5.6.4. Improving computational efficiency

Mining spatial patterns is often computationally expensive. For example, the estimation of the parameters for the spatial autoregressive model is an order of magnitude more expensive than that for linear regression in classical data mining. Similarly, the co-location mining algorithm is more expensive than the *a priori* algorithm for classical association rule mining (Agrawal and Srikant, 1994). Research is needed to reduce the computational costs of spatial data mining algorithms by a variety of approaches including the classical data mining algorithms as potential filters or components.

5.6.5. Modeling semantically rich spatial properties, such as topology

The spatial relationship among locations in a spatial framework is often modeled via a contiguity matrix using a neighborhood relationship defined using adjacency and distance. However, spatial connectivity and other complex spatial topological relationships in spatial networks are difficult to model using the continuity matrix. Research is needed to evaluate the value of enriching the continuity matrix beyond the neighborhood relationship. Another area with research potential is modeling of 3D topographic data (Penninga, 2005).

5.6.6. Statistical interpretation models for spatial patterns

Spatial patterns, such as spatial outliers and co-location rules, are identified in the spatial data mining process using unsupervised learning methods. There is a need for an independent measure of the statistical significance of such spatial patterns. For example, we may compare the co-location model with dedicated spatial statistical measures, such as Ripley's K-function, characterize the distribution of the participation index interest measure under spatial complete randomness using Monte Carlo simulation, and develop a statistical interpretation of co-location rules to compare the rules with other patterns in unsupervised learning.

Another challenge is the estimation of the detailed spatial parameters in a statistical model. Research is needed to design effective estimation procedures for the continuity matrices used in the spatial autoregressive model and Markov random field-based Bayesian classifiers from learning samples.

5.6.7. Effective visualization of spatial relationships

Visualization in spatial data mining is useful to identify interesting spatial patterns. As we discussed in section 5.2, the data inputs of spatial data mining have both spatial and non-spatial features. To facilitate the visualization of spatial relationships, research is needed on ways to represent both spatial and non-spatial features.

For example, many visual representations have been proposed for spatial outliers. However, we do not yet have a way to highlight spatial outliers within visualizations of spatial relationships. For instance, in variogram cloud (Figure 5.9(a)) and scatterplot (Figure 5.10(b)) visualizations, the spatial relationship between a single spatial outlier and its neighbors is not obvious. It is necessary to transfer the information back to the original map in geographic space to check neighbor relationships. Since a single spatial outlier tends to flag not only the spatial location of local instability but also its neighboring locations, it is important to group flagged locations and identify real spatial outliers from the group in the post-processing step.

5.6.8. Preprocessing spatial data

Spatial data mining techniques have been widely applied to the data in many application domains. However, research on the preprocessing of spatial data has lagged behind. Hence, there is a need for preprocessing techniques for spatial data to deal with problems such as treatment of missing location information and imprecise location specifications, cleaning of spatial data, feature selection, and data transformation.

5.7. SUMMARY

This chapter discussed major research accomplishments and techniques in spatial data mining, especially those related to four important output patterns: predictive models, spatial outliers, spatial co-location rules, and spatial clusters. Research needs in the area of spatial data mining were also identified.

ACKNOWLEDGMENTS

This work was supported in part by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014, the content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

We are particularly grateful to our collaborators Prof. Vipin Kumar, Prof. Paul Schrater, Dr. Sanjay Chawla, Dr. Chang-Tien Lu, Dr. Weili Wu, and Prof. Uygar Ozesmi for their various contributions. We also thank James Kang, Mete Celik, Jin Soung Yoo, Betsy George and anonymous reviewers for their valuable feedback on earlier versions of this chapter. We would also like to express our thanks to Kim Koffolt for improving the readability of this chapter.

REFERENCES

Agarwal, T., Imielinski, R. and Swami, A. (1993). Mining Association Rules between sets of items in large databases. In: *Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., May 1993.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for Mining Association Rules. In: *Proc. of Very Large Databases*, May 1994.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht, Netherlands: Kluwer.

Anselin, L. (1994). Exploratory spatial data analysis and geographic information systems. In: Painho, M. (ed.), *New Tools for Spatial Analysis*, pp. 45–54.

Anselin, L. (1995). Local indicators of spatial association: LISA. *Geographical Analysis*, **27**(2): 93–115.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. 3rd edn. New York: John Wiley.

Besag, J.E. (1974). Spatial interaction and statistical analysis of latice systems. *Journal of Royal Statistical Society, Ser. B*, **36**: 192–236.

Besag, J.E. (1977). Comments on Ripley's paper. *Journal of the Royal Statistical Society*.

Bolstad, P. (2002). *GIS Fundamentals: A First Text on GIS*. Eider Press.

Celik *et al.* (2006). *NORTHSTAR: A Parameter Estimation Method for the Spatial Auto-regression Model*.

Civilis, S.P.A. and Jensen, C.S. (2005). Techniques for efficient road-network-based tracking of moving objects. *IEEE Transaction on Knowledge and Data Engineering*, **17**(5).

Cressie, N.A. (1993). *Statistics for Spatial Data*, New York: Wiley.

Han, J., Kamber, M. and Tung, A. (2001). Spatial clustering methods in data mining: a survey. In: Miller, H. and Han, J. (eds). *Geographic Data Mining and Knowledge Discovery*. New York: Taylor and Francis.

Hawkins, D. (1980). *Identification of Outliers*. New York: Chapman and Hall.

Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data*. New York: Prentice Hall.

Jhung, Y. and Swain, P.H. (1996). Bayesian contextual classification based on modified M-estimates and Markov random fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **34**(1): 67–75.

Kolaczyk, G.S., Eric, D. and Ju Junchang, J. (2005). Multiscale, Multigranular Statistical Image Segmentation.

Kou, Y., Lu, C.T. and Chen, D. (2003). Algorithms for spatial outlier detection. *IEEE International Conference on Data Mining*.

Li, S. (1995). Markov random field modeling. *Computer Vision.* Berlin: Springer Verlag.

Morimoto, Y. (2001). Mining frequent neighboring class sets in spatial databases. In: *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Mamoulis, N., Cao, H. and Cheung, D.W. (2005). Mining frequent spatio-temporal sequential patterns. *Fifth IEEE International Conference on Data Mining*.

Penninga, F. (2005). *3D Topographic Data Modelling: Why Rigidity Is Preferable to Pragmatism*.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. New York: Kaufmann Publishers.

Roddick, J.-F. and Spiliopoulou, M. (1999). A bibliography of temporal, spatial and spatio-temporal data mining research. *SIGKDD Explorations*, **1**(1): 34–38.

Rogers, J.P., Shine Celik, J.A. and Shekhar, S. (2006). *Discovering Emerging Spatio-Temporal Co-occurrence Patterns: A Summary of Results*.

Shekhar, S. and Chawla, S. (2003). *A tour of spatial databases*. New York: Prentice Hall.

Shekhar, S. and Huang, Y. (2001). Co-location rules mining: a summary of results. *Proc. of Spatio-temporal Symposium on Databases*.

Shekhar, S., Lu, C.T. and Zhang, P. (2003). A unified approach to detecting spatial outliers. *GeoInformatica*, **7**(2).

Shekhar, S., Schrater, P.R., Vatsavai, R.R., Wu, W. and Chawla, S. (2002). Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transaction on Multimedia*, **4**(2).

Shekhar, S., Srivastava, J., Mane, A. and Murray, C. (2005). Spatial clustering of chimpanzee locations for neighborhood identification. *Fifth IEEE International Conference on Data Mining*, pp. 773–740.

Solberg, A. H., Taxt, T. and Jain, A. (1996). A Markov random field model for classification of multisource satellite imagery. *IEEE Transaction on Geoscience and Remote Sensing*, **34**(1): 100–113.

Taylor, G. H. (xxxx). Impacts of El Nino on Southern Oscillation on the Pacific Northwest.

Tobler, W.R. (1979). In: Gale and Olsson (eds), *Cellular Geography, Philosophy in Geography*. Dordrecht: Reidel.

Vatsavai, T.E.B. and Shekhar, S. (2005). A semi-supervised learning method for remote

sensing data mining. *IEEE International Conference on Tools with Artificial Intelligence*, pp. 207–211.

Warrender, C. E. and Augusteijn, M. F. (1999). Fusion of image classifications using Bayesian techniques with Markov rand fields. *International Journal of Remote Sensing*, **20**(10): 1987–2002.

Yoo, S.S. (2004). A partial join approach for mining co-location patterns. In: *Proc. 12th International Symposium on Advances in Geographic Information Systems*.

Yoo, S.S. (2006). A join-less approach for mining spatial co-location patterns. *IEEE Transaction on Knowledge and Data Engineering*.

Zhang, P., Huang, Y., Shekhar, S. and Kumar, V. (2003). Exploiting spatial auto-correlation to efficiently process correlation-based similarity queries. In: *Proc. 8th Intl. Symp. on Spatial and Temporal Databases*.

6

Spatial Autocorrelation

Marie-Josée Fortin and Mark R.T. Dale

6.1. INTRODUCTION

Objects in natural systems (e.g., tree species in a forest) are rarely randomly distributed over space. In fact, they usually have some degree of patchiness (i.e., they are spatially clustered). Spatial aggregation of objects produces a variety of distinct spatial patterns that can be characterized by the size and shape of the aggregations, and can be quantified according to the degree of similarity between the objects in their attributes or quantitative values. These properties of spatial patterns can be indicative of the underlying processes and factors that generate and modify them through time. This is why in most disciplines (geography, economics, ecology, evolution, epidemiology, environmental science, genetics, etc.), the first step toward the understanding of phenomena is to determine whether the actual locations (coordinates) of observational data matter in explaining the spatial arrangement. So, the primary quest is to investigate and to test whether nearby objects tend to have similar attributes or to be more clustered (Figure 6.1(a)) than expected from randomness alone (Figure 6.1(b)). The presence of spatial structure in quantitative data means that similarity varies with distance between the locations and how this variation is affected by distance is known as the structure of the variable's spatial autocorrelation. In natural systems, it is the norm to have a mosaic of patches with different spatial autocorrelation structures (Figure 6.1). As spatial structures have their own intensity (magnitude) and size (extent) that make them distinct, they are usually easy to detect. In fact, it is the presence of spatial patterns that creates scale; if there were only spatial randomness around us there would be no need to determine the spatial sampling design (Delmelle, in this volume; Fortin *et al.*, 1989). We therefore need to detect the spatial patterns and determine the scope of their temporal and spatial scales.

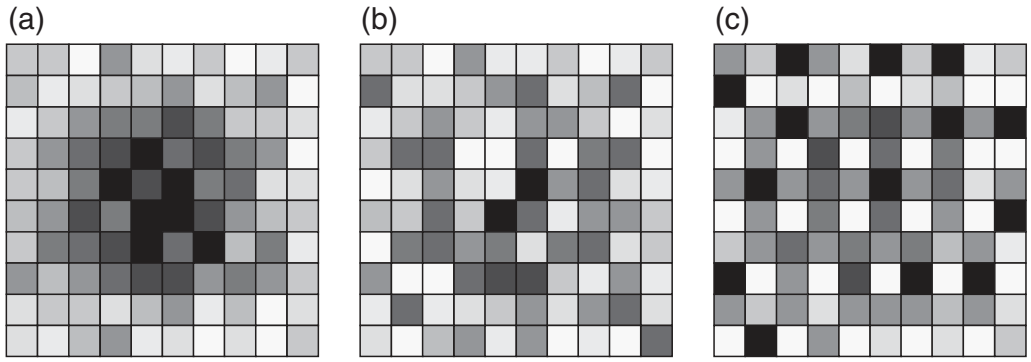


Figure 6.1 Spatial patterns. (a) Positive spatial autocorrelation where cells with similar values (gray tones) are nearby forming a patch. (b) Spatial randomness. (c) Negative spatial autocorrelation where nearby cells have dissimilar values showing spatial repulsion.

Just as there is 'no smoke without fire', there is no spatial pattern without the underlying processes that create it. Hence, spatial patterns in data can act as indicators of the processes that have occurred over a given region. If it was simple and there was a straightforward match between the spatial pattern and the generating process, we could identify and understand immediately the phenomenon under study. Of course it is not that simple. In fact, it is quite complex because several processes can take place through time each operating at a given spatial and temporal scale (Fortin and Dale, 2005; Green *et al*., 2005). Moreover, spatial patterns are shaped by a sequence of processes, all varying in duration and in intensity. Consequently, data are the end-product of an amalgam of interacting processes (Fortin and Dale, 2005; Wagner and Fortin, 2005).

The factors and processes that affect data can be coarsely divided into two kinds: those that induce spatial dependence (most spatially distributed environmental factors) and those that generate spatial autocorrelation (Figure 6.2). In our grouping of environmental factors, we include all physical initial conditions of a region based on geology, geomorphology, topography, and hydrology. Similarly, we refer loosely to processes as any events (such as disturbance, dispersal, species interactions) that change the spatial pattern or the state of the variable under study. As illustrated in Figure 6.2, seed spatial aggregation can be mainly due to their need for a specific soil type. This will be a case of spatial dependency where the seeds are patchy at the scale of the study area but not necessarily at the scale of the soil patch. Seeds' abundance at a given location i, j can be modeled by a regression function of the effects of environmental factors where the error terms ($\varepsilon_{i,j}$) are independent:

$$\begin{aligned}\text{seeds}_{i,j} &= f(\text{soil}_{i,j}, \text{moisture}_{i,j}, \text{topography}_{i,j}, \text{etc.}) \\ &\quad + \varepsilon_{i,j}.\end{aligned}$$

Non-uniform seed dispersal will also result in seed patchiness (Figure 6.2). Seed patchiness is due to the process of dispersal and refers to the degree of spatial autocorrelation of the seeds. Here, the distribution of seeds can be modeled, including the

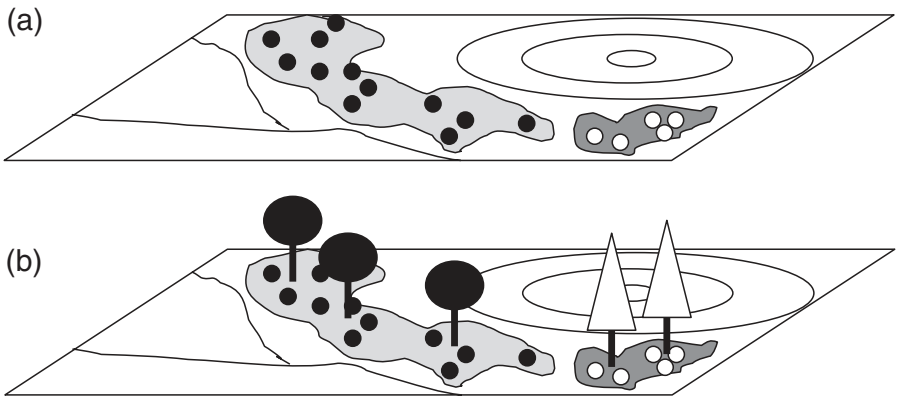


Figure 6.2 Sources of spatial structures. (a) Spatial dependence where the spatial distribution of environmental factors (here soil types A and B) constrained seeds spatial distribution. The gray polygons represent different soil types where black seeds (circles) can grow only soil type A (light gray polygon) and white seeds on soil type B (dark gray polygon). (b) Spatial autocorrelation, on top of the spatial dependence described in (a), where the seeds are dispersed by the trees.

effects of the environmental factors, and adding the effect of spatial dependence (ρ) among seed abundance values as function of distance ($d_{((i,j),(\sim i,\sim j))}$) between a location (i,j) and nearby locations $(\sim i,\sim j)$, as an autoregressive component (Anselin, in this volume; Lichstein *et al.*, 2002), and having independent errors ($\varepsilon_{i,j}$):

$$\begin{aligned}\text{seeds}_{i,j} &= f(\text{soil}_{i,j}, \text{moisture}_{i,j}, \text{topography}_{i,j}, \text{etc.}) \\ &\quad + \rho_{d((i,j),(\sim i,\sim j))}(\text{seeds}_{\sim i,\sim j}) + \varepsilon_{i,j}.\end{aligned}$$

As in the case of spatial dependence of the environmental variables, the seed dispersal process created spatial patchiness at the scale of the study area, as well as at the scale of the soil type.

Furthermore, seed abundance as a function of environmental factors and processes and the errors could include some degree of spatial dependency (Haining, 2003; Henebry, 1995):

$$\begin{aligned}\text{seeds}_{i,j} &= f(\text{soil}_{i,j}, \text{moisture}_{i,j}, \text{topography}_{i,j}, \text{etc}) \\ &\quad + (\rho_{d(i,j,\sim i,\sim j)}\varepsilon_{\sim i,\sim j} + \varepsilon_{i,j}).\end{aligned}$$

This effect results in a statistical problem because the error terms are not independent of locations. Spatially dependent errors impair the use of both parametric significance testing and randomization tests (Cliff and Ord, 1981; Fortin and Dale, 2005; Fortin and Jacquez, 2000; Haining, 2003).

Ideally, to understand natural systems we would like to be able to separate the spatial structure due to the environmental factors from that due to spatial autocorrelation generated by the processes themselves. This worthwhile task is complicated because there are feedback effects between existing spatial patterns and processes that act on them that modify both the spatial patterns and

the processes (Wagner and Fortin, 2005). These legacies of the spatial patterns on the processes (Peterson, 2002) can either promote the spread of disturbances and disease or impede animal movement (e.g., fragmentation due to roads). The result of the sequence of processes and feedbacks are included in the observed data (Haining, this volume). The question is then: At which scale should we spatial analyze the data when it is the scale itself that we want to determine?

Spatial statistics will save us, right?

No!

Spatial statistics were developed to quantify the degree of spatial aggregation (join count, Ripley's K), spatial autocorrelation (Moran's I, Geary's c) or spatial variance (semi-variance γ; see Atkinson and Lloyd, in this volume) over a study area where the mean and variance of the function describing the process are constant with distance and direction between locations. Thus, spatial statistics can quantify patterns but cannot identify their origin.

So what are spatial statistics good for?

To answer this fundamental question, we summarize first how the most commonly used spatial statistics estimate spatial patterns and spatial autocorrelation. We stress how spatial analyses of larger areas where there is more than one process impair the direct use of spatial statistics and parametric statistics. We present the statistical issues and the recent developments aiming to address them. Then, we conclude by commenting on some unresolved challenges in the field of spatial statistics.

6.2. SPATIAL STATISTICS IN A NUTSHELL

Arising from time series statistics and the more familiar parametric statistics, spatial statistics quantify the degree of self-similarity of a variable as a function of distance. These spatial statistics assume that, within the study area, the parameters of the function defining the underlying process, such as the mean and the variance, are constant regardless of the distance and direction between the sampling locations. This property of the random function is known as spatial stationarity (Cressie, 1993). Then the goal of spatial statistics is to test the null hypothesis of absence of 'spatial pattern'. For each spatial statistic 'spatial pattern' is either spatial aggregation or segregation (Ripley's K; join count statistics) or spatial autocorrelation (Moran's I and Geary's c). The null hypothesis implies that nearby locations (or attributes, measures) do not affect one another such that there is independence and spatial randomness (Figure 6.2(b)). The alternatives are that there is clustering and thus positive spatial autocorrelation (Figure 6.2(a)) or repulsion and negative spatial auotocorrelation (Figure 6.2(c)).

The mathematical commonality of the various spatial statistics is that they use the cross-product between a weighted function relating the degree of distance (w_{ij}) among the sampling locations (n) and a function (Y) quantifying the degree of similarity among the values of the variable (x_{ij}) at these sampling locations (Dale *et al.*, 2002; Getis, 1991; Getis and Ord, 1992):

$$\text{Statistic}_{(d)} = \frac{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} w_{ij}(d) Y_{ij}(x)}{C(w, d)}$$

where d is the spatial distance lag (or search window size or calculation template of radius d), between the sampling locations at which the spatial statistic is computed. The divisor of $C(w, d)$ is just a correction of the overall magnitude of the statistic calculated.

Table 6.1 Similarity functions and significance test procedures according to the spatial statistics

	Global spatial statistics	*Significance test*
Point data	Ripley's K: for each radius t, the statistic sums the indicator function $I_t(i, j)$ that counts, at each point, the number of points within a circle of radius t (Figure 6.3).	Assess using a confidence envelope based on a randomization procedure (complete spatial randomness).
	Join count statistics: The statistics count the number of links of matching J_{rr} and mismatching J_{rs} categories.	Assess by comparing the observed frequencies of links to those expected under the null hypothesis of randomness.
Polygon data	Join count: The statistics count the number of links of matching J_{rr} and mismatching J_{rs} categories.	Assess by comparing the observed frequencies of links to those expected under the null hypothesis of randomness.
Quantitative data	Moran's I: The statistic sums the deviation of the values at a given distance lag from the mean of the variable, $Y_{ij}(x) = (x_i - \bar{x})\ (x_j - \bar{x})$.	Assess using either a randomization procedure or a normal distribution approximation test where the expected value of absence of spatial autocorrelation is $E_N(I) = E_R(I) = -(n-1)^{-1}$.
	Geary's c: The statistic sums the squared deviation of the values at a given distance lag, $Y_{ij}(x) = (x_i - x_j)^2$.	Assess using either a randomization procedure or a normal distribution approximation test where the expected value of absence of spatial autocorrelation is $E_N(c) = E_R(c) = 1$.

Each spatial statistic is characterized by a particular way to determine its search window type (links, areas), size and shape (circular, square), as well as the way it calculates the degree of relationship among the values of the variables (Table 6.1; for mathematical details see Cliff and Ord, 1981; Cressie, 1993; Diggle, 1983; Epperson, 2003; Fortin and Dale, 2005; Haining, 2003; Ripley, 1981).

These spatial statistics provide, for each search window size, a single value representing the average degree of spatial autocorrelation over the study area. This is why it is important to assume spatial stationarity. This property of stationarity of the random function attributed to the process is also paramount because it allows significance testing of the null hypothesis of spatial randomness (Cliff and Ord, 1981; Cressie, 1993). For instance, by assuming normality the expected mean and variance of the spatial pattern can be estimated (Table 6.1). To overcome the issue of assumed normality, randomization tests can be used to determine the significance of the spatial pattern (Bjørnstad and Falck, 2001). This is how the significance of Ripley's K is assessed by using a complete spatial randomness procedure (Poisson process). Several researchers have suggested that complete spatial randomness is not useful as a hypothesis for comparison and have proposed other more realistic point patterns such as Poisson–Poisson or Cox–Poisson for comparison (Haining, 2003; Fortin and Dale, 2005). Other spatially restricted randomization procedures can be used for spatial statistics to reflect the spatial dynamics of the process or of the spatial distribution of the environmental factors (Goovaerts and Jacquez, 2004; Wiegand and Moloney, 2004).

To emphasize the essential property of the spatial statistics that are based on the assumption of spatial stationarity, the qualifier 'global' can be added when referring

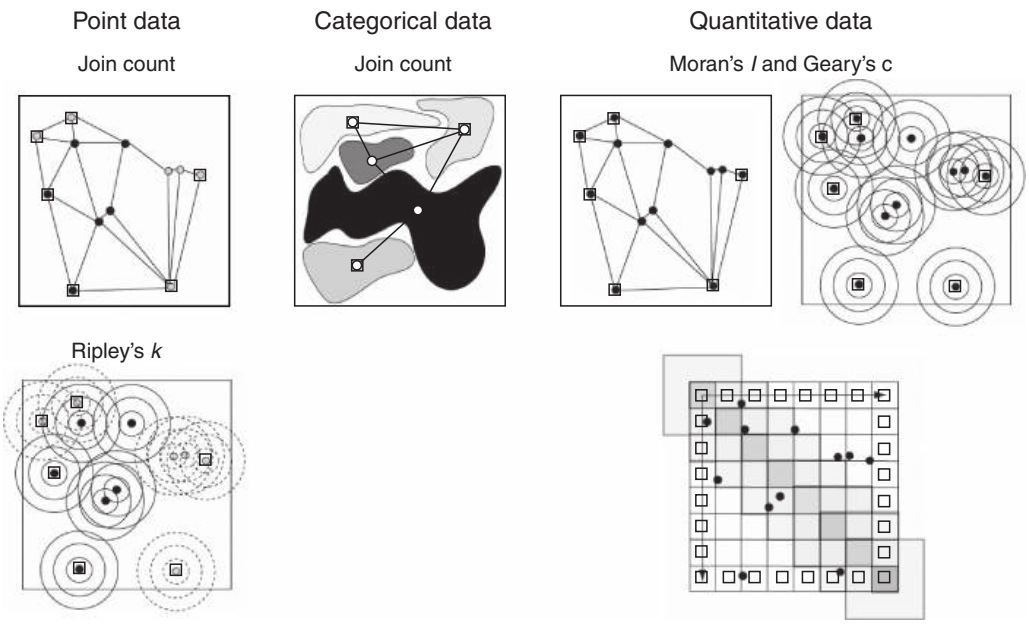


Figure 6.3 Search window types to determine the distance weight among sampling locations according to the data types and spatial statistics. For points data, the geographical coordinates of objects as well as their attributes (black or gray) need to be surveyed for the entire study area. Join count statistics require first to establish the link network among the sampling locations. Here we used a Delaunay tessellation network (Fortin and Dale, 2005; Okabe *et al.*, 2000) to determine the links. Ripley's *K* is using circles of radius *t* at each point. For categorical data from polygons, join count statistics can be used where the links are determined using the centroid of the polygons. For quantitative data, spatial autocorrelation coefficients (Moran's *I*, Geary's *c*) can be computed using either a link network among the sampling locations, a search window from each sampling location or from each cell (quadrat) from a grid (quadrats, cells).

to them. Because spatial stationarity is assumed, the shape of the search window is isotropic (Figure 6.3) and the intensity of the spatial pattern is measured as if it were the same, whatever the direction. In natural systems, this assumption is often not realistic, as water flow and wind are mostly directional processes. Such directional processes generate anisotropic patterns for which the characteristics depend on direction (Figure 6.2). Isotropic search windows are not able to detect anisotropic patterns and therefore, weights are needed to compute spatial autocorrelation according to direction as well as distance (Dubin, in this volume; Fortin and Dale, 2005). While this feature was used early on in geostatistics (Atkinson, in this volume; Journel and Huijbregts, 1978), it took longer to become common practice in other applications of spatial statistics (Oden and Sokal, 1986). The use of these directional weights still assumes that the process can occur over the entire area. It is not always the case, as in studying fish pools for example, where the spatial patterns we need to consider are only those of the aquatic network itself. In addition, proximity in an aquatic network (Figure 6.4) cannot be determined using Euclidean distance as in terrestrial systems (Figure 6.2), but rather

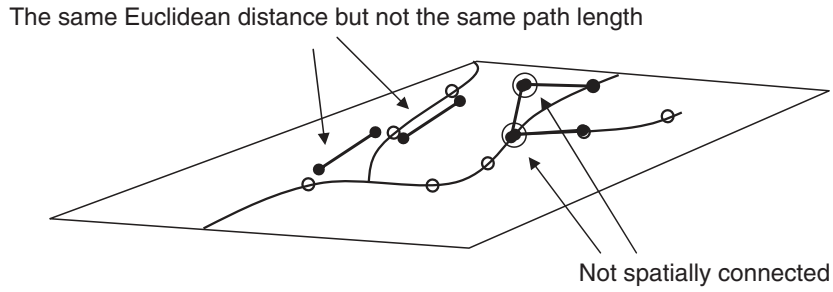


Figure 6.4 Aquatic network path length does not match the Euclidean distance among sampling locations.

requires a topological basis for proximity (Fortin and Dale, 2005; Okabe *et al*., 2000). Okabe and Yamada (2001) used such network weights to account for the particular topology of spatial networks for computing Ripley's *K*.

6.3. EFFECTS OF THE EXTENT ON GLOBAL SPATIAL STATISTICS

As we are looking for spatial patterns in natural systems, several decisions will affect our ability to detect and quantify spatial structures: how the data are gathered (Dungan *et al*., 2002; Fortin and Dale, 2005; Legendre *et al*., 2002), the size of the study area (the 'extent'), and the size of the sampling units (the 'grain'). Here we will focus on the change of extent size as it has a direct effect on the spatial stationarity of the area and the validity of the global spatial statistics. The change of grain size is also important and it is known as the modifiable area unit problem (MAUP). A whole chapter is dedicated to MAUP (Wong, in this volume), and so we refer the readers to that part of this handbook.

The extent of the study area affects our ability to detect spatial patterns: too small, and it could not include enough data to characterize the pattern; too large, and it could cover several patterns from various sources and at different scales as already mentioned above (Dungan *et al*., 2002; Fortin and Dale, 2005). Increasing the extent of the study area implies that more processes and environmental factors may alter the variable of interest. Usually, however, it is rare that we know in advance at which extent to study a phenomenon. In the absence of prior knowledge, researchers should perform a pilot study to determine it (Dungan *et al*., 2002; Legendre *et al*., 2002). Unfortunately, the wealth of available data captured by remote sensing over large areas is tempting and we often succumb to the temptation. We use all the data available to us. We 'go fishing' for spatial patterns. The problem is that the larger the extent, the more likely it is that several environmental factors and processes operate on the variable under investigation, resulting in spatial nonstationarity with the spatial patterns of several scales intermingled, or that some processes have greater effects in some sub-regions than in others. The consequence of the resulting estimation by global spatial statistics of spatial autocorrelation at various distances is that the average values of spatial autocorrelation may not reflect any spatial pattern as the

spatial structures may be cancelling out each other's signals.

Even when the extent of the study area is appropriate for the phenomena under study, our ability to determine adequately the spatial pattern can be altered due to sampling issues, statistical issues, or a combination of both. One sampling issue is the mismatch between the location of the extent and the process under study: if the actual location of the study area is a few meters north or south, it can cause the detected spatial pattern to vary (Plante *et al*., 2004). From a statistical point of view, the number of neighboring points at the edge of the study area is always smaller than at the center (as illustrated in Figure 6.3 by the sampling locations, at the centroid of patches or at the centroid of quadrats marked by squares). This edge effect is known and several edge correction algorithms have been proposed to adjust either for the edge, the corner or both (Goreaud and Pélissier, 1999; Haase, 1995; Wiegand and Moloney, 2004). Similarly, rectangular study areas will have pairs of locations at the larger distance classes only in one direction (Fortin 1999). To have a more comparable number of pairs of sampling locations to estimate spatial autocorrelation, it is recommended to use distance classes no larger than half or two thirds of the smallest side of the study area (Fortin and Dale, 2005) or to use equifrequent classes where the number of pairs is kept constant rather than the thresholds of Euclidean distances for succeeding classes (Sokal and Wartenberg, 1983).

6.4. LOCAL SPATIAL STATISTICS: ONE STEP IN A GOOD DIRECTION

The previous section presented some of the most common sampling and statistical issues that affect the reliability of global spatial statistics in estimating spatial autocorrelation and how to minimize them within the context of global analyses over the entire study area. Another approach to deal with these issues is to measure spatial autocorrelation locally using local spatial statistics (Table 6.2). Local indicators of spatial autocorrelation (or spatial association, called LISA, Anselin, 1995) measure the degree of spatial autocorrelation using, for example, Moran's I algorithm for sampling locations based only on the neighborhood around a given sampling location. The neighborhood search window can be based either on a link network or on distance classes as in the global Moran's I approach. Several variants of LISA having been developed in the same spirit of measuring local spatial association rather than autocorrelation such as the local Getis and the local Ord statistics (Boots, 2002; Fotheringham *et al*., 2000; Getis and Ord, 1996; Ord and Getis, 1995, 2001). One of the advantages of these local spatial statistics is that the values of spatial autocorrelation (or spatial association) can be mapped at each sampling location allowing the identification of sub-regions within the study area having positive (called 'hot spots') or negative (called 'cold spots') autocorrelation values (Wulder and Boots, 1998). This is very useful when large study areas are analyzed to determine how homogeneous (or not) a region is. One drawback, however, is that the significance test for each sampling location is based on the global estimate of spatial autocorrelation for the entire study area and that assumes spatial stationarity. In the absence of spatial stationarity, the advantage of using local spatial statistics over larger areas is cancelled by the lack of significance test. This is why recently researchers have been developing new procedures to assess local significance that account for the global estimate of spatial autocorrelation (Ord and Getis, 2001; Kabos

Table 6.2 Recent developments related to each kind of spatial statistics

Spatial statistics	*Developments*
Point pattern	Dale and Powell (2001): asymmetric point pattern analysis permits the detection of centers of low and high density regions (univariate) or of segregation and aggregation (bivariate or multivariate). Dixon, (2002): multivariate point pattern analysis using counts of nearest neighbors.
Ripley's *K*	Edge correction algorithms to adjust for points close to the edge and corner of the study area that have less likely to have points nearby than those at the center (Goreaud and Pélissier, 1999; Haase, 1995). Spatial network weights accounting for particular topology as in roads or aquatic stream network (Okabe and Yamada, 2001). Restricted randomization accounting for spatially heterogeneous study area (Wiegand and Moloney, 2004).
Join count	Significance test accounting for the presence of global spatial autocorrelation (Kabos and Csillag, 2002).
Global spatial statistics	Global non-parametric spatial covariance (Bjørnstad and Falck, 2001). Local indicators of spatial autocorrelation or association (Anselin, 1995). Local spatial aggregation statistics (Boots, 2002, 2003; Getis and Ord, 1992; Ord and Getis, 1995; Wulder and Boots, 1998).
Local spatial statistics	Statistics that account for the presence of global spatial autocorrelation: Ord's *O* (Ord and Getis, 2001). Local indicators of spatial autocorrelation (Anselin, 1995). Local spatial aggregation statistics (Boots, 2002, 2003; Getis and Ord, 1992, Ord and Getis, 1995, Wulder and Boots, 1998).

and Csillag, 2002). Even with these newer methods to test significance, one cannot apply a Bonferroni's correction to adjust for the multiple tests for each coefficient as for the global spatial statistics (Fortin and Dale, 2005) because the tests may be highly correlated, and there are usually too many sampling locations so often no coefficients would appear significant. However, the mapping of a local spatial coefficient value at each sampling location has been found a very informative tool for exploring the characteristics of spatial data (Fotheringham, 1997; Fotheringham and Brunsdon, 1999; Pearson, 2002; Sokal *et al*., 1998). In the same spirit of analyzing locally spatial pattern and the underlying factor or process responsible for it, geographically weighted regression can be used (see Fotheringham, in this volume).

6.5. SPATIAL AUTOCORRELATION IMPLICATIONS FOR PARAMETRIC AND RANDOMIZATION SIGNIFICANCE TESTING

One important feature of spatial dependence in data is that positive spatial autocorrelation makes parametric statistical tests too liberal, in that they produce more apparently significant results than the data actually justify. A simple intuitive explanation is that because of the lack of independence, at least some of the information of sample i is contained in adjacent samples and so instead of having the information of n independent samples, we have the information appropriate to fewer samples, n, called the 'effective sample size' (cf. Cressie, 1993). It is tempting to suggest

Table 6.3 Correction procedures for the presence of spatial autocorrelation for parametric tests

General concepts	*Correction procedure*
Parametric tests	Cressie (1993)
Univariate tests	
No general solution: model & Monte Carlo	Mizon (1995); Dale and Fortin (2002)
Bivariate tests	
Correlation	Modified *t*-test (Clifford *et al.*, 1989; corrected by Dutilleul (1993)
Linear regression	Alpargu and Dutilleul (2003)
Partial correlation	Alpargu and Dutilleul (2006)
2×2 contingency table	Cerioli (1997)
$R \times C$ contingency table	Cerioli (2002)
Cochran–Armitage	Cerioli (2003)
Multivariate	
Following Dutilleul–Cerioli approach	Speculation!

that, based on the work of Cressie and others (see below), we should be able to use the autocorrelation structure of the data in order to calculate the correct effective sample size for testing. For univariate tests, this approach does not seem to work well, and Dale and Fortin (2002) suggest the approach of modeling the data by refining a general ARMA ('Auto-Regressive Moving Average') model followed by the Monte Carlo generation of artificial 'data' sets with similar autocorrelation structure for comparison. For bivariate data, the effective sample size method seems to work well for a broad range of statistics (see Table 6.3), and we speculate that it will work for multivariate data as well.

To avoid having to deal with the estimation of the effective sample size, the use of randomization tests also seems attractive. Randomization tests (also called permutation, resampling or computer intensive tests) are convenient when the goal of the study is to assess the significance of the sample itself. When the goal is to make inferences about the sampling population, a Monte Carlo procedure should be used instead (Good, 2000). (Permutation re-orders the original data, whereas a Monte Carlo procedure produces 'new' data of similar structure.) In either case, the presence of spatial autocorrelation in the data impairs the fundamental assumption of randomization tests which is that each labeling (attributes, values) can be exchangeable randomly (Figure 6.5 (a,b)). Depending on the type of spatial autocorrelation, modified randomization procedures (or simply restricted randomization tests) can be used where the data are randomized with some specific spatial restriction. For example, in Figure 6.5(a–c), the data show marked regional differences along the south-west–north-east diagonal. With such a spatial structure, a complete randomization test cannot be used, as illustrated in Figure 6.5(a), and a restricted one is more appropriate. One way to account for this type of spatial structure is to have the study area partitioned into two regions (Figure 6.5(d)) and then the randomization is applied in each region separately. When the data show spatial dependence due to underlying environmental factors, restricted randomization procedures that generate a comparable degree of spatial autocorrelation as that observed in the data can be helpful

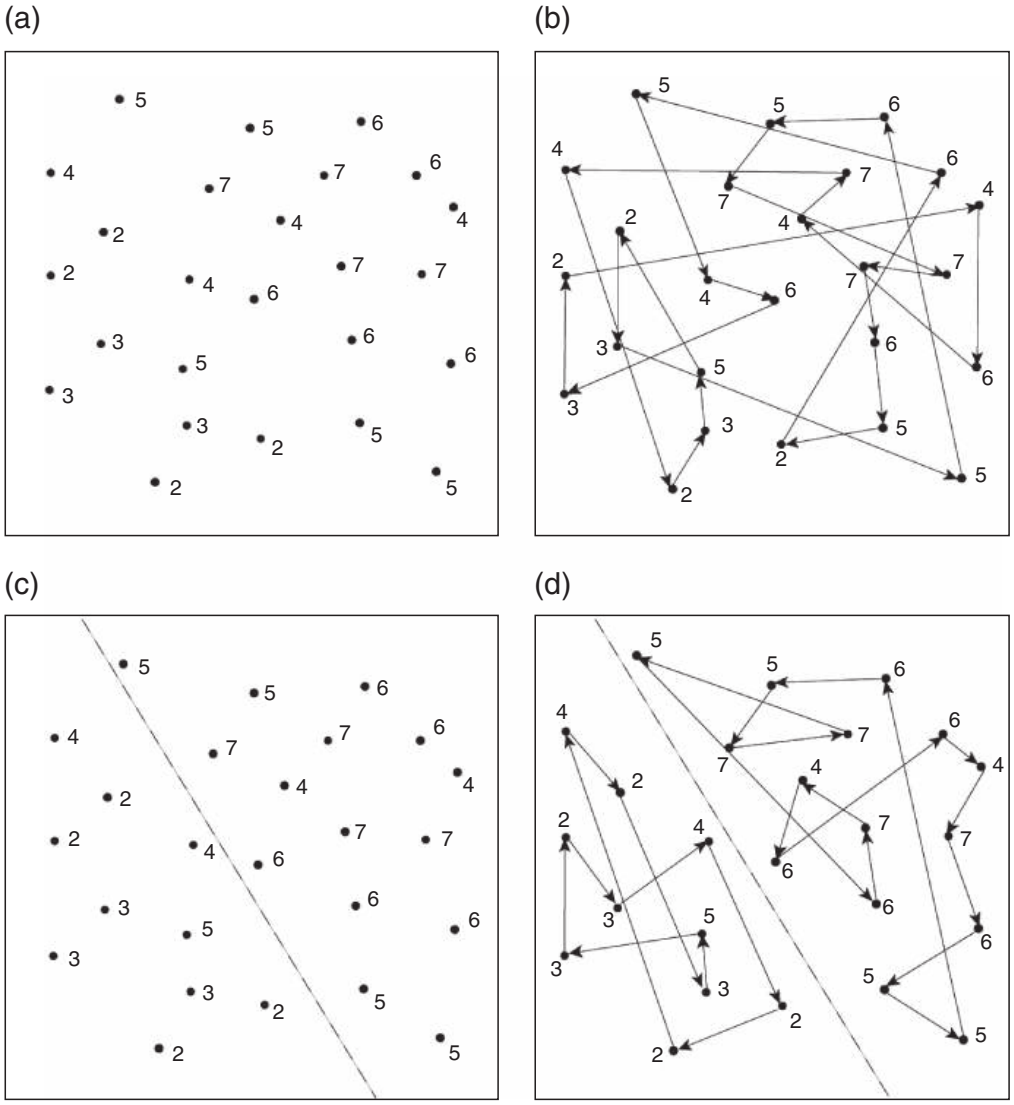


Figure 6.5 Randomization procedure. (a) Sampling locations with the quantitative values of a given variable. (b) Complete spatial randomness of the values of the variable over the sampling locations. (c) Same study area as in (a) where the dashed line delineates two sub-regions having different mean values. (d) Restricted randomization within each region.

(Fortin *et al*., 2003). To assess significance with more complex spatial patterns in which there is more than one spatial scale, Goovearts and Jacquez (2004) proposed a typology of increasing levels of spatial restrictions, that they called neutral models, to simulate more spatially realistic reference distributions.

6.6. HOW MANY SPATIAL SCALES?

A good practice to analyze larger regions involves assessing first whether the spatial patterns of the data involve more than one spatial scale, and then relate each scale to a key factor or process. This was easy to say but not so easy to do until recently. Two new

approaches have been proposed to use spatial scales as spatial predictors in regression or canonical analysis models (Borcard and Legendre, 2002; Keitt and Urban, 2005). Borcard and Legendre (2002) determined spatial predictors using principal coordinates of neighbor matrices (PCNM) that decomposed spatial scales into orthogonal spatial predictors based on the eigenvectors of the positive eigenvalues of the principal coordinates. The advantage of this method lies in the fact that neighborhoods, i.e., spatial scales, can be determined using the Euclidean distances among irregularly spaced sampling locations. Keitt and Urban (2005) used the wavelet-coefficient of the wavelet transform at each decomposition level as spatial predictor in a multiple regression model. Unlike the PCNM approach, the wavelet decomposition requires that the data are surveyed in a contiguous way as is the case with remotely sensed and GIS raster data. These new approaches have a lot of potential to determine the relative importance of environmental factors and processes in explaining the patterns of data.

6.7. NEW ERA OF SPATIAL ANALYSIS: CATEGORICAL DATA

Spatial analysis of data requires *a priori* knowledge about the data and the underlying processes. It requires as well good understanding of possibilities and limitations of the various spatial statistics available (Figure 6.6; see also Csillag and Boots, 2005). The issues presented in this chapter deal mostly with the context of spatial analysis of quantitative data. Over larger study areas, it is rare however that quantitative data are available and it is more likely that we need to rely only on qualitative data. The spatial analysis of categorical data requires often that the questions are revised (Figure 6.6) as well as the type of spatial statistical tools. GIS packages offer a series of simple spatial descriptions of qualitative data (e.g., area, number of patches) and several landscape metrics are available to refine the spatial characterization of categorical data (Gustafson, 1998). More work is still needed, however, to be able to determine the significance of these metrics so that they can be compared through time

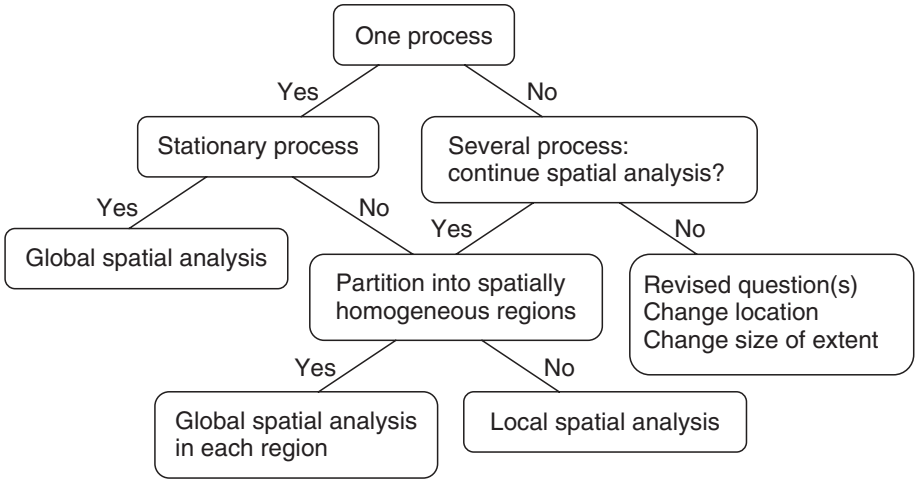


Figure 6.6 Flow chart to decide which spatial statistics to use.

and between sites (Fortin *et al*., 2003; Remmel and Csillag, 2003, 2006). As for spatial statistics for categorical data *per se*, recent methods were at the global level by assessing spatial variance using a transiogram (Weidong, 2006) and at the local level developing new local measures of spatial association (Boots, 2003). The use of mark connection functions (Stoyan and Penttinen, 2000) is also a promising area of further investigation, perhaps where the mosaic of patches is converted into a network of points with 'marks' which identify connections to first-order neighbors. Finally, there remains the large problem of incorporating time, creating a spatio-temporal analysis to assess the changes in spatial characteristics.

REFERENCES

Alpargu, G. and Dutilleul, P. (2003). To be or not to be valid in testing the significance of the slope in simple quantitative linear models with autocorrelated errors. *Journal of Statistical Computation and Simulation*, **73**: 165–180.

Alpargu, G. and Dutilleul, P. (2006). Stepwise regression in mixed quantitative linear models with autocorrelated errors. *Communications in Statistics – Simulation and Computation*, **35**: 79–104.

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis*, **27**: 93–115.

Atkinson and Lloyd (Chapter 9 – Geostatistics)

Bjørnstad, O.N. and Falck, W. (2001). Nonparametric spatial covariance functions: estimation and testing. *Environmental and Ecological Statistics*, **8**: 53–70.

Boots, B. (2002). Local measures of spatial association. *Écoscience*, **9**: 168–176.

Boots, B. (2003). Developing local measures of spatial association for categorical data. *Journal of Geographical Systems*, **5**: 139–160.

Borcard, D. and Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling,* **153**: 51–68.

Cerioli, A. (1997). Modified tests of independence in 2×2 tables with spatial data. *Biometrics*, **53**: 619–628.

Cerioli, A. (2002). Testing mutual independence between two discrete-valued spatial processes: a correction to Pearson chi-squared. *Biometrics*, **58**: 888–897.

Cerioli, A. (2003). The Cochran–Armitage trend test under spatial autocorrelation. *Proceedings of the Conference 'Complex Models and Computational Methods for Estimation and Prediction*'. Treviso, Italy, September 2003.

Cliff, A.D. and Ord, J.K. (1981). *Spatial Processes: Models and Applications*. London: Pion.

Clifford, P., Richardson S. and Hémon, D. (1989). Assessing the significance of correlation between two spatial processes. *Biometrics*, **45**: 123–134.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*, Revised Edition. New York: Wiley.

Csillag, F. and Boots, B. (2005). A framework for statistical inferential decisions in spatial pattern analysis. *The Canadian Geographer*, **49**: 172–179.

Dale, M.R.T. and Fortin, M.-J. (2002). Spatial autocorrelation and statistical tests in ecology. *Écoscience*, **9**: 162–167.

Dale, M.R.T. and Powell, R.D. (2001). A new method for characterizing point patterns in plant ecology. *Journal of Vegetation Science*, **12**: 597–608.

Dale, M.R.T., Dixon, P., Fortin, M.-J., Legendre, P., Myers, D.E. and Rosenberg, M. (2002). The conceptual and mathematical relationships among methods for spatial analysis. *Ecography*, **25**: 558–577.

Delmelle (Chapter 10 – Spatial sampling).

Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns*. London: Academic Press.

Dixon, P.M. (2002). Nearest-neighbor contingency table analysis of spatial segregation for several species. *Écoscience*, **9**: 142–151.

Dubin (Chapter 8 – Spatial Weights).

Dubin, R.A. (1998). Spatial autocorrelation: A primer. *Journal of Housing Economics*, **7**: 304–327.

Dungan, J.L., Perry, J.N., Dale, M.R.T., Legendre, P., Citron-Pousty, S., Fortin, M.-J., Jakomulska, A., Miriti, M. and Rosenberg, M.S. (2002). A balanced view of scale in spatial statistical analysis. *Ecography*, **25**: 626–640.

Dutilleul, P. (1993). Modifying the *t* test for assessing the correlation between two spatial processes. *Biometrics*, **49**: 305–314.

Epperson, B.K. (2003). Covariances among join-count spatial autocorrelation measures. *Theoretical Population Biology*, **64**: 81–87.

Fortin, M.-J. and Dale, M.R.T. (2005). *Spatial Analysis. A Guide for Ecologists*. Cambridge: Cambridge University Press.

Fortin, M.-J. and Jacquez, G.M. (2000). Randomization tests and spatially autocorrelated data. *Bulletin of the Ecological Society of America*, **81**: 201–205.

Fortin, M.-J., Boots, B., Csillag, F. and Remmel, T.K. (2003). On the role of spatial stochastic models in understanding landscape indices in ecology. *Oikos*, **102**: 203–212.

Fortin, M.-J., P. Drapeau, P. and Legendre, P. (1989). Spatial autocorrelation and sampling design. *Vegetatio*, **83**: 209–222.

Fotheringham (Chapter 13 – GWR).

Fotheringham, A.S. (1997). Trends in quantitative methods I: stressing the local. *Progress in Human Geography*, **21**: 88–96.

Fotheringham, A.S. and Brunsdon, C. (1999). Local forms of spatial analysis. *Geographical Analysis*, **31**: 340–358.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis*. London: Sage Publications.

Getis, A. (1991). Spatial interaction and spatial autocorrelation: a cross product approach. *Environment and Planning A*, **23**: 1269–1277.

Getis, A. and Ord, J.K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, **24**: 189–206.

Getis, A. and Ord, J.K. (1996). Local spatial statistics: an overview. In: Longley, P and Batty, M. (eds), *Spatial Analysis: Modelling in a GIS Environment*, pp. 261–277. Cambridge: GeoInformation International.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd edn. New York: Springer-Verlag.

Goovaerts, P. and Jacquez, G.M. (2004). Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *International Journal of Health Geographics*, **3**: 14.

Goreaud, F. and Pélissier, R. (1999). On explicit formulas of edge effect correction for Ripley's *K*-function. *Journal of Vegetation Science*, **10**: 433–438.

Green, J.L., Hastings, A., Arzberger, P., Ayala, F.J., Cottingham, K.L., Cuddington, K., Davis, F.D., Dunne, J.A., Fortin, M.-J., Gerber, L. and Neubert, M. (2005). Complexity in ecology and conservation: mathematical, statistical, and computational challenges. *BioScience*, **55**: 501–510.

Gustafson, E.J. (1998). Quantifying landscape spatial pattern: What is the state of the art? *Ecosystems*, **1**: 143–156.

Haase, P. (1995). Spatial pattern analysis in ecology based on Ripley's *K*-function: Introduction and methods of edge correction. *Journal of Vegetation Science,* **6**: 575–582.

Haining (Chapter 2 – Nature of Spatial Data).

Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.

Journel, A.G. and Huijbregts, C. (1978). *Mining Geostatistics*. London: Academia Press.

Kabos, S. and Csillag, F. (2002). The analysis of spatial association on a regular lattice by join-count statistics without the assumption of first-order homogeneity. *Computers and Geosciences*, **28**: 901–910.

Keitt, T.H. and Urban, D.L. (2005). Scale-specific inference using wavelets. *Ecology*, **86**: 2497–2504.

Legendre, P., Dale, M.R.T., Fortin, M.-J., Gurevitch, J., Hohn, M. and Myers, D.E. (2002). The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, **25**: 601–615.

Lichstein, J.W., Simons, T.R., Shriner, S.A. and Franzreb, K.E. (2002). Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**: 445–463.

Mizon, G.E. (1995). A simple message for autocorrelation correctors: don't. *Journal of Econometrics*, **69**: 267–289.

Oden, N.L. and Sokal, R.R. (1986). Directional autocorrelation: an extension of spatial correlograms to two dimensions. *Systematic Zoology*, **35**: 608–617.

Okabe, A. and Yamada, I. (2001). The *K*-function method on a network and its computational implementation. *Geographical Analysis*, **33**: 270–290.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S.N. (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd edn. Chichester: John Wiley.

Ord, J.K. and Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, **27**: 286–306.

Ord, J.K. and Getis, A. (2001). Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science*, **41**: 411–432.

Pearson, D.M. (2002). The application of local measures of spatial autocorrelation for describing pattern in north Australian landscapes. *Journal of Environmental Management*, **64**: 85–95.

Perry, J.N., Liebhold, A.M., Rosenberg, M.S., Dungan, J., Miriti, M., Jakomulska, A. and Citron-Pousty, S. (2002). Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. *Ecography*, **25**: 578–600.

Peterson, G.D. (2002). Contagious disturbance, ecological memory, and the emergence of landscape pattern. Ecosystems, **5**: 329–338.

Plante, M., Lowell, L., Potvin, F., Boots, B. and Fortin, M.-J. (2004). Studying deer habitat on Anticosti Island, Québec: Relating animal occurrences and forest map information. *Ecological Modelling*, **174**: 387–399.

Remmel, T.K. and Csillag, F. (2003). When are two landscape pattern indices significantly different? *Journal of Geographical Systems*, **5**: 331–351.

Remmel, T.K. and Csillag, F. (2006). Mutual information spectra for comparing categorical maps. *International Journal of Remote Sensing*, **27**: 1425–1452.

Ripley, B.D. (1981). *Spatial Processes*. New York: John Wiley.

Sokal, R.R., Oden, N.L. and Thomson, B.A. (1998). Local spatial autocorrelation in biological variables. *Biological Journal of the Linnean Society*, **65**: 41–62.

Sokal, R.R. and Wartenberg, D.E. (1983). A test of spatial autocorrelation using an isolation-by-distance model. *Genetics*, **105**: 219–237.

Stoyan, D. and Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statistical Science*, **15**: 61–78.

Wagner, H.H. and Fortin, M.-J. (2005). Spatial analysis of landscapes: concepts and statistics. *Ecology*, **86**: 1975–1987.

Weidong, L. (2006). Transiogram: A spatial relationship measure for categorical data. *International Journal of Geographical Information Science*, **20**: 693–699.

Wiegand, T. and Moloney, K.A. (2004). Rings, circles and null-models for point pattern analysis in ecology. *Oikos*, **104**: 209–229.

Wong (Chapter 7 – MAUP).

Wulder, M. and Boots, B. (1998). Local spatial autocorrelation characteristics of remotely sensed imagery assessed with the Getis statistic. *International Journal of Remote Sensing*, **19**: 2223–2331.

7

The Modifiable Areal Unit Problem (MAUP)

David Wong

7.1. INTRODUCTION

Geographical space is continuous in nature – there is no perfect discontinuity on the Earth's surface. Therefore, in modeling geographical space, the raster model, which depicts the Earth's surface with small grid cells, is often used to mimic the continuous nature of space as closely as possible. But the geographical space is also occupied by locations of identical characteristics (e.g., locations along a concrete pedestrian walkway) and objects or features. In the latter case, we will use geometric primitives of points, lines or arcs, and polygons to represent those objects or features either in drawings (maps) or data (vector data). In the former situation where areas have homogeneous characteristics, we often want to demarcate the areas by discrete boundaries to define homogeneous or formal regions. In both situations, boundaries are drawn, but the former situation, i.e., defining a region, is the essence of the analytical issue known as the modifiable areal unit problem (MAUP).

The term MAUP was coined by Openshaw and Taylor (1979) when they experimented with how correlation coefficient values changed when smaller areal units were aggregated to form larger areal units either hierarchically or non-hierarchically. They concluded that the correlation coefficient could carry a range of values over different levels of spatial aggregation. The source of the problem is that boundaries of areal units are often created artificially or in an *ad hoc* manner and thus can be changed. When boundaries are drawn in a different manner, analyses of data tabulated according to different boundaries will provide different results. Therefore, Openshaw and Taylor referred to this inconsistency of analytical

results due to different spatial configuration or partitioning schemes as the modifiable areal unit problem.

This chapter is intended to provide an overview of the MAUP. Although several overviews of the MAUP exist, they are dated (e.g., Openshaw, 1984; Wong, 1995). I explain the MAUP and its two sub-problems in more detail in section 7.2. While existing literature has already elaborated upon the impacts and scope of the MAUP, I provide an overview of some of its fundamental impacts in section 7.3. In section 7.4, I use simulated data and empirical datasets to illustrate the processes creating the two MAUP sub-problems. In section 7.5, I summarize the research developments pertaining to the MAUP, with emphases upon the most recent decades. Different directions in handling and searching for solutions for the MAUP are reviewed in section 7.6, and this is followed by a concluding remark.

7.2. WHAT IS THE MAUP?

The essence of the MAUP is that there are many ways to draw boundaries to demarcate space into discrete units to form multiple spatial partitioning systems. These units may serve administrative purposes, such as the counties in the U.S., or statistical or data gathering purposes, such as the census enumeration units of tracts, block groups and blocks below the county level. Although these boundaries are often drawn along some physical features (such as rivers or roads) that may serve as physical barriers separating areal units, there are multiple ways to draw those boundaries. Thus multiple datasets of the same area can be created and they will offer different descriptions of the areas and different analytical results.

But the process of 'drawing boundaries' should be interpreted beyond the literal sense. In remote sensing or raster modeling, the basic areal units are pixels or grid cells. Each pixel or cell can be regarded as a spatially discrete unit. These units can be of different sizes or resolutions. Where the edges of the pixels or cells are located is somewhat arbitrary. By shifting the grid system slightly over space or changing the size of the pixels or cells, a new dataset can be created. Thus, numerous raster-based datasets can be created and they will give us different results. Therefore, the MAUP is not limited to polygon or vector data, but also exists in raster data.

There are two dimensions through which we can partition space or draw boundaries. One is to focus on the spatial dimension by using different configurations to partition space and fixing the number of areal units to be derived in the study region. As discussed earlier, there are many ways to partition a region even if the number of areal units is kept constant. In reality, we often encounter this in the form of re-partitioning or rezoning processes. A common example is the rezoning of school districts at the local scale. In some cases, the number of schools or districts does not change. But because of changes in population distribution across the districts and/or in the capacities of school facilities (such as through school renovation or addition of structures), the school district boundaries have to be redrawn to accommodate the change. With new school boundaries, the student compositions of some schools according to the new boundaries may be different from the original ones. Therefore, data tabulated according to the old and new school boundary systems will give different results. Note that the number of districts and the population could be the same before and after the rezoning. Changes in the data occur when the population is spatially regrouped into different sub-units in the region.

Another common example is congressional redistricting. Although redistricting may not

Figure 7.1 107th Congressional Districts for Georgia; (b) 109th Congressional Districts for Georgia; (c) 109th Congressional Districts around the Atlanta Region, Georgia; (d) Census tracts, block groups and blocks of Washington, DC, 2000 Census.

be a perfect example, since the number of congressional districts is likely different after the redistricting process due to population growth, it provides a good example to illustrate the significance and magnitude of this dimension of the MAUP. Figure 7.1(a) and (b) show two maps of the 107th and 109th congressional districts (CDs) in Georgia. Figure 7.1(a) is the map of the 107th CDs and Figure 7.1(b) is for the 109th CDs. It is obvious that the two partitioning systems have very different spatial patterns, although only two districts were added in the 109th Congress for Georgia. No old district in the 107th Congress in Georgia maintained its territory in the 109th. Another obvious change is that the area around the Atlanta metropolitan area has become much more spatially fragmented to accommodate more CDs. Because of the spatial

complexity within that region, Figure 7.1(c) was provided to show the details. It is impressive how some districts have such an irregular or non-compact shape. For instance, CD 8 essentially has two sectors, very much breaking up CD 11. CD 13 seems to have several pieces scattered around the city of Atlanta and stretched outward narrowly from the city. The case of Atlanta CDs is a possible case of gerrymandering, and is a good example to illustrate how space can be partitioned in a seemingly infinite number of ways (Openshaw, 1996; Fotheringham, 2000).

When the number of areal units is fixed or relatively stable, but boundaries are redrawn to accommodate changes, this is basically a zoning process. Data gathered according to different zoning systems of the same region will give us different depictions of the region and different analytical results when the data are analyzed. The inconsistency of the results based upon data from different zoning systems is known as the *zoning problem*, one of the two sub-problems of the MAUP.

Another dimension through which we can partition space is the scale dimension. Given a study region, we can partition the region to different levels of detail. For instance, the U.S. is divided into four census regions, and each region is further subdivided into divisions, giving the entire U.S. nine divisions. Under each division are states, which are essentially political and administrative units (Wong and Lee, 2005, p. 8). Under each state, we have counties and then census tracts, block groups and census blocks. Under counties, those census units are enumeration or statistical units created for census data gathering, tabulation and dissemination purposes. But they provide information about the region at a more geographically detailed level than the state or county levels.

Note that when the U.S. is partitioned according to the levels of census geography described above (region–division–state–county–tract-block–group-block), they form a geographical hierarchy such that subdivisions at the more detailed level are found only within, but not across, the larger units involved. When other census units, such as metropolitan areas, are involved, the situation will not conform to a geographical hierarchy. Still, the general idea is that the region can be subdivided to different levels of detail or spatial resolution, as in raster data. Figure 7.1(d) offers such an example using Washington, DC, while only tracts, block groups, and blocks are shown here.

Data are available at all of these census geography levels. Census tract and census block groups data are commonly used in demographic and socioeconomic analyses. But one cannot assume that analysis results from the census tract data will be consistent with the results based on the block group data. This inconsistency due to the use of data at different geographical scales or spatial resolutions is known as the *scale problem*, the second sub-problem of the MAUP. In the next section, I will use simple examples to illustrate some fundamental inconsistencies of analytical results due to the zoning and scale problems.

7.3. FUNDAMENTAL IMPACTS OF THE MAUP

To date, most of the literature on the MAUP has been focused on the impacts of the problems. Before I provide a review of the literature, I will illustrate some simple effects of the MAUP using the Congressional Districts data of Georgia and the census data of Washington, DC.

In the Georgia example, the boundaries of congressional districts (CDs) changed quite

significantly between the 107th and 109th Congresses. The maps in Figure 7.1 show only the boundary changes without demonstrating the potential impacts on analysis due to this zoning effect. The redistricting of the 109th Congress was based upon the 2000 Census data. The 2000 Census data can also be tabulated according to the boundaries of the 107th Congressional Districts in order to assess how the rezoning affected the characteristics of the CDs. Using simple GIS procedures, some population variables of the 2000 Census were tabulated according to the 107th CD boundaries. Figure 7.2 shows the percent black in each congressional district in Georgia in the 2000 Census, according to the boundaries of the two Congresses.

The two maps show very different spatial patterns of the African–American population. The congressional district in southeast Georgia has a lower black concentration according to the 109th when compared with that in the 107th. On the other hand, CDs along the southern side of the city of Atlanta became more populated by blacks when the boundaries changed from the 107th to the 109th, while whites tended to be more numerous in CDs surrounding the outskirts of Atlanta and the northeast part of the state.

When one examines the legends of the two maps in Figure 7.2, it is easy to note that: (1) the different visual patterns of the two maps are not due to using different classification values; and (2) data tabulated according to the two CDs have different statistical distributions such as minimum and maximum values. Table 7.1 shows some of the statistics in detail. Numerically, the means from the two Congresses are different, although they are quite close. The 109th CDs have a smaller range than the 107th CDs, but the standard deviation is slightly larger. When the correlation of percent white and percent black is evaluated for the two Congresses, the correlation for the 107th

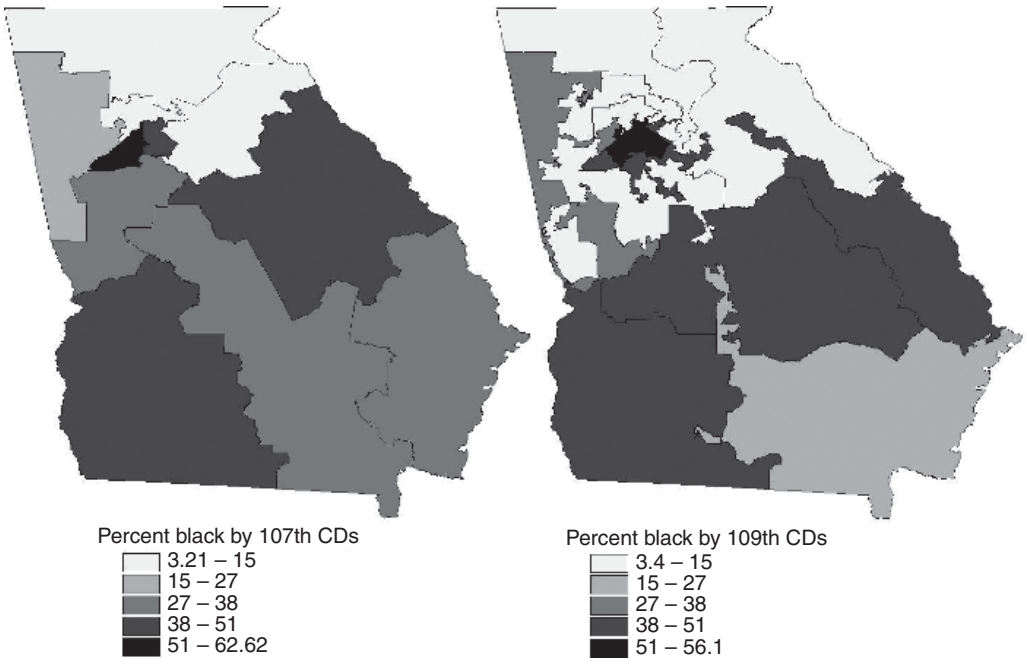


Figure 7.2 Percent blacks in 2000 Census according to the boundaries of the 107th and 109th Congressional Districts (CDs).

Table 7.1 Selected statistics for the variable percent black for the 107th and 109th CDs

Variable: percent black	*Mean*	*Minimum*	*Maximum*	*Standard deviation*
107th CDs	30.19	3.21	62.62	17.60
109th CDs	28.70	3.40	56.10	18.70

Congress was -0.987 ($p < 0.001$) and for the 109th Congress it was -0.9898 ($p < 0.001$), too close to tell that they are different. Although statistically the two means of percent black for the two Congresses are not significantly different, numerical difference in statistical values does raise some concerns about the consistency of analysis results using data tabulated according to different spatial partitioning systems. This inconsistency attributable to zonal differences is part of the impact of the zoning effect.

The most effective way to illustrate the scale effect of the MAUP is to use data at different levels of a geographical hierarchy. Figure 7.1(d) shows three levels of the census geography of the Washington, DC area. The lowest level, census block, has limited socioeconomic variables. Therefore, only census tract and block group data are used here. Figure 7.3 shows the variable per capita income (PCI) for blacks at the two census levels. The overall income distributions depicted by the two maps are quite similar – higher levels in the northwest and lower levels in the southeast. But the block group map provides refined details that are otherwise concealed at the census tract level. Some of the block groups in the western part of the region have relatively low PCI values. Because their neighboring block groups had reasonable PCI levels for blacks, the overall tract level PCI values are medium to high. Similarly, there is one small block group on the southeastern side that had a moderately high value. But because all neighboring block groups had lower PCI values, the aggregated value for that tract was relatively low.

When a small number of low value areas are surrounded by a large number of high value areas, the scale effect tends to inflate the low value areas. On the other hand, when a small number of high value areas are surrounded by a large number of low value areas, the scale effect tends to deflate the high value areas. To summarize, a general characteristic of the scale effect is to smooth out extreme values so that the range of the values is narrower. To verify this general impact, Table 7.2 shows selected statistics of the variable at the census tract and block group levels. Although the means for the two levels are not dramatically different, their maximum values and standard deviations are quite different, supporting the argument that more aggregated data (tract) tend to have less variation, since the aggregation process over scale smooths the variability.

If one follows the logic that more spatially aggregated data are less variable, and this logic is extended to analyze correlation between variables, it is not difficult to come to the conclusion that data at the higher aggregation levels will likely have higher correlation than more spatially disaggregated data. By picking two variables, per capita income for black and median house value, we can evaluate their correlation at the tract and block group levels. At the tract level, Pearson's correlation coefficient for the two variables is 0.6806 ($p < 0.001$). At the block group level, the correlation is only 0.3867 ($p < 0.001$). Apparently, the correlation at the block group level was much lower than that at the census tract level. This impact of scale effect has long been recognized in the literature for many decades (Gehlke and

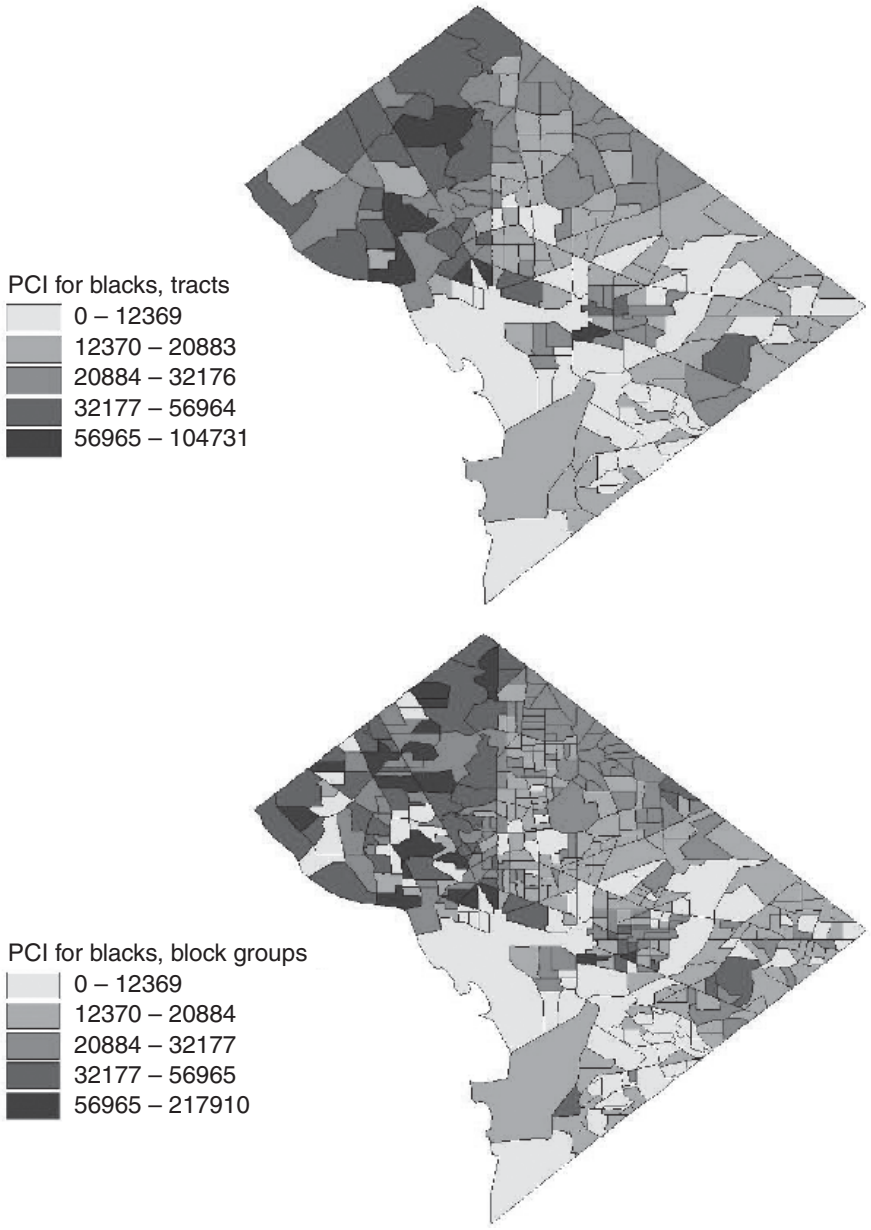


Figure 7.3 Per capita income of blacks, census tract and block group levels.

Table 7.2 Selected statistics for the variable per capita income for blacks at the census tract and block group levels

Variable: per capita income for blacks	*Mean*	*Minimum*	*Maximum*	*Standard deviation*
Census tracts	21879	0	104731	15053
Block groups	23390	0	217910	20073

Biehl, 1934; Openshaw and Taylor, 1979; Robinson, 1956). Fotheringham and Wong (1991) provided a more detailed statistical explanation to the spatial scale-variant nature of the correlation coefficient.

7.4. THE MAUP PROCESSES

7.4.1. The scale effect

The above analyses have demonstrated that the MAUP is relevant to even simple mapping. Maps are often used to explore and visualize spatial patterns. The Georgia example shows that to a large extent, the spatial pattern is a function of the partitioning system. Adopting different partitioning systems can generate different patterns for the same area, despite using the same variable. The impacts on mapping for the scale effect, in the specific example of Washington, DC, are not very dramatic. The overall pattern is quite persistent across different scales. However, it is dangerous to assume that the scale effect has minimal impacts on mapping. In fact, many experiments and studies have shown that using data from different scale levels can portray very different spatial patterns.

The impacts of the MAUP on mapping are quite obvious, but its impacts on statistical analysis seem to be quite difficult to comprehend and generalize. That is why most of the literature on the MAUP has been on assessing its impacts on different subject areas (population, urban, vegetation, soil, etc.) and on different techniques (general statistics, spatial statistics, and mathematical models). But the simple correlation analysis above using Washington, DC census tract and block group level data offers some insights on the processes related to the scale effect and its potential impacts. As smaller areal units (such as block groups) are aggregated to form larger units (such as tracts), original values of the smaller units with some level of variability are summarized or replaced by a representative value, which, in most cases, is a measure of central tendency such as the mean or median. Extreme values among the smaller units are now removed and therefore data more aggregated are becoming less varied or more similar. Thus, the correlations among variables tend to be higher with higher levels of spatial aggregation. This nature of the scale effect was best exemplified by the work of Openshaw and Taylor (1979), which shows that the correlation coefficient could carry a wide range from a moderate level for relatively disaggregated data, to a very high correlation level for highly aggregated data. Sometimes, slightly negative relationships at the individual or disaggregated level can turn into moderate positive correlations when data are aggregated into larger areal units (Fotheringham and Wong, 1991).

Although these correlation analyses are quite straightforward, their results and general patterns have significant implications for conducting general statistical and spatial analyses on data that may be tabulated and disseminated at different spatial aggregation levels. With most multivariate statistics, the relationships between variables are often summarized by the correlation matrix, or the variance–covariance matrix, and these serve as the foundation for analysis (Griffith and Amrhein, 1997). Data at higher levels of aggregation tend to inflate correlation as compared to the disaggregated levels. Therefore, we can expect that analyses using more aggregated data will likely show stronger relationships than analyses using more disaggregated data. To some extent, the process of the scale effect and its general impacts are quite predictable. Also, because the variance–covariance matrix is the core of almost all multivariate analyses, the impacts of the MAUP on this matrix are propagated to various multivariate statistical techniques

(e.g., Perle (1977) on factor analysis; Hunt and Boots (1996) on principal component analysis).

7.4.2. The zoning effect

For the zoning effect, its process and its general impacts seem to be more difficult to assess and comprehend. There are several variables acting both independently and together to determine the impacts of the zoning effect. To illustrate the roles of some of these variables, I have created two artificial landscapes (Figure 7.4a and b). Both landscapes have the same number of areal units (100) and the same set of values. For Figure 7.4(a), similar values (variable 1) tend to locate close to each other, exhibiting strong positive spatial autocorrelation, a situation quite common in reality (Odland, 1988; Griffith, 1987). Figure 7.4(b) was created by randomly shuffling the original 100 values and assigning them to different cells to create variable 2. As a result, the pattern is somewhat random. In addition, I have created two zoning patterns: the first pattern, Configuration 1 in Figure 7.4, follows closely the patterns of Variable 1 in Figure 7.4(a); the second pattern, Configuration 2 in Figure 7.4, cuts through different zones of Variable 1.

When Configuration 1 is applied to Variable 1, we expected that the general spatial trend will likely be preserved, while this

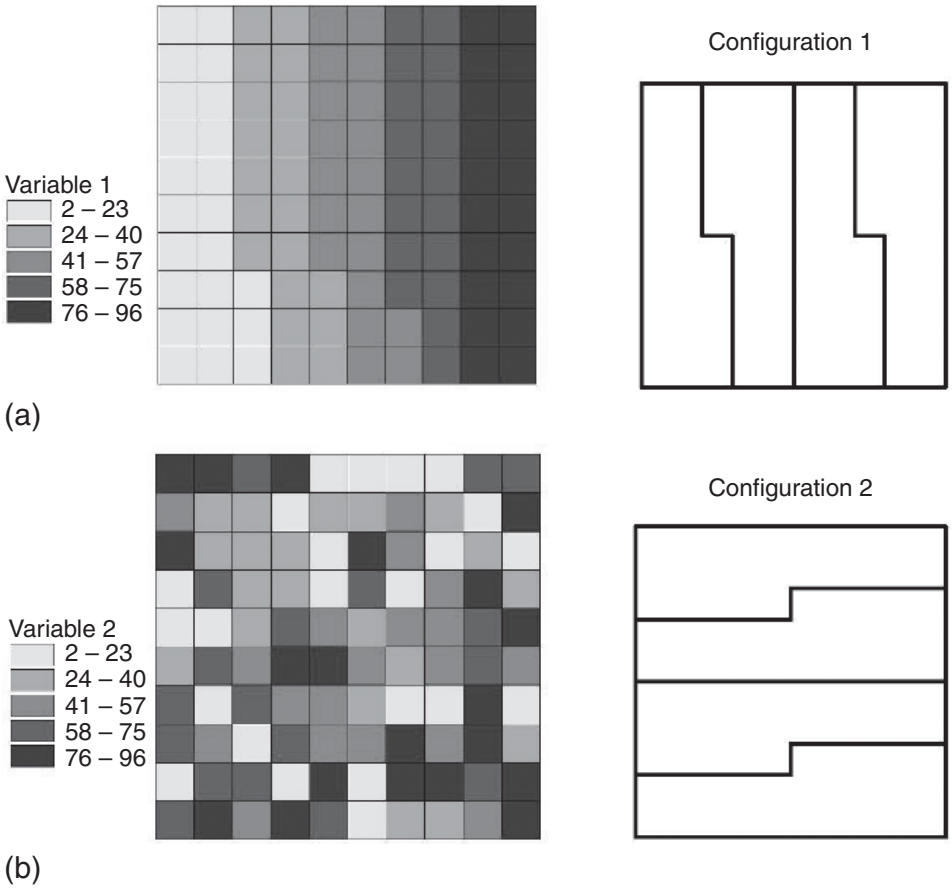


Figure 7.4 100 units with (a) positive spatial autocorrelation and (b) random pattern; and two hypothetical zoning systems: Configuration 1 and Configuration 2.

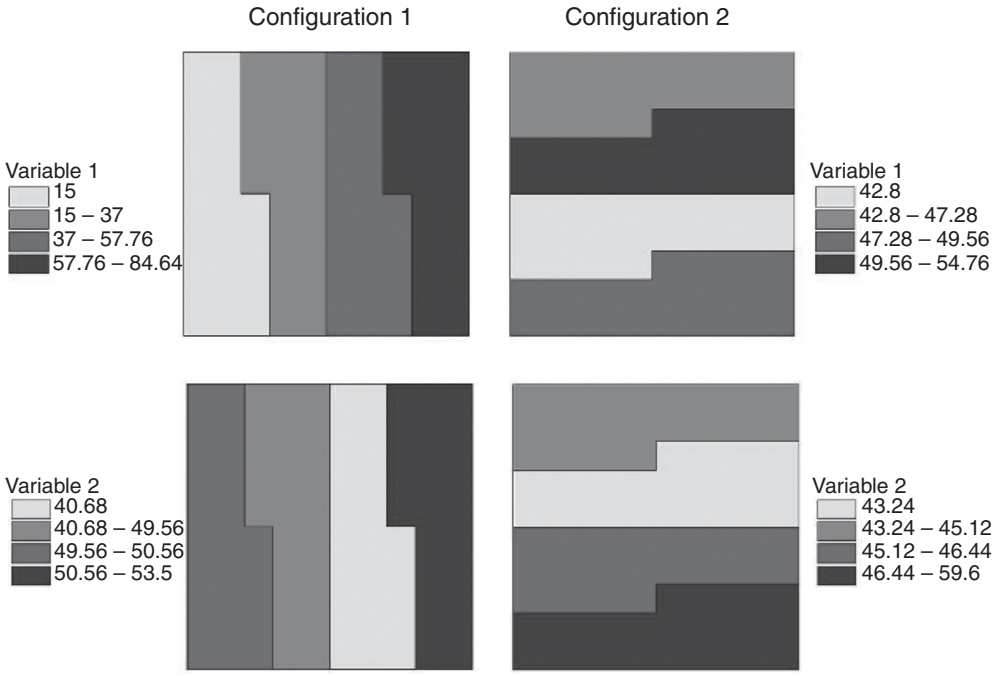


Figure 7.5 Configurations 1 and 2 applied to Variables 1 and 2.

will not be the case when Configuration 2 is applied to Variable 1. On the other hand, because the spatial distribution of Variable 2 is somewhat random, imposing Configurations 1 or 2 will unlikely create major differences. Figure 7.5 shows the results; besides Variable 1–Configuration 1, we cannot identify any pattern. Note that the ranges of values in other situations are much smaller than that in Variable 1–Configuration 1.

Table 7.3 shows the details for some of the statistics. The first row in Table 7.3 (V1 and V2) lists the statistics of the original values. Assuming that we aggregate the original 100 areal units into four larger units by taking the averages of the original values, the first batch of rows shows the results from the averaging process. The only situation that can preserve some of the statistics (standard deviation and to some extent maximum) of the original values reasonably well is V1–C1, when the spatial configuration coincides with the spatial pattern. When Configuration 2 was applied to Variable 1, the minimum value was inflated, but the standard deviation was greatly suppressed. For Variable 2, we see no obvious differences in statistics when different configurations were applied, as the values are spatially random. In other words, the zoning effect will be minimal if the phenomenon exhibits a somewhat random pattern. But if the phenomenon exhibits strong positive spatial autocorrelation, then we should expect some significant impacts due to the zoning effect.

Besides the spatial distribution of the data, another major factor in determining the impacts of the MAUP is the spatial aggregation mechanism, or the process used to derive a representative value for the aggregated units. The above example used averaging as the process, i.e., the average value of the original data will be used for the aggregated unit. But there are other possible choices for the representative values, such

Table 7.3 Selected statistics for using the two hypothetical configurations (1 and 2) to aggregate Variable 1 and Variable 2 in Figure 5

Variables (V1 and V2) and configurations (C1 and C2)	*Mean*	*Minimum*	*Maximum*	*Standard deviation*
V1 and V2	49.00	2.00	96.00	27.00
V1–C1 (Average)	48.60	15.00	84.00	29.70
V1–C2 (Average)	48.60	42.80	54.76	4.98
V2–C1 (Average)	48.60	40.68	53.60	5.55
V2–C2 (Average)	48.60	43.24	59.60	7.45
V1–C1 (Minimum)	37.00	2.00	71.00	29.00
V1–C2 (Minimum)	7.00	2.00	11.00	4.00
V2–C1 (Minimum)	5.00	2.00	10.00	3.00
V2–C2 (Minimum)	7.00	2.00	12.00	5.00

as median, minimum, maximum and others. The second batch of rows in Table 7.3 shows the aggregation results when the minimum values are taken as the representative values.

Again, applying Configuration 1 to Variable 1 best preserves the original information, but still the results are quite different from using the averaging process and the original values. Therefore, how values of sub-units are aggregated to larger units will also affect the magnitude of the MAUP. Although our discussion focuses on the zoning effect, both the spatial distribution of the data and the aggregation mechanism are also applicable to explain the scale effect.

7.5. STAGES OF THE MAUP RESEARCH

In the following section, I attempt to provide an account of MAUP research over the past several decades. To facilitate the discussion, I divide the history into two periods characterized by the major types of MAUP research appearing during those periods: discovering and assessing the impacts of the problem; and conceptualizing and formulating solutions. One needs to recognize that this division is somewhat artificial and not exclusive in nature, since their labels simply reflect the dominant types of research during those periods.

7.5.1. Discovery and impact assessment

The impacts of MAUP have been documented thoroughly. Given that changing the correlation among variables is a typical and fundamental impact of the MAUP, it is not surprising to find that most statistical analyses are subject to the MAUP. In addition, non-statistical-based mathematical models or quantitative methods are also likely impacted by the MAUP. Although Openshaw and Taylor coined the term, many researchers prior to them had documented some aspects of the MAUP. The earliest seems to be the work by Gehlke and Biehl (1934), who reported patterns of correlation coefficient changes when census tracts were grouped differently. Another early work by Robinson (1956) moved a step forward by arguing that a weighting scheme was necessary to correct the correlation coefficient to

account for different numbers of observations among areal units. While not targeted at the MAUP specifically, Moellering and Tobler (1972) offered a better understanding of the smoothing process of the scale effect by explaining how variance changes over scale levels. Sawicki (1973), and later Clark and Avery (1976), launched among the earliest attempts to assess the MAUP effects on general statistical analyses. Perle (1977) explicitly links the MAUP to the issue of ecological fallacy (Wong, 2007), although the potential problems of using ecological correlation to infer individual behavior were well documented (Robinson, 1950). Parallel to these developments, some British geographers, including Openshaw, focused on a related issue of developing optimal zonal systems, partly for regionalization purposes and partly to deal with the MAUP problem. As Batty (1976) adopted the information approach to handle spatial aggregation, others aimed at designing the best zonal systems to support spatial interaction modeling (Masser and Brown, 1975; Openshaw, 1977a, b, 1978a, b). Creating zones or regions is often needed in regional analysis, and these zones or regions provide the basis for location–allocation models. Goodchild (1979) first recognized the MAUP effect on location–allocation modeling. Mathematical modelers occasionally picked up this topic (Fotheringham *et al*., 1995; Hodgson *et al*., 1997; Murray and Gottsegen, 1997; Horner and Murray, 2002), but these studies were limited to assessing the impacts of the MAUP.

After Openshaw and Taylor coined the term MAUP in 1979, the next major concerted effort in addressing the MAUP started around 1989, partly due to the research initiative of the National Center for Geographic Information Analysis (NCGIA) on data accuracy. In between, there were intermittent developments in identifying different aspects of the MAUP. Batty continued his entropy-based approach to deal with the aggregation problem in the context of developing gravity-based models (Batty and Sikdar, 1982a–d). Putman and Chung (1989) also joined the British geographers to address zonal design issues for spatial interaction models. Blair and Miller (1983) demonstrated the impacts of MAUP on input–output models.

The formation of the NCGIA and the launching of the spatial data accuracy research initiative created a boost for the MAUP research since 1989. Fotheringham (1989) called for the recognition of scale sensitivity issues in spatial analysis, as well as the need to perform multi-scale analyses. In the same volume, Tobler (1989) argued that the MAUP is a spatial problem and therefore the solution has to be spatial in nature. Subsequently, he proposed a migration modeling framework that was not sensitive to scale changes, probably the first scale-independent spatial analytical technique to be introduced. Unrelated to the development of NCGIA, Arbia (1989) published a highly in-depth monograph addressing the MAUP.

7.5.2. From conceptualization to problem solving

With the NCGIA research initiative on spatial data accuracy as the platform, a new wave of research activities on the MAUP began in the early 1990s, starting with the paper by Fotheringham and Wong (1991), a frequently cited paper, systematically addressing the impacts of the MAUP on correlation analysis and regression models. While researchers were still interested in, and to some extent obsessed with the impacts of the MAUP, the community had gradually moved toward finding solutions to the MAUP. This search for solutions was in parallel to the effort of several researchers who had provided

evidence that the MAUP effects may not be as pervasive as some others claimed (e.g., Fotheringham and Wong, 1991). Amrhein and Flowerdew (1989) show that the MAUP has limited impact on Poisson regressions. Trying to identify when MAUP will be significant, Amrhein (1995) and Amrhein and Reynolds (1996) conducted a series of simulation, controlling for various statistical properties of the data, including various levels of spatial autocorrelation. They concluded that the MAUP effects may not be significant given certain levels of aspatial and spatial correlation among variables, but their relationships are extremely complex. While most impact analyses of the MAUP focused on statistical or mathematical modeling, some analyses were more narrowly focused on index formulations, particularly using indices to measure segregation (Wong, 1997; Wong *et al*., 1999). Besides conceptualizing the scale effect on measuring segregation, this line of research also shows that spatial measures are likely more sensitive to changing scale than aspatial measures (Wong, 2004).

A coordinated effort during this phase of the MAUP research was the publishing of a special issue of *Geographical Systems* (Wong and Amrhein, 1996). In this special issue, some researchers still focused on the MAUP effects (e.g., Okabe and Tagashira, 1996; Hunt and Boots, 1996), but others delved deeper into the sources of the MAUP (e.g., Amrhein and Reynolds), including the change-of-support concept in geostatistics (Cressie, 1996). A clear direction was to develop solutions. Holt *et al*. (1996) argued that the source of the scale effect was the changes in correlation between variables and thus they proposed a framework to model the changes of correlation over scale by taking into account spatial autocorrelation implicitly. Unfortunately, the complexity of the computational method was beyond a practical solution to the problem. Creating optimal zoning was firmly believed to be a potential solution to the MAUP in the past (Openshaw, 1977a), and this direction was still an interest at this stage of the research (Openshaw and Schmidt, 1996).

Most of the research on the MAUP mentioned above focused on aggregating polygon feature data, a popular operation in manipulating vector format data in GIS and frequently used in the handling of socioeconomic phenomena. However, the impacts of the MAUP are also present in physical geography, environmental modeling and in general, the analysis of raster format data. Outside of human geography, some landscape ecologists and physical geographers started developing an appreciation of the MAUP problems (Jelinski and Wu, 1996), and a series of research followed this direction. While Arbia *et al*. (1996) might have been the first linking the scale effect in raster or remote sensing data analysis to the MAUP explicitly, the scale effect or scale dependency issue was definitely not new to remote sensing scientists (e.g., Bian and Walsh, 1993) since remote sensing data are often available and can be tabulated easily to multiple scale levels (Bian, 1997). Part of the issue, which has historically been a problem in remote sensing analysis, is to select the resolution appropriate for the analysis (e.g., Townshend and Justice, 1988). Lam and Quattrochi (1992) reviewed several concepts related to scale and resolution, attempting to address the issue of choosing the optimal scale or resolution to analyze a particular phenomenon. Some researchers also recognized that the scale effect is essentially a change-of-support problem in geostatistics (Atkinson and Curran, 1995). The edited volume by Quattrochi and Goodchild (1997) collected papers partly focusing on the impacts of the MAUP on remote sensing, and also on modeling the scale effect and developing solutions (e.g., Bian, 1997; Xia and

Clarke, 1997). Still, no clear solutions have been identified.

Outside of the geographical literature, the MAUP attracted additional attention after the appearance of King's monograph (1997), which focused on ecological inference issues across social science disciplines, but also addressed the related MAUP. He made a bold claim that an error-bound approach can solve the scale effect, part of the MAUP and is conceptually related to the ecological fallacy problem. His claim triggered reactions from the geographic realm, and some of these reactions were aired through a series of coordinated comments (Sui, 2000), although the focus was still on the ecological fallacy issue. But geographers' responses (Fotheringham, 2000; Anselin, 2000; O'Loughlin, 2000) were not too optimistic that King's solutions can solve the ecological fallacy issue and specifically the MAUP. On the other hand, Johnston and Pattie (2001) rebuffed the claim that geographers have not spent adequate effort in dealing with the ecological fallacy by citing previous research on entropy maximization, which offers promising results in dealing with the ecological inference problem.

7.6. POTENTIAL SOLUTIONS

The recent exchanges between geographers and King raise doubt that King's solutions can solve the MAUP. Even though the early phase of the MAUP research was fascinated by the pervasiveness of the MAUP effects and overwhelmed by 'impact-analysis' type of studies, researchers have never stopped searching for solutions since the very beginning. Robinson (1956) suggested simplistic weighting methods to overcome some of the MAUP effects on correlation analysis. Tobler (1989) argued that because the MAUP is a spatial problem, solutions have to be spatial in nature. Thus, he called for the development of scale-insensitive or frame-independent spatial analytical techniques to deal with the MAUP and he employed a population migration model that was relatively insensitive to scale changes (Tobler, 1989). Tobler's migration model is one of the very few analytical tools that are relatively scale-insensitive. Another one that has demonstrated some level of stability in correlation over different scale levels is location-specific correlation analysis (Wong, 2001). But all of these potentially scale-insensitive tools have limited applications.

A popular spatial 'solution' to the MAUP even before Openshaw and Taylor coined the term was to create optimal zoning systems (Openshaw, 1977a, b, 1978a, b; Openshaw and Baxter, 1977; Openshaw and Rao, 1995; Openshaw and Schmidt, 1996). Given that most aggregation problems involve multiple variables, derivations of zonal systems have to be based upon multiple variables and multiple objectives. In general, the principle is to create zonal systems to minimize the intra-zonal variances and to maximize inter-zonal variances. But often there is no unique solution and therefore, heuristic processes seem to be quite promising (Bong and Wang, 2004).

Recently, the edited volume by Tate and Atkinson (2001) pointed to three directions of MAUP research: impacts of the scale effects, the potential of fractal analysis in dealing the scale issue and the use of geostatistics, specifically kriging and related methods such as variograms to handle and model the scale effect. Although the intended coverage of the volume included both vector and raster data, the impact assessment tended to focus more on vector data while the modeling and 'solutions' were geared more toward raster data. Fractals have a strong relationship historically with the scale effect as remote sensing data can be tabulated and analyzed at multiple scale levels and fractal geometry

is a powerful mathematical tool to handle multiscale phenomena (Lam and Quattrochi, 1992; Lam and De Cola, 1993; Pecknold *et al*., 1997; Quattrochi *et al*., 1997). The volume by Tate and Atkinson (2001) includes several papers on using fractals to handle the scale problem. But so far, although fractal analysis has been demonstrated to be effective in describing and modeling phenomena at multiple scales, it has not yet been proven to be a viable solution to the MAUP, or more specifically the scale effect.

Tate and Atkinson (2001) also suggested geostatistical analysis as a potential solution to the scale problem. Geostatistical tools, especially variograms, can identify the geographical range of spatial autocorrelation. This is an important piece of information to understand and model the scale effect. They claimed that geostatistical tools are not used to rescale the data themselves, but to rescale statistics describing the data (Atkinson and Tate, 2000). This is an interesting idea, but has not been fully validated or operationalized. More recently, following the introduction of Geographically Weighted Regression (GWR), the potential for using GWR to depict spatial heterogeneity related to the MAUP was alluded to (Fotheringham *et al*., 2000). Because a major source of the scale effect is spatial heterogeneity and GWR can model local variability reasonably well, it is believed that GWR may be more robust than other global models and less sensitive to the scale effect (Fotheringham *et al*., 2002, pp. 144–158). Still GWR cannot really be regarded as a solution to the scale effect or the MAUP.

Somewhat similar to the geostatistical approach to rescale statistics over multiple scale levels was the direction taken by a group of social statisticians (Holt *et al*., 1996; Steel and Holt, 1996; Tranmer and Steel, 1998). They realized that the scale effect can be kept to a minimum when the aggregated areas have a high degree of internal homogeneity (low variance), and the magnitude of the scale effect will be partly a function of the internal homogeneity. As a result, one may model the scale effect or statistics describing the data at different scale levels as long as we can establish the rules of aggregation and how the scale effect is related to the level of internal homogeneity. Since the foundation of most classical statistics is the variance–covariance matrix, this group of researchers proposed using the correlation at the individual level to estimate the correlation at the aggregated level and thus can estimate the variance–covariance matrix at the aggregate level. The statistical derivations involved were very sophisticated and the computation was very demanding. As a result, this has not been a practical solution to the MAUP.

Although tremendous efforts have been spent to deal with the scale problem, to many researchers, the zoning problem seems to be easier to handle. Flowerdew and Green (1989, 1992) treated the zoning problem in the same way as resolving incompatible zonal systems. The general approach is to use spatial interpolation methods to transform data gathered according to one zonal pattern to another pattern. Fisher and Langford (1995, 1996) have evaluated the reliability of this technique in handling the zoning problem. A related technique, dasymmetric mapping, was also shown to be effective to handle incompatible zonal patterns from a cartographic perspective (Fisher and Langford, 1996; Mennis, 2003). An older smoothing or interpolation technique, the smooth pycnophylactic interpolation introduced by Tobler (1979), has also been revisited and is believed to be a solution candidate for the MAUP, specifically in addressing the problem from the change-of-support perspective (Gotway and Young, 2002).

To summarize, the MAUP effects can possibly be tackled by sophisticated models and computationally intensive techniques,

while their practical and operational potentials are yet to be affirmed. Relatively simple techniques can handle the zoning problem, but not the scale problem. Thus, without generally feasible methods to handle the MAUP, the old call for recognizing the MAUP is still the most affordable approach to deal with this long-term stubborn problem (Fotheringham, 1989). Given the advances in GIS technology and computational tools, and the availability of digital data at various scales, repeating the same analysis but using different scales or partitioning schemes is within reach of most researchers. This approach is probably the minimum standard in handling the MAUP given where we are on this topic.

Taking one step further, using segregation indices as examples, Wong (2003) disaggregated segregation at different geographical levels to demonstrate that one can document the sources of the MAUP effects. This accounting framework is to identify and quantify the amount of the MAUP effects contributed by different locations at different scale levels. This detailed mapping of the MAUP effects by scale and space is not just informative, but also sheds light on where the MAUP effects may be the most acute in the geographic hierarchy and highlights locations that deserve more attention.

7.7. CONCLUDING REMARK

Many methodological or technical problems can be found in the geographical literature. Some have broad impacts and are very complex, while some are confined to certain areas and are more manageable. Two very stubborn but pervasive problems in statistical analysis of spatial data are spatial autocorrelation and the MAUP. The past two decades of research in spatial statistics and spatial econometrics have moved the field forward to the stage that some very promising and operational modeling techniques are available to handle spatial autocorrelation quite effectively (e.g., Griffith, 2003). For the MAUP, we have accumulated pieces of knowledge and developed some comprehensive understanding and conceptualizations of the problems. But a systematic research agenda seems to be needed in order to bring significant advancements along this direction. Assessing the impacts of the MAUP should be a topic confined to the past, and the future should focus on developing operational solutions.

REFERENCES

Amrhein, C.G. (1995). Searching for the elusive aggregation effect: evidence from statistical simulations. *Environment and Planning A*, **27**: 105–119.

Amrhein, C.G. and Flowerdew, R. (1989). The effect of data aggregation on a Poisson regression model of Canadian migration. In: Goodchild, M.F. and Gopal, S. (eds), *Accuracy of Spatial Databases*, pp. 229–238. London: Taylor and Francis.

Amrhein, C.G. and Reynolds, H. (1996). Using spatial statistics to assess aggregation effects. *Geographical Systems*, **3**(2/3): 143–158.

Anselin, L. (2000). The alchemy of statistics, or creating data where no data exist. *Annals, Association of American Geographers*, **90**(3): 586–592.

Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.

Arbia, G., Benedetti, R. and Espa, G. (1996). Effects of the MAUP on image classification. *Geographical Systems*, **3**(2/3): 123–141.

Atkinson, P.M. and Curran, P.J. (1995). Defining an optimal size of support for remote sensing investigations. *IEEE Transactions on Geosciences and Remote Sensing*, **33**(3): 768–776.

Atkinson, P.M. and Tate, N.J. (2000). Spatial scale problems and geostatistical solutions: a review. *The Professional Geographer*, **52**(4): 607–623.

Batty, M. (1976). Entropy in spatial aggregation. *Geographical Analysis*, **8**: 1–21.

Batty, M. and Sikdar, P.K. (1982a). Spatial aggregation in gravity models: 1. An information-theoretic framework. *Environment and Planning A*, **14**: 377–405.

Batty, M. and Sikdar, P.K. (1982b). Spatial aggregation in gravity models: 2. One-dimensional population density models. *Environment and Planning A*, **14**: 525–553.

Batty, M. and Sikdar, P.K. (1982c). Spatial aggregation in gravity models: 3. Two-dimensional trip distribution and location models. *Environment and Planning A*, **14**: 629–658.

Batty, M. and Sikdar, P.K. (1982d). Spatial aggregation in gravity models: 4. Generalisations and large-scale applications. *Environment and Planning A*, **14**: 795–822.

Blair, P. and Miller, R.E. (1983). Spatial aggregation in multiregional input–output models. *Environment and Planning A*, **15**: 187–206.

Bian, L. (1997). Multiscale nature of spatial data in scalling up environment models. In: Quattrochi, D.A. and Goodchild, M.F. (eds). *Scale in Remote Sensing and GIS.* pp. 13–27. Lewis Publishers.

Bian, L. and Walsh, S. (1993). Scale dependencies of vegetation and topography in a mountainous environment of Montana. *The Professional Geographer*, **45**(1): 1–11.

Bong, C.W. and Wang, Y.C. (2004). A multiobjective hybrid metaheuristic approach for GIS-based spatial zoning model. *Journal of Mathematical Modelling and Algorithms*, **3**: 245–261.

Clark, W.A.V. and Avery, K.L. (1976). The effects of data aggregation in statistical analysis. *Geographical Analysis*, **8**: 428–438.

Cressie, N. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems*, **3**(2/3): 159–180.

Fisher, P.F. and Langford, M. (1995). Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, **27**(2): 211–224.

Fisher, P.F. and Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery: a study of areal interpolation by dasymetric mapping. *The Professional Geographer*, **48**(3): 299–309.

Flowerdew, R. and Green, M. (1989). Statistical methods for inference between incompatible zonal systems. In: Goodchild, M. and Gopal, S. (eds), *The Accuracy of Spatial Data Bases*, pp. 239–247. London: Taylor and Francis.

Flowerdew, R. and Green, M. (1992). Developments in areal interpolating methods and GIS. *Annals of Regional Science*, **26**: 67–78.

Fotheringham, A.S. (1989). Scale-independent spatial analysis. In: Goodchild, M. and Gopal, S. (eds), *Accuracy of Spatial Databases*. pp. 221–228. London: Taylor and Francis.

Fotheringham, A.S. (2000). A bluffer's guide to a solution to the ecological inference problem. *Annals, Association of American Geographers*, **90**(3): 582–586.

Fotheringham, A.S., Brunsdon, C. and Charlton, M.E. (2000). *Quantitative Geography: Perspectives on Spatial Analysis*. London: Sage.

Fotheringham, A.S., Brunsdon, C. and Charlton, M.E. (2002). *Geographically Weighted Regression*. England: Wiley & Sons.

Fotheringham A.S., Densham P.J. and Curtis A. (1995). The zone definition problem in location-allocation modeling. *Geographical Analysis*, **27**: 60–77.

Fotheringham A.S. and Wong, D.W.S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, **23**: 1025–1044.

Gehlke, C.E. and Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, **29**: 169–170.

Gotway, C.A. and Young, L.J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, **97**: 632–648.

Griffith, D.A. (1987). *Spatial Autocorrelation: a Primer*. Washington, D.C.: Association of American Geographers.

Griffith, D.A. (2003). *Spatial Autocorrelation and Spatial Filtering*. Berlin: Springer-Verlag.

Griffith, D.A. and Amrhein, C.G. (1997). *Multivariate Statistical Analysis for Geographers*. Upper Saddle River, NJ: Prentice Hall.

Goodchild, M.F. (1979). Aggregation problem in location-allocation. *Geographical Analysis*, **11**: 240–255.

Hodgson, M.J., Shmulevitz, F. and Körkel, M. (1997). Aggregation error effects on the discrete-space *p*-median model: the case of Edmonton, Canada. *Canadian Geographer*, **41**: 415–429.

Holt, D., Steel, D.G. and Tranmer, M. (1996). Area homogeneity and the Modifiable Areal Unit Problem. *Geographical Systems*, **3**(2/3): 181–200.

Horner, M.W. and Murray, A.T. (2002). Excess commuting and the modifiable areal unit problem. *Urban Studies*, **39**: 131–139.

Hunt, L. and Boots, B.N. (1996). MAUP effects in the principal axis factoring technique. *Geographical Systems*, **3**(2/3): 101–122.

Jelinski, D.E. and Wu, J. (1996). The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology*, **11**: 129–140.

Johnston, R. and Pattie, C. (2001). On geographers and ecological inference. *Annals, Association of American Geographers*, **91**(2): 281–282.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

Lam, N.S.-N. and De Cola, L. (eds) (1993). *Fractals in Geography*. Englewood Cliffs, NJ: Prentice-Hall.

Lam, N.S.-N. and Quattrochi, D.A. (1992). On the issues of scale, resolution, and fractal analysis in the mapping sciences. *The Professional Geographer*, **44**(1): 88–98.

Masser, I. and Brown P.J.B. (1975). Hierarchical aggregation procedures for interaction data. *Environment and Planning A*, **7**: 509–523.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer*, **55**(1): 31–42.

Moellering, H. and Tobler, W.R. (1972). Geographical variances. *Geographical Analysis*, **4**: 34–42.

Murray, A. and Gottsegen, J. (1997). The influence of data aggregation on the stability of *p*-median location model solutions. *Geographical Analysis*, **29**: 200–213.

Odland, J. (1988). *Spatial Autocorrelation*. London: Sage.

Okabe, A. and Tagashira, N. (1996). Spatial aggregation bias in a regression model containing a distance variable. *Geographical Systems*, **3**(2/3): 77–99.

O'Loughlin, J. (2000). Can King's ecological inference method answer a social scientific puzzle: who voted for the Nazi party in Weimar Germany? *Annals, Association of American Geographers*, **90**(3): 592–601.

Openshaw, S. (1977a). Geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling. *Transactions of the Institute of British Geographers*, **2**: 459–472.

Openshaw, S. (1977b). Optimal zoning systems for spatial interaction models. *Environment and Planning A*, **9**: 169–184.

Openshaw, S. (1978a). An optimal zoning approach to the study of spatially aggregated data. In: Masser, I. and Brown, P.J.B. (eds), *Spatial Representation and Spatial Interaction*, pp. 93–113. Leiden: Martinus Nijhoff.

Openshaw, S. (1978b). Empirical-study of some zone-design criteria. *Environment and Planning A*, **10**: 781–794.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem. CATMOG, 38*. Norwich, England: Geobooks.

Openshaw, S. (1996). Developing GIS-relevant zone-based spatial analysis methods. In Longley, P. and Batty, M. (eds), *Spatial Analysis: Modelling in a GIS Environment*, pp. 55–73. Cambridge, U.K.: GeoInformation International.

Openshaw, S. and Baxter, R.S. (1977). Algorithm 3 – procedure to generate pseudo-random aggregations of *n*-zones into *m*-zones, where *m* is less than *n*, *Environment and Planning A*, **9**: 1423–1428.

Openshaw, S. and Rao L. (1995). Algorithms for reengineering 1991 census geography. *Environment and Planning A*, **27**: 425–446.

Openshaw, S. and Schmidt, J. (1996). Parallel simulated annealing and genetic algorithms for re-engineering zoning systems. *Geographical Systems*, **3**(2/3): 201–220.

Openshaw, S. and Taylor, P.J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley, N. (ed.), *Statistical Applications in the Spatial Sciences*, pp. 127–144. London: Pion.

Pecknold, S., Lovejoy, S., Schertzer, D. and Hooge, C. (1997). Multifractals and resolution dependence of remotely sensed data: GSI to GIS. In: Quattrochi, D.A. and Goodchild, M.F. (eds), *Scale in Remote Sensing and GIS*, pp. 361–394. Lewis Publishers.

Perle, E.D. (1977). Scale changes and impacts on factorial ecology structures. *Environment and Planning A*, **9**: 549–558.

Putman, S.H. and Chung, S.H. (1989). Effects of spatial system-design on spatial interaction models.1. the spatial system definition problem. *Environment and Planning A*, **21**: 27–46.

Quattrochi, D.A. and Goodchild, M.F. (eds) (1997). *Scale in Remote Sensing and GIS*. Lewis Publishers.

Quattrochi, D.A., Lam, N.S.-N., Qiu, H-L. and Zhao, W. (1997). ICAMS: A geographic information system for the characterization and modeling of multiscale remote sensing data. In: Quattrochi, D.A. and Goodchild, M.F. (eds), *Scale in Remote Sensing and GIS*, pp. 295–308. Lewis Publishers.

Robinson, A.H. (1956). The necessity of weighting values in correlation analysis of areal data. *Annals, the Association of American Geographers*, **46**: 233–236.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, **15**: 351–357.

Sawicki, D.S. (1973). Studies of aggregated areal data – problems of statistical inference. *Land Economics*, **49**: 109–114.

Steel, D.G. and Holt, D. (1996). Rules for random aggregation. *Environment and Planning A*, **28**: 957–978.

Sui, D. (2000). New directions in ecological inference: an introduction. *Annals, Association of American Geographers*, **90**(3): 579–582.

Tate, N.J. and Atkinson, P.M. (eds) (2001). *Modelling Scale in Geographical Information Sciences*. London: Wiley & Sons.

Tobler, W. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, **74**: 519–536.

Tobler, W. (1989). Frame independent spatial analysis. In: Goodchild, M. and Gopal, S. (eds), *The Accuracy of Spatial Data Bases*, pp. 115–122. London: Taylor and Francis.

Townshend, J.R.G. and Justice, C.O. (1988). Selecting the spatial resolution of satellite sensors required for global monitoring of land transformation. *International Journal of Remote Sensing*, **9**: 187–236.

Tranmer, M. and Steel, D.G. (1998). Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A*, **30**: 817–831.

Wong, D.W.S. (1995). Aggregation effects in geo-referenced data. In: Arlinghaus, S.L. and Griffith, D.A. (eds), *Practical Handbook of Spatial Statistics*, pp. 83–106. Boca Raton, FL: CRC Press.

Wong, D.W.S. (1997). Spatial dependency of segregation indices. *The Canadian Geographer*, **41**(2): 128–136.

Wong, D.W.S. (2001). Location-specific cumulative distribution function (LSCDF): an alternative to spatial correlation analysis. *Geographical Analysis*, **33**(1): 76–93.

Wong, D.W.S. (2003). Spatial decomposition of segregation indices: a framework toward measuring segregation at multiple levels. *Geographical Analysis*, **35**(3): 179–194.

Wong, D.W.S. (2004). Comparing traditional and spatial segregation measures: a spatial scale perspective, *Urban Geography*, **25**(1): 66–82.

Wong, D.W.S. (2007). Ecological fallacy, In: B. Warf (ed), *Encyclopedia of Human Geography*, Sage Publications, pp. 117–118.

Wong, D.W.S. and Amrhein, C.G. (1996). Research on the MAUP: old wine in a new bottle or real breakthrough? *Geographical Systems*, **3**(2/3): 73–77.

Wong, D.W.S., Lasus, H. and Falk, R.F. (1999). Exploring the variability of segregation index *D* with scale and zonal systems: an analysis of thirty US cities. *Environment and Planning A*, **31**: 507–522.

Wong, D.W.S. and Lee, J. (2005). *Statistical Analysis of Geographic Information.* New York: Wiley and Sons.

Xia, Z-G. and Clarke, K.C. (1997). Approches to scaling of geo-spatial data, In: Quattrochi, D.A. and Goodchild, M.F. (eds), *Scale in Remote Sensing and GIS*, pp. 309–360. Lewis Publishers.

8

Spatial Weights

Robin Dubin

8.1. WEIGHT MATRICES

A weight matrix summarizes the spatial relationships in the data. In particular, the i'th row of a weight matrix shows observation i's relationship to all of the other observations. By convention, the main diagonal of this matrix consists of zeros. Because the weight matrix shows the relationships between all of the observations, its dimension is always $N \times N$, where N is the number of observations. In most applications, the weight matrix itself is treated as exogenous; that is, it is assumed that the researcher knows how the observations are related to each other. Note that the space in which the observations are located need not be geographic; any type of space is acceptable, as long as the researcher can specify the spatial interactions.

Spatial data can appear in many forms. The data can come from regions (e.g., counties) or points (e.g., houses). The data may be located on a regular grid or lattice, but this is not necessary. The numbers in the weight matrix can indicate whether or not a relationship is present or they can indicate the strength of the relationship. The former weighting schemes are called discrete, and the latter, continuous.

It is common, but not necessary, to row normalize the weight matrix. This means that the matrix is transformed so that each of the rows sums to one. Row normalizing gives the weight matrix some nice theoretical properties. For example, row normalizing allows the weight matrices from different weighting schemes to be compared, since all elements must lie between 0 and 1 (inclusive). Row normalizing also allows λ (a parameter discussed later in the chapter) to be bounded by -1 and 1. All of the weight matrices presented in this chapter will be row normalized. The cost of row normalizing is that it may interfere with the interpretation of the weights. For example, in the case of inverse distance weighting,

row normalizing will change the weights so that they sum to one. Thus pairs with the same separation distance can have different weights, depending on the number of nearby observations.

In the remainder of this chapter, I will explore weight matrices for the following cases: regular lattice data for points, regular lattice data for areas, irregularly located data for points, and irregularly located data for areas.

Consider the data presented in Figure 8.1. This is a map of 25 regions arranged on a regular lattice. The borders of the regions are shown with solid lines, the centroids are shown with heavy black points, and the lattice itself is shown with dashed lines. Each region is identified by a number between 1 and 25.

The most natural way to represent the spatial relationships with areal data is through the concept of contiguity. That is, regions will be considered to be related if their boundaries share common points. There are three types of contiguity that are commonly considered: rook contiguity, bishop contiguity, and queen contiguity. Contiguity is determined by imagining that the regions form a chess board; neighbors are determined by the regions that the appropriate chess piece could reach.

8.1.1. Rook contiguity

With rook contiguity, the neighbors are due north, south, east and west. Region 7's neighbors are regions 2, 6, 8 and 12 and

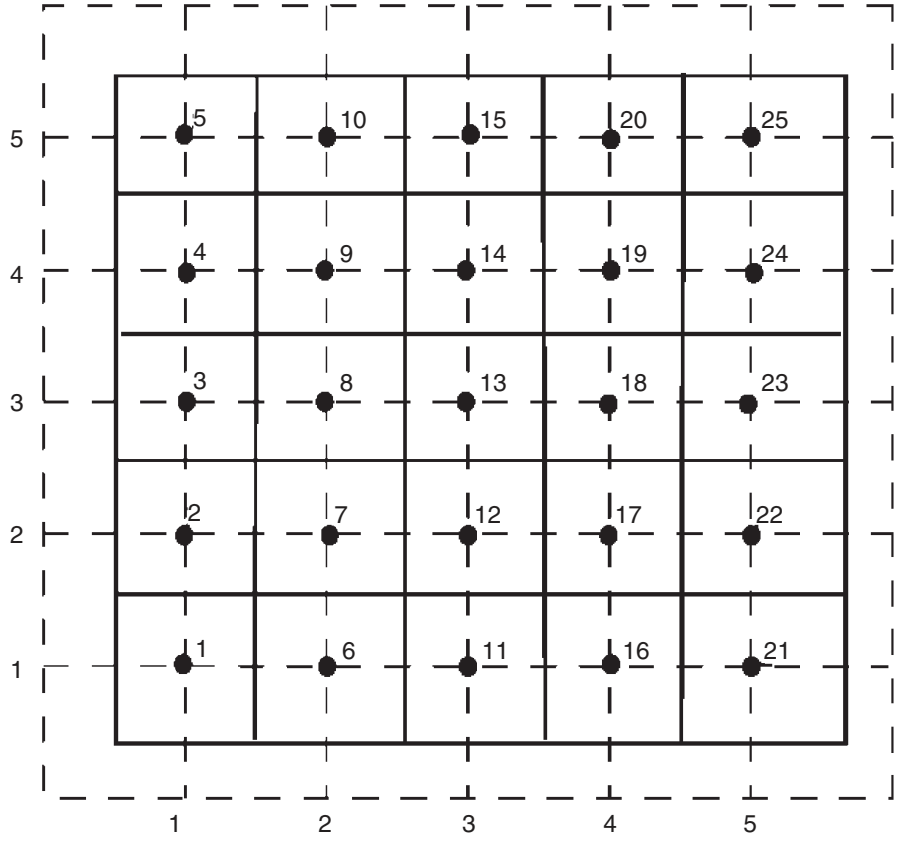


Figure 8.1 Map of regular lattice areas.

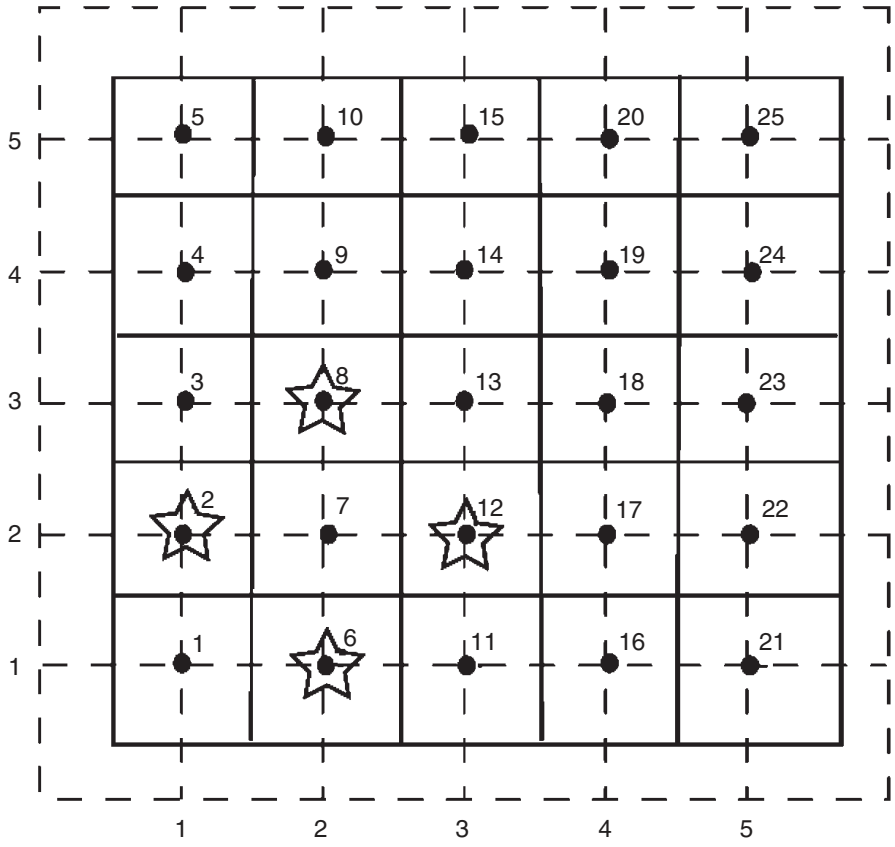


Figure 8.2 Neighbors in rook contiguity.

are indicated with stars in Figure 8.2. The weight matrix for this data will have 25 rows and columns. The first 10 rows and columns are of the unstandardized weight matrix are shown in Figure 8.3.

This symmetric matrix has zeros on its main diagonal. A one indicates that regions i and j are neighbors. Regions in the interior of the study area will have four ones in their rows. For example, the seventh row of the weight matrix contains four ones, because region 7 has four neighbors (only three are shown in Figure 8.3 because the fourth neighbor is region 12). Regions on the periphery will have fewer neighbors. For example, the first row (representing region 1) has only two ones. These are in the second and sixth cells, indicating that region 1 has only two neighbors: region 2 and region 6.

To obtain the row normalized version of this weight matrix, divide each row by the number of neighbors (ones). Thus in rows with 4 neighbors, the entries will be 0.25, and in rows with only two, the entries will be 0.5. This is a common occurrence: row normalizing often makes symmetric weight matrices asymmetric.

8.1.2. Bishop contiguity

In bishop contiguity, region i's neighbors are located at its corners. Figure 8.4 shows the neighbors for region 7 under this scheme. The neighbors are regions 1, 3, 11 and 13.

Again, regions in the interior will have four neighbors, while those on the periphery will have fewer. Figure 8.5 shows the first 10 rows

0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
1.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00
0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
0.00	1.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00
0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	1.00
0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00

Figure 8.3 Subset of unstandardized weight matrix for rook contiguity.

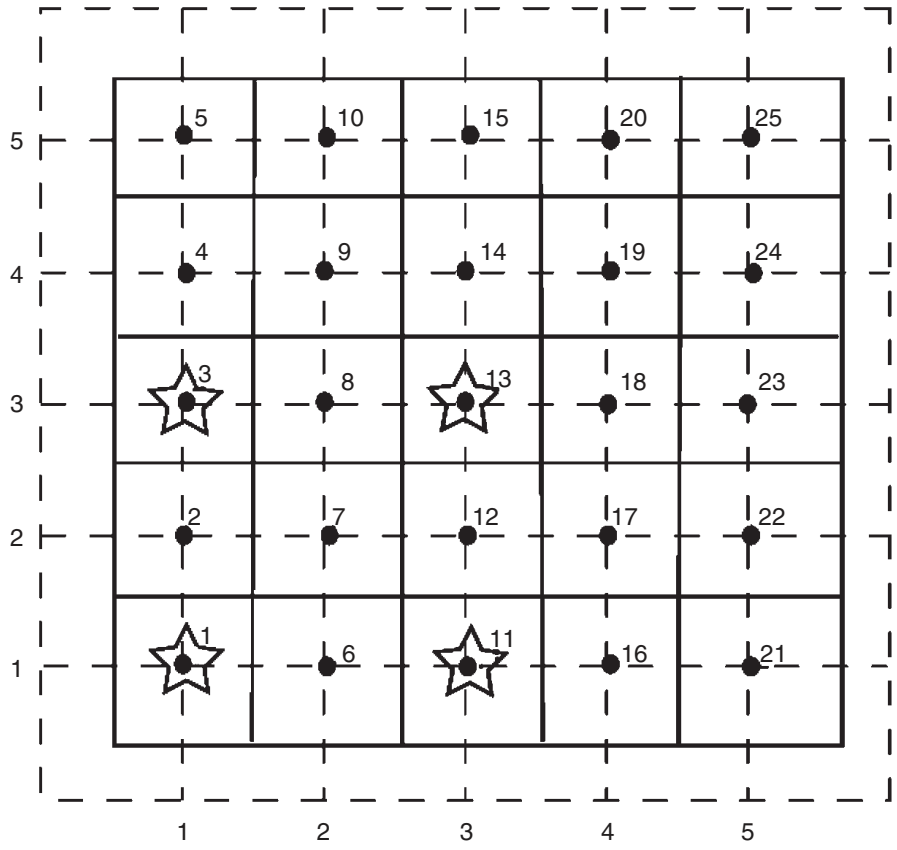


Figure 8.4 Neighbors in bishop contiguity.

and columns of the unstandardized weight matrix for this case.

Examination of the first row of Figure 8.5 shows a 1 in position 7, indicating that region 7 is region 1's (only) neighbor. The second row shows that region 2 has two neighbors: regions 6 and 8.

8.1.3. Queen contiguity

In queen contiguity, any region that touches the boundary of region *i*, whether on a side or a corner, is considered to be a neighbor. The maximum number of neighbors for this case is eight. In Figure 8.6, stars indicate

0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 8.5 Subset of unstandardized weight matrix for bishop contiguity.

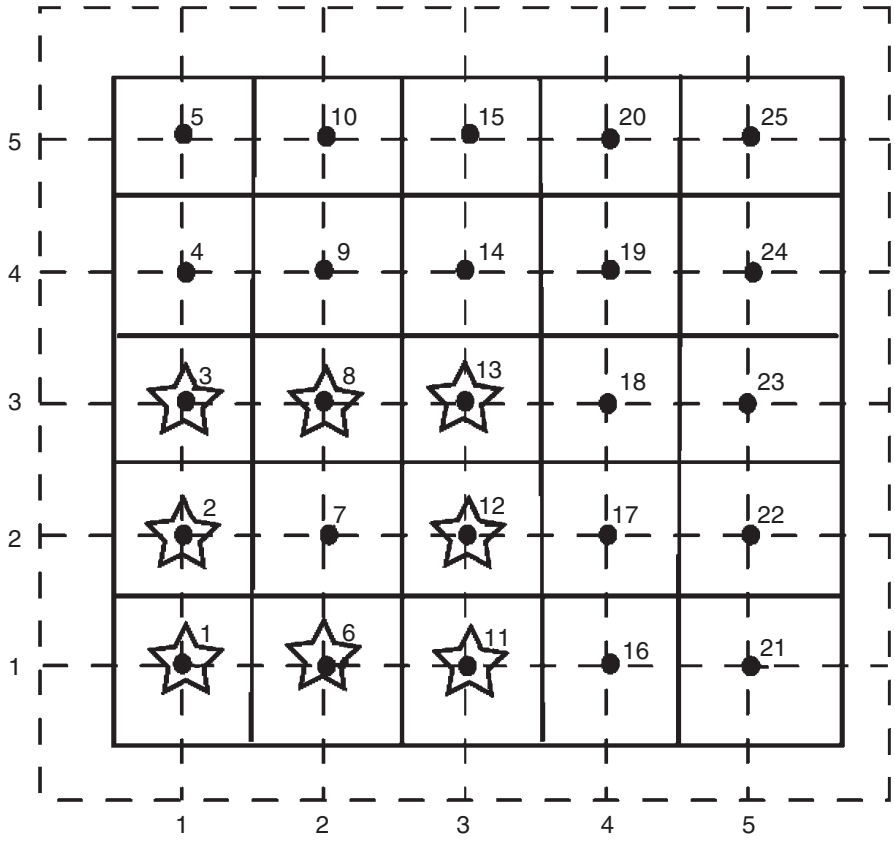


Figure 8.6 Neighbors in queen contiguity.

the neighbors for region 7 under queen contiguity.

The weight matrix for queen contiguity is the sum of the weight matrices for rook and bishop contiguity. The first 10 rows and columns of the unstandardized weight matrix are shown in Figure 8.7.

Comparing Figure 8.7 to Figures 8.5 and 8.6 shows that Figure 8.7 can be obtained by summing the other two weight matrices.

0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
1.00	0.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
0.00	1.00	0.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00
0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	1.00	1.00
0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	1.00
1.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
1.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
0.00	1.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00
0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00
0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00

Figure 8.7 Subset of unstandardized weight matrix for queen contiguity.

For example, the first row now has three ones, in positions 2, 6 and 7, showing that these three regions are neighbors of region 1.

8.2. CORRELATION MATRICES

Suppose that the data has been generated by the following process:

$$Y = \mu + \varepsilon \tag{8.1}$$

$$\varepsilon = \lambda W \varepsilon + e \tag{8.2}$$

where W is the weight matrix, and e is a white noise error term (that is, the elements of e are assumed to be independent and have zero mean and constant variance, σ^2). In this system, Y is a random variable with constant mean, μ, and a spatially correlated error term, ε. The error term in the first equation is spatially correlated because the error term for one observation depends on the value of the errors of its neighbors, as shown in equation 8.2. The parameter λ shows the strength of the spatial autocorrelation and must lie between -1 and 1 for a normalized weight matrix.

The variance-covariance matrix for the data is given by the following formula:

$$\Omega = \sigma^2 \left[(I - \lambda W)' (I - \lambda W) \right]^{-1} \tag{8.3}$$

where I is the identity matrix. The variance–covariance matrix can be converted into a correlation matrix, K, as shown in equation 8.4.

$$K_{ij} = \frac{\Omega_{ij}}{\sqrt{\Omega_{ii}\Omega_{jj}}} \tag{8.4}$$

As equations (8.3) and (8.4) show, the correlation matrix depends on the choice of W. In what follows, I examine the correlation matrices that result from using the three types of contiguity: bishop, rook and queen. In the examples λ is set to 0.67, σ^2 to 1, and all weight matrices are row normalized.

The correlation matrices for the example set of regions will be 25×25. Because it is difficult to look at so many numbers at one time, I will present the correlation matrices using symbols, rather than numbers. A diamond will represent correlations that are equal to or greater than 0.8. A cross will indicate that the correlation is less

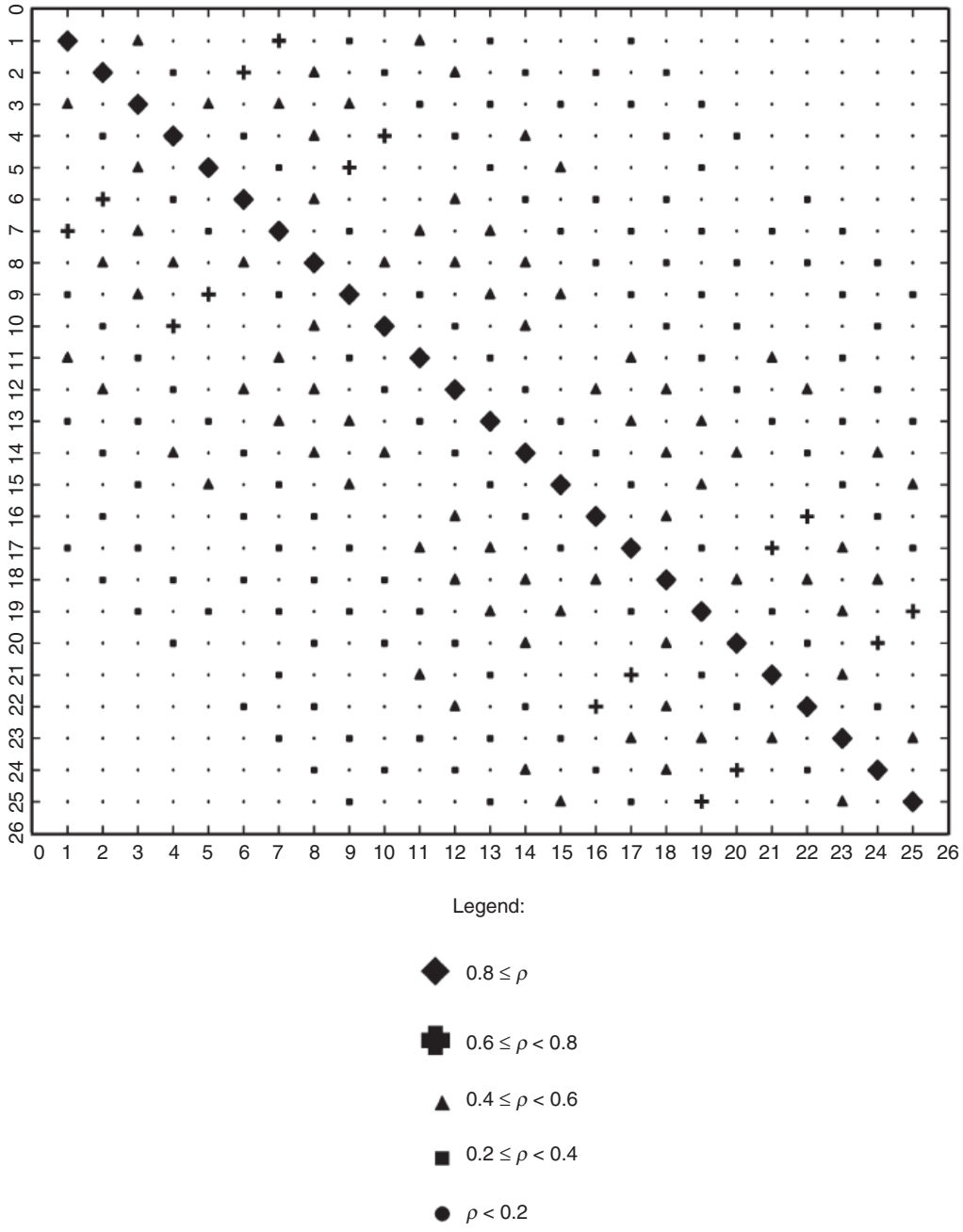


Figure 8.8 Correlation matrix for bishop contiguity.

than 0.8 but greater than or equal to 0.6. A triangle shows that the correlation is between 0.4 and 0.6. A square shows that the correlation is between 0.2 and 0.4, while a dot shows that the correlation is less than 0.2. Figures 8.8, 8.9 and 8.10 show the correlation matrices for bishop, rook and queen contiguity, respectively.

8.2.1. Correlation matrix for bishop contiguity

Even a brief examination of the three correlation matrices shows that the three weighting schemes produce very different correlations in the data. The bishop's case produces a particularly interesting correlation matrix.

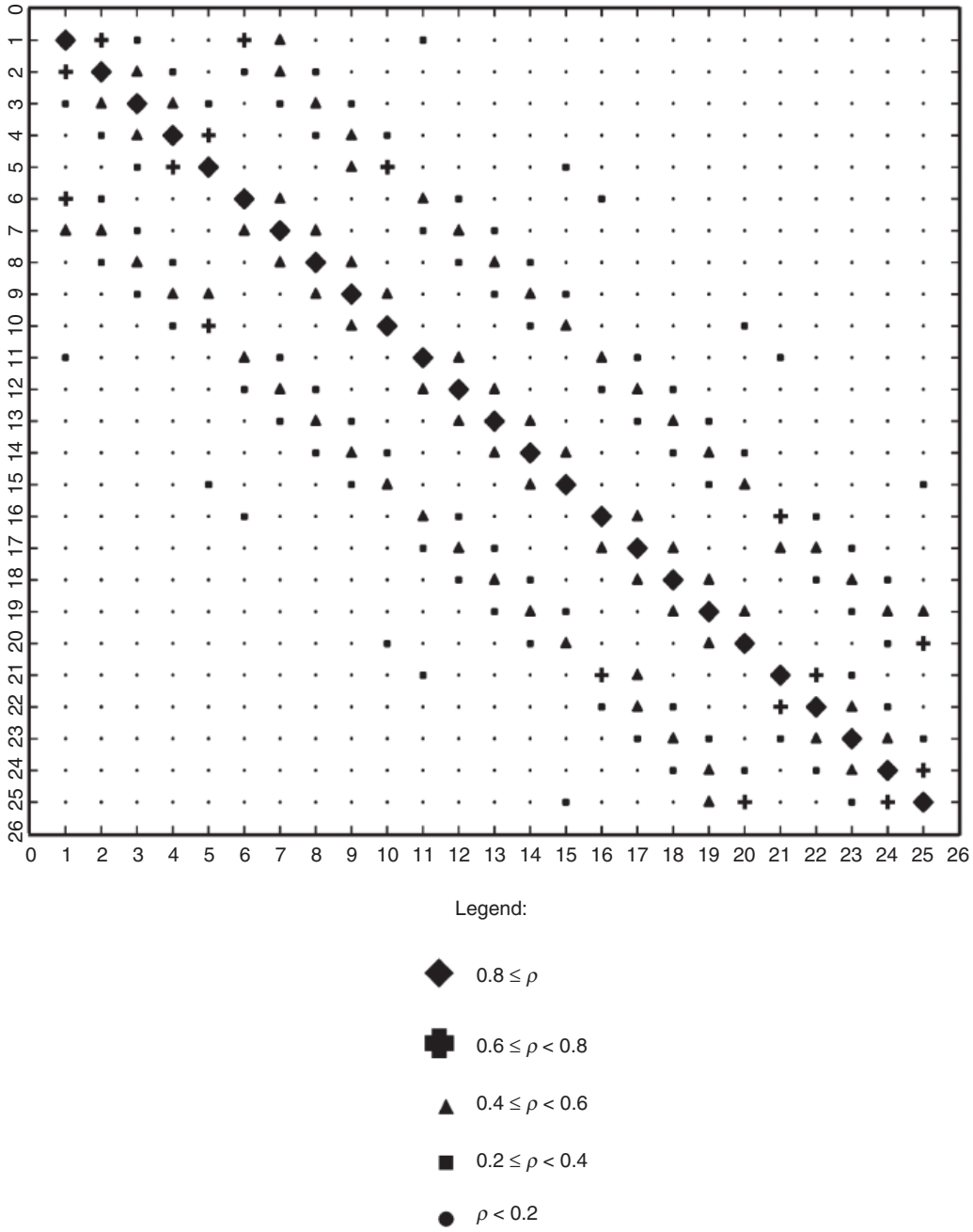


Figure 8.9 Correlation matrix for rook contiguity.

In what follows, I will use region 7 for illustration purposes. Figure 8.11 shows the seventh row of the correlation matrix.

Recall that under bishop contiguity the neighbors of region seven are regions 1, 3, 11 and 13. One might then reasonably expect that the correlation between all of the direct neighbors and region 7 would be the same. However, Figure 8.11 shows that the correlations are 0.76, 0.59, 0.59 and 0.48, respectively. The correlations differ because these regions have different numbers of neighbors themselves, as shown in Table 8.1. The general rule is that the

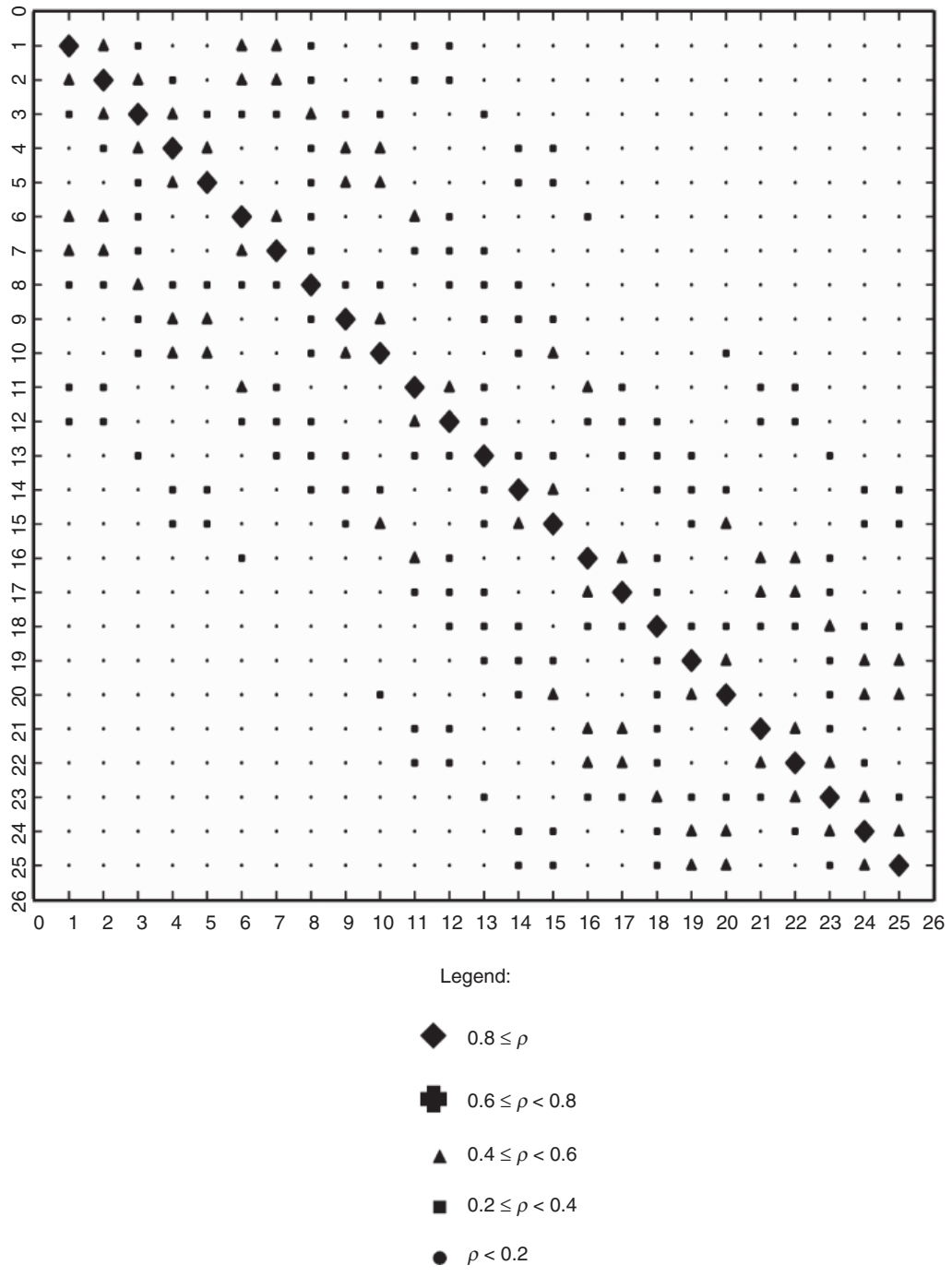


Figure 8.10 Correlation matrix for queen contiguity.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0.76	0.00	0.59	0.00	0.22	0.00	1.00	0.00	0.31	0.00	0.59	0.00	0.48	0.00	0.20	0.00	0.31	0.00	0.20	0.00	0.22	0.00	0.20	0.00	0.14

Figure 8.11 Row 7 from the bishop's correlation matrix.

Table 8.1 Neighbors of region 7 using bishop contiguity

Neighbors	*Correlation with region 7*	*Number of neighbors of neighbors (excluding region 7)*
1	0.76	0
3	0.59	1
11	0.29	1
13	0.48	3

more 'connected' a region is, the lower its correlation with another region. This makes sense because the more neighbors a region has, the greater the different influences on it.

Although sensible, this 'connectedness' property will increase the severity of 'edge effects'. Edge effects occur when the spatial processes continue outside of the study area. Regions with fewer neighbors are assigned higher correlations; however, these regions occur on the boundaries of the study area, where they will be influenced by regions not included in the study.

Also of note in the correlation matrix is the large number of zeros. These are shown as dots in Figure 8.8 and can be seen explicitly in Figure 8.11. This occurs because there are regions which are impossible to reach from region *i* using bishop's moves. For example, as Figure 8.12 shows, it is impossible to reach regions 8 or 10 from region 7, and so Figure 8.11 shows zeros for these cells.

Figure 8.12 also shows that, although not direct neighbors, regions 9 and 5 can be reached from region 7. It takes two

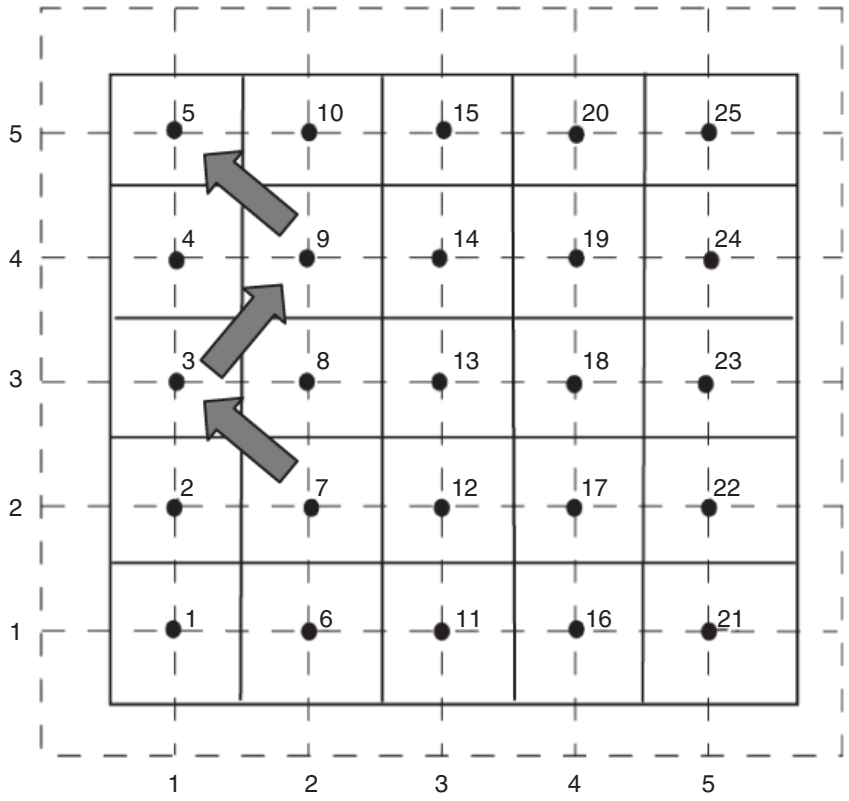


Figure 8.12 Relationship paths for bishop contiguity.

'moves' to reach region 9 and three to reach region 5. Therefore, we should see non-zero correlations in these cells, with a larger correlation in cell 9 than in cell 5, and this is confirmed by Figure 8.11.

8.2.2. Correlation matrix for rook contiguity

An examination of Figure 8.9 reveals a much different pattern of correlations for the rook's case. The 7th row of the correlation matrix is shown in Figure 8.13.

The same principal of greater connectivity leading to smaller correlations continues to be demonstrated by the rook contiguity, as shown in Table 8.2.

In the rook's case, it is possible to get from one region to any other region, although many 'moves' may be required. This means that there are no unrelated regions, as in the bishop's case, and hence no zeros in the correlation matrix. There are, however, many small values, and these are shown as dots in Figure 8.9. The greater the number of 'moves' required, the smaller the correlation. For example, starting from region 7, three moves are required to get to region 10 but only two to get to region 9. Figure 8.13 shows that the correlation between regions 7 and 9 is larger than that between regions 7 and 10. Likewise, it is also possible to get to region 25, but this requires 6 moves, and the correlation here is very small (0.02).

Finally, it is of interest to examine regions 9 and 13. Both are two moves away from region 7. However, two of region 13's neighbors are also neighbors of region 7. Region 9 has only one neighbor in common with region 7. This means that region 13 should have the stronger relationship with region 7, and this is borne out by Figure 8.13.

Table 8.2 Neighbors of region 7 using rook contiguity

Neighbors	*Correlation with region 7*	*Number of neighbors of neighbors (excluding region 7)*
2	0.51	2
6	0.51	2
8	0.44	3
12	0.44	3

8.2.3. Correlation matrix for queen contiguity

Although the weight matrix for the queen's case is the sum of the rook's and the bishop's case, the same cannot be said for the correlation matrix. However, the same principals noted above apply: more connected neighbors have lower correlations and shared neighbors increase the correlations. The seventh row of the correlation matrix is shown in Figure 8.14.

Clearly, in the queen's case, it is possible to reach one region from any other region, and so there are no isolated regions. However, there are again some very small correlations, which appear as dots in Figure 8.10. Table 8.3 shows the relationship between

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0.40	0.51	0.31	0.14	0.07	0.51	1.00	0.44	0.16	0.07	0.31	0.44	0.24	0.10	0.05	0.14	0.16	0.10	0.05	0.03	0.07	0.07	0.05	0.03	0.02

Figure 8.13 Row 7 from the rook's correlation matrix.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0.52	0.49	0.40	0.18	0.11	0.49	1.00	0.39	0.19	0.11	0.40	0.39	0.32	0.15	0.09	0.18	0.19	0.15	0.10	0.07	0.11	0.11	0.09	0.07	0.05

Figure 8.14 Row 7 from the queen's correlation matrix.

Table 8.3 Neighbors of region 7 using queen contiguity

Neighbors	*Correlation with region 7*	*Number of neighbor's neighbors*	*Number of shared neighbors*
1	0.52	2	2
2	0.49	4	4
3	0.40	4	2
6	0.49	4	4
8	0.39	7	4
11	0.40	4	2
12	0.39	7	4
13	0.32	7	2

connectedness and the correlations. The queen's case is somewhat more complex than the rook and the bishop because now there are more neighbors.

Examination of Table 8.3 shows that, as before, the correlations with region 7 fall as its neighboring regions have more neighbors themselves. For example, region 1 has the smallest number of neighbors (two) and the highest correlation (0.52). However, we see that the relationship is more complex than before, because regions 2, 3, 6, and 11 all have four neighbors but their correlations differ. The answer can be found in the last column of Figure 8.17, which shows the number of shared neighbors. That is, these are the number of regions that are neighbors to both region 7 and the region in the first column. Thus, regions 3 and 11 have the same correlation because both the number of neighbors and the number of shared neighbors is the same. Similarly, region 6 has a higher correlation because all of its four neighbors are also neighbors to region 7.

8.3. CORRELOGRAMS

The final tool that I will use to analyze differences between these weighting schemes is the correlogram. A correlogram shows how the correlations change as the distance between the regions increases. Thus, the correlation is graphed on the vertical axis, and separation distance is graphed on the horizontal axis. In general, we expect that the correlations will fall as separation distance increases. Figure 8.15 shows the correlograms for the three cases under consideration.

Since the data is on a regular lattice, and hence the centroids of the regions are evenly spaced, one might think that the correlations would be the same for any given separation distance. However, Figure 8.15 shows that this is not the case: there is a range of correlations for each separation distance, although this range gets smaller as the separation distance increases. The range of correlations comes from the dependence of the correlations on the connectedness of the neighbors and the number of shared neighbors as discussed above.

The correlations for both the rook's and the queen's cases do tend to fall with separation distance. For the bishop's case, there is a tendency for the correlations to decline with separation distance, but this decline is not monotonic. Relatively large correlations are interrupted by the zero correlations of the isolated regions, as discussed previously.

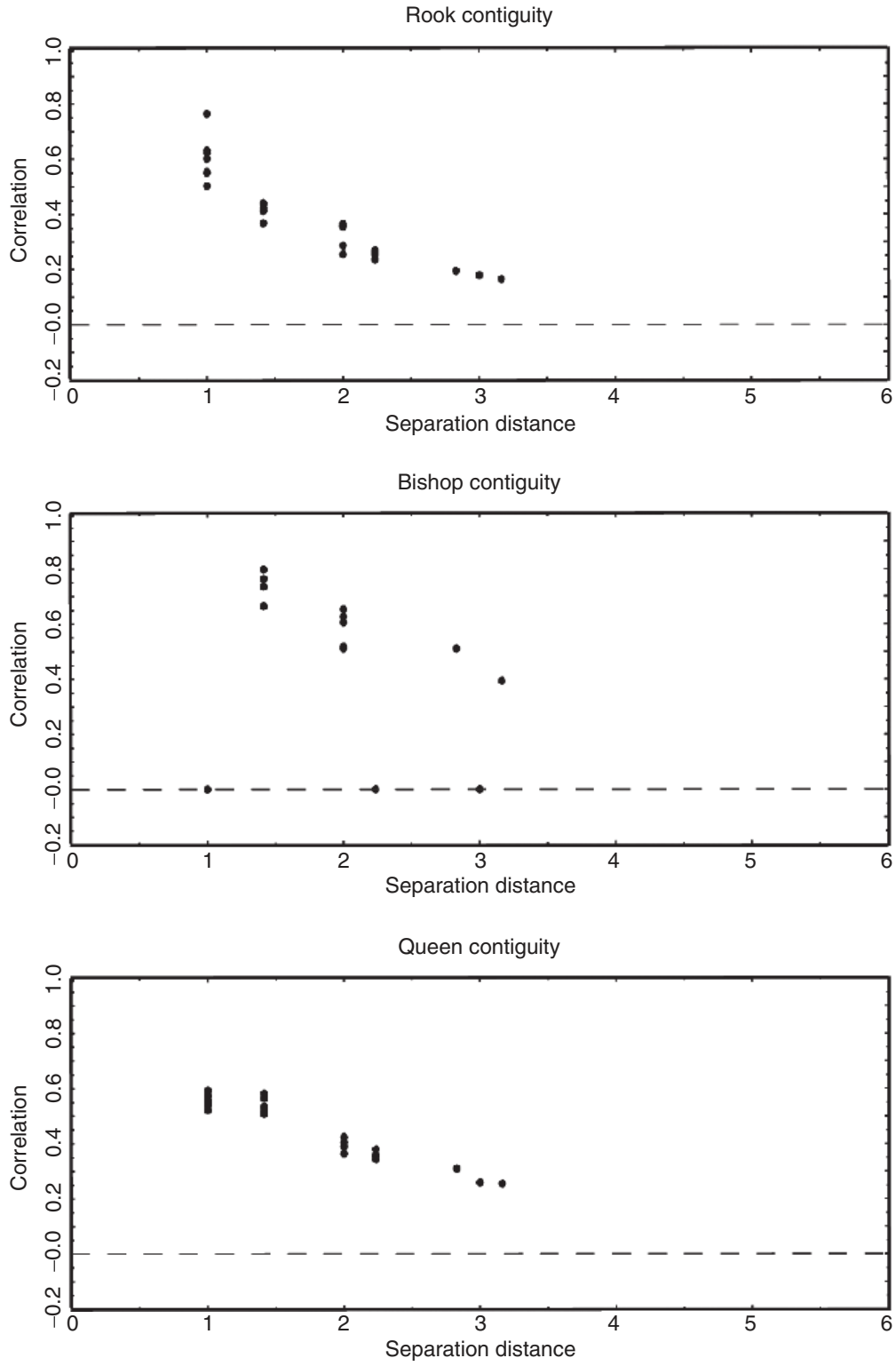


Figure 8.15 Correlograms for regular lattice data.

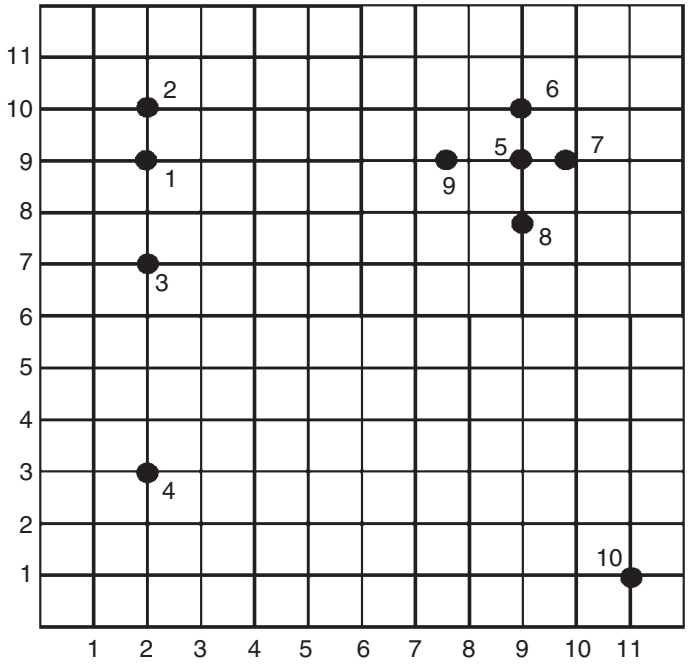


Figure 8.16 Irregularly located point example data.

8.4. REGULAR LATTICE POINT DATA

This is point data that is located at the intersection points of a regular grid. The data of the previous section can be used here by simply considering the centroids of the regions to be the data points. This means that applicable weighting schemes include: rook, bishop and queen contiguity. Also, weighting schemes that are used primarily for irregularly located point data can be used here as well. These will be discussed in the next section.

8.5. IRREGULARLY LOCATED POINT DATA

The discussion thus far has pertained to data located at regular intervals along a grid. However, spatial data is not always located so conveniently, and it is to this case that we now turn. In what follows, I will use ten points, located as shown in Figure 8.16, for purposes of illustration. The coordinates of these points are given in Table 8.4.

I have chosen to use a small number of points to keep the weight and correlation matrices small. Cluster 1 consists of observations 1 though 4. This cluster

Table 8.4 Coordinates for irregularly located point example data

Observation	*X Coordinate*	*Y Coordinate*
1	2	9
2	2	10
3	2	7
4	2	3
5	9	9
6	9	10
7	9.75	9
8	9	7.75
9	7.5	9
10	11	1

is somewhat dispersed, with observation 4 having the weakest link. Cluster 2 consists of observations 5 though 9. Cluster 2 is much tighter than Cluster 1. Observation 10 is an isolated point and not part of any cluster. The example data makes it clear that this type of data is very different from the regular lattice data, in which no clusters could appear. There are a number of weighting schemes that can be used for this type of data; the analyst must be skillful in choosing the weighting scheme that best represents the spatial interactions in the data.

Since the coordinates of the data points are known, the distances separating each pair of observations can be calculated. These distances can be stored in an $N \times N$ distance matrix. The distance matrix for the example data is shown in Table 8.5. All of the weighting schemes discussed in this chapter will be functions of separation distance.

The relatively small numbers in the shaded upper left portion of Table 8.5 reveal Cluster 1. The very small numbers in the shaded lower right portion reveal Cluster 2. The large numbers in the last row and column show that observation 10 is isolated from the rest of the data.

In what follows, I explore the properties of five different weighting schemes, which can be characterized as discrete or continuous. A discrete weighting scheme will have a non-normalized weight matrix consisting of ones and zeros, with the ones indicating the interactions. In the continuous weighting schemes, the cells will consist of numbers which indicate the strength of the interactions. Each of these weighting schemes has a parameter, the value of which must either be determined by the researcher or estimated. The weighting schemes are summarized in Table 8.6 and described below. Note that in most of the presented matrices, the

Table 8.5 Distance matrix for irregularly located point example data

0.00	1.00	2.00	6.00	7.00	7.07	7.75	7.11	5.50	12.04
1.00	0.00	3.00	7.00	7.07	7.00	7.81	7.35	5.59	12.73
2.00	3.00	0.00	4.00	7.28	7.62	8.00	7.04	5.85	10.82
6.00	7.00	4.00	0.00	9.22	9.90	9.80	8.46	8.14	9.22
7.00	7.07	7.28	9.22	0.00	1.00	0.75	1.25	1.50	8.25
7.07	7.00	7.62	9.90	1.00	0.00	1.25	2.25	1.80	9.22
7.75	7.81	8.00	9.80	0.75	1.25	0.00	1.46	2.25	8.10
7.11	7.35	7.04	8.46	1.25	2.25	1.46	0.00	1.95	7.04
5.50	5.59	5.85	8.14	1.50	1.80	2.25	1.95	0.00	8.73
12.04	12.73	10.82	9.22	8.25	9.22	8.10	7.04	8.73	0.00

Table 8.6 Weighting schemes

Scheme	*Type*	*Parameter*
Nearest neighbors	Discrete	Number of neighbors (*NN*)
Limit	Discrete	Distance limit (*L*)
Pace and Gilley's nearest neighbors	Continuous	Exponent (α), Maximum number of neighbors (k^*)
Inverse distance	Continuous	Exponent (*P*)
Negative exponential	Continuous	Denominator (*A*)

elements representing the pairs in the two clusters will be shaded. If the text refers to specific cells, these will be highlighted instead.

8.5.1. Nearest neighbors

A nearest neighbor weight matrix is defined so that:

$$W_{ij} = 1 \text{ if } j \text{ is } i\text{'s nearest neighbor}$$

$$= 0 \text{ otherwise}$$

A nearest neighbor is the observation that is the closest to observation i. Nearest neighbors can be generalized to include any number of neighbors. For example, if the number of nearest neighbors is set to five, then the non-normalized W will have five ones in each row, indicating the five closest observations to i. The number of neighbors (NN) is the parameter of this weighting scheme. Table 8.7 shows the weight matrix when NN is set to 1.

An examination of this table shows that the weight matrix is not symmetric. For example, (3, 1) = 1 but (1, 3) = 0. This is because observation 1 is observation 3's nearest neighbor, but the reverse is not true (observation 2 is 1's nearest neighbor). Also note that this weighting scheme gives observation 4 the same relationship with observation 3 that 3 has with 1, even though 3 is much closer to 1 than 4 is to 3. For one nearest neighbor, the unstandardized weight matrix will have one 1 in each row; in general, the number of ones per row will be equal to NN.

Figure 8.17 shows the correlation matrix for 1 nearest neighbor. The two clusters show up clearly. Note that observation 10 appears to be part of Cluster 2, even though it is distant from the other points in that cluster. Figure 8.18 provides a closer look at Cluster 2.

Let a doublet be a pair such that each is the other's nearest neighbor, and a singlet be a pair in which one member is the other's nearest neighbor, but not the reverse. The only doublet shown in Figure 8.18 is the pair (5, 7), and this pair has the highest correlation shown, at 0.92. Observations 6, 8 and 9 have only singlet connections to observation 5, and their correlations are lower at 0.83. Observation 10 has a singlet connection to 8, but this correlation is even lower at 0.76. This is because 8 (the nearest neighbor) is less well connected to the rest of the cluster than is point 5.

It may seem that there is a contradiction between the correlation patterns in the

Table 8.7 Weight matrix for example data with one nearest neighbor

0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00

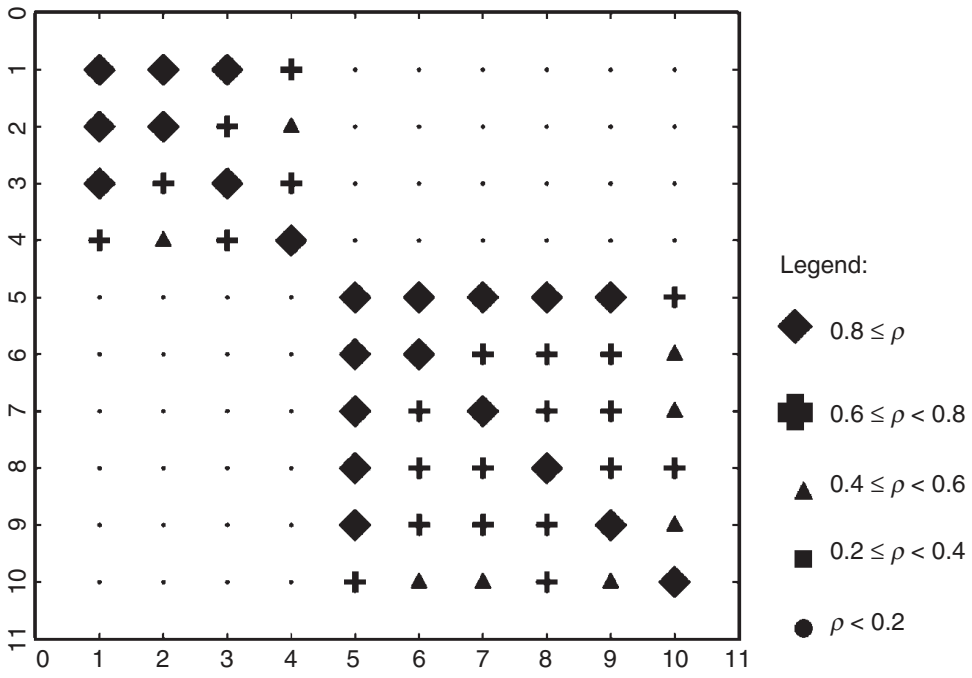


Figure 8.17 Correlation matrix for one nearest neighbor ($\lambda = 0.67$).

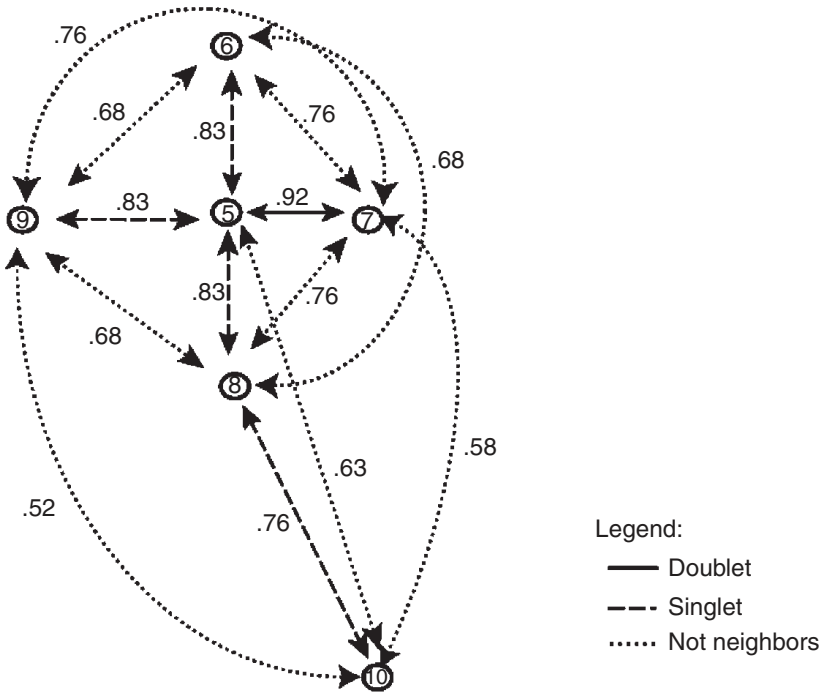


Figure 8.18 Correlations between selected points for cluster two, one nearest neighbor.

regular and irregular lattice cases: greater connectedness causes lower correlations in the regular lattice case and higher correlations here. The apparent contradiction is caused by differences in what is being held constant in each case. For the regular lattice, the number of neighbors can vary but the relationships between the data points is fixed. For the nearest neighbors weighting scheme with irregularly spaced data, the number of neighbors is fixed, but the spatial relationships can change. For the regular lattice, more neighbors means more influences and therefore lower correlations. Here, having a central neighbor (as indicated by a high correlation) causes point *i* to be more central as well.

With nearest neighbors, points can only be related through nearest neighbor pairs (whether doublet or singlet). Thus a 'move' here is a step along a path connecting nearest neighbor pairs. As in the regular lattice case, as the number of moves increases, the correlations fall. Observation 9 has a correlation of 0.52 with point 10. This is because 6 is related to 10 through observation 5 (path: 9, 5, 8, 10). Observation 7 has a slightly higher correlation with 10 (0.58) even though the same number of moves are involved, because the path to 10 goes through the doublet (5, 7) (path: 7, 5, 8, 10).

8.5.2. Two nearest neighbors

Table 8.8 shows the standardized weight matrix and Figure 8.19 the symbolic correlation matrix for two nearest neighbors. Increasing the number of neighbors to two makes the weight matrix more symmetric in the clusters, which means that there will be more doublets. Since there are more doublets, the impact of the most central points (1 and 5) is reduced. This has the effect of making the clusters appear more compact, as shown in Figure 8.19.

Figure 8.20 explores Cluster 2 more closely. Since there are two neighbors, as opposed to one, Figure 8.20 shows more doublets and singlets than Figure 8.18. The general rules for the magnitudes of the correlations are that: (a) doublets have higher correlations than singlets and (b) the more connected the 'partners', the higher the correlations (this rule holds for both doublets and singlets). For example, $\rho_{8,10}$ is smaller than $\rho_{8,5}$ because point 8 is better connected than point 10.

The pairs shown by the dotted lines are not nearest neighbor pairs, and therefore are reached by steps along a path. As before, if the path goes through a doublet, the correlations are higher than if it goes through a singlet. Also, the larger the number of steps, the lower the correlation.

Table 8.8 Standardized weight matrix for two nearest neighbors

0.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.50	0.50	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.50	0.50	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.00	0.00

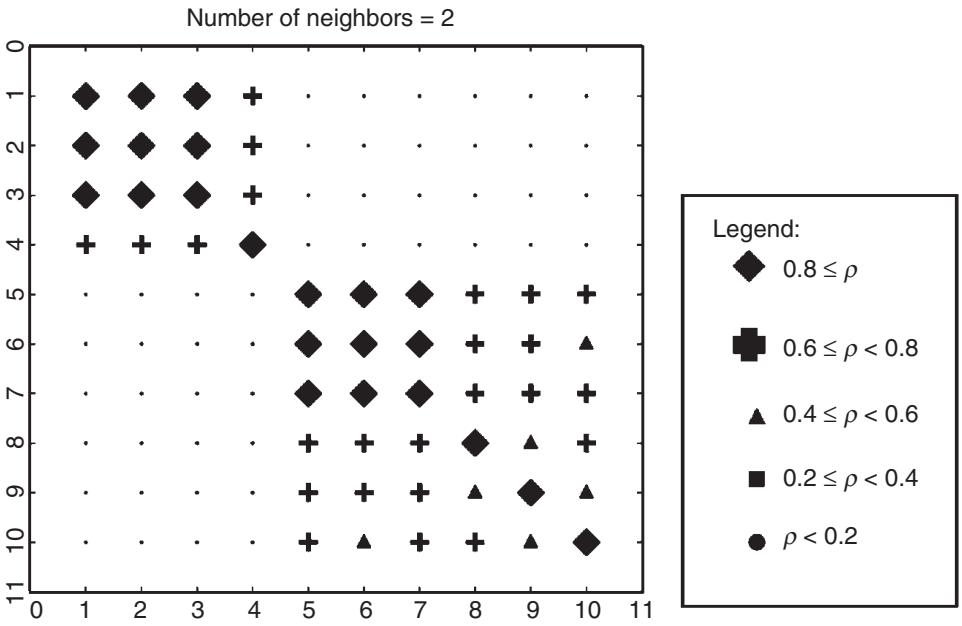


Figure 8.19 Correlation matrix for two nearest neighbors ($\lambda = 0.67$).

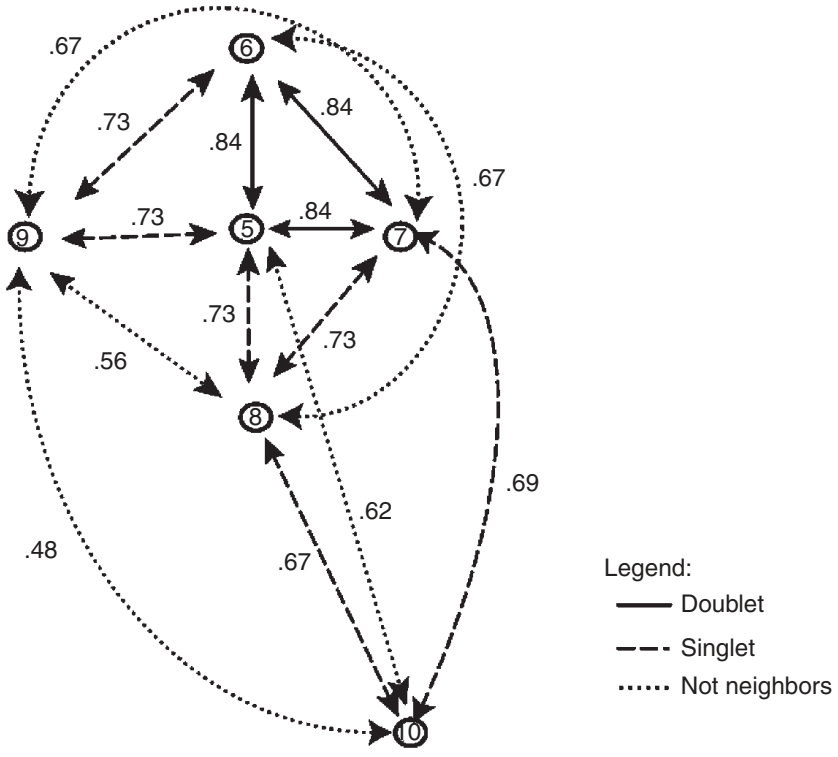


Figure 8.20 Correlations between selected points for cluster two, two nearest neighbors.

8.5.3. Three nearest neighbors

Table 8.9 shows the standardized weight matrix for this case and Figure 8.21 shows the correlation matrix. Because there are more doublets and singlets, more pairs become well connected or central. However, the influence of any individual connection is diminished, since there are more connections. Thus the clusters look more diffused and the two clusters begin to affect each other, as shown in Figure 8.21.

Figure 8.22, which illustrates the correlations for Cluster 2, shows that pair (5, 7) has regained its dominant position, having the highest correlation at 0.74. This pair is the most connected because all of its connections are doublets. Even though this pair has the highest correlation, it had a much higher correlation in the one nearest neighbor case (0.92). This is because there are many more doublets here, so each has a smaller impact.

Comparing Figure 8.22 to Figures 8.18 and 8.20 shows that most of the dotted lines

Table 8.9 Standardized weight matrix for three nearest neighbors

0.00	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.33	0.00
0.33	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.33	0.00
0.33	0.33	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00
0.33	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.33	0.33	0.33	0.00	0.00
0.00	0.00	0.00	0.00	0.33	0.00	0.33	0.00	0.33	0.00
0.00	0.00	0.00	0.00	0.33	0.33	0.00	0.33	0.00	0.00
0.00	0.00	0.00	0.00	0.33	0.00	0.33	0.00	0.33	0.00
0.00	0.00	0.00	0.00	0.33	0.33	0.00	0.33	0.00	0.00
0.00	0.00	0.00	0.00	0.33	0.00	0.33	0.33	0.00	0.00

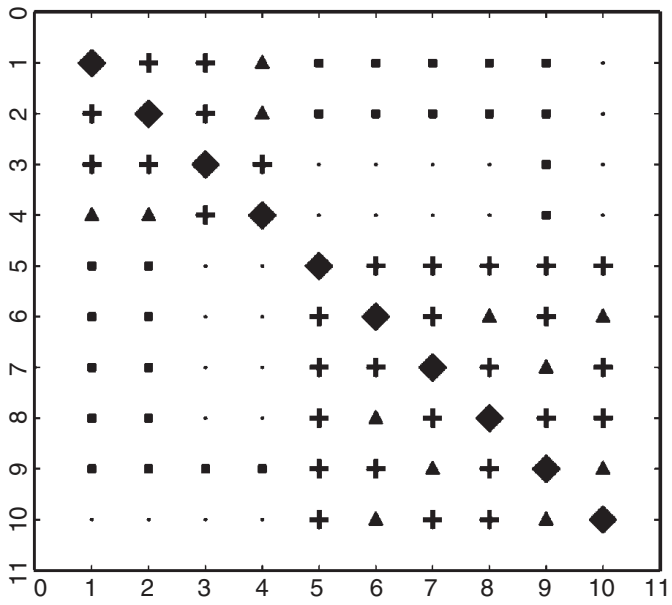


Figure 8.21 Correlation matrix for three nearest neighbors.

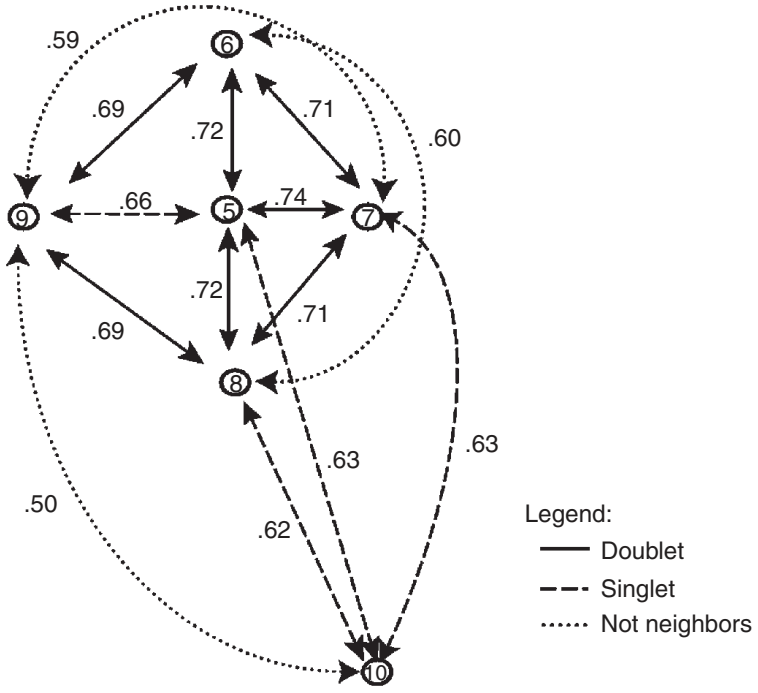


Figure 8.22 Correlations between selected points for cluster two, three nearest neighbors.

have been replaced by solid and dashed lines, because there are more neighbors. Thus, most of the points in this cluster are related, which leads to the diffused pattern of correlations shown in Figure 8.21.

8.5.4. Correlograms for nearest neighbors

The correlograms for nearest neighbors are shown in Figure 8.23. Examination of this figure shows that the correlations do not decline monotonically with separation distance. For one nearest neighbor, although there is a slight diminution with distance, the basic pattern is that the correlations are either very strong or zero. Furthermore, the strong correlations are interspersed with the zeros. This means that some pairs can be highly correlated, while others, that are closer together, are not. However, all of the small separation distance pairs are highly correlated and all of the large separation distance pairs are not related to each other.

Increasing the number of neighbors to two reduces the size of the large correlations, and hence diminishes the diminution with separation distance. The same pattern of large correlations interspersed with zeros persists. Finally, when the number of neighbors is increased to three, all points are related at least weakly (keep in mind that there are only 10 observations). The upper end of the strong (greater than 0.5) correlations has been reduced further. There is still a separation distance range in which strong correlations are interspersed with weak ones, and so there is no monotonic relationship between separation distance and correlation. It remains true, however, that the strongest correlations are associated with the smallest separation distances and the largest separation distances have the smallest correlations.

All of these results show that the number of neighbors is an important parameter for this weighting scheme; the number of neighbors

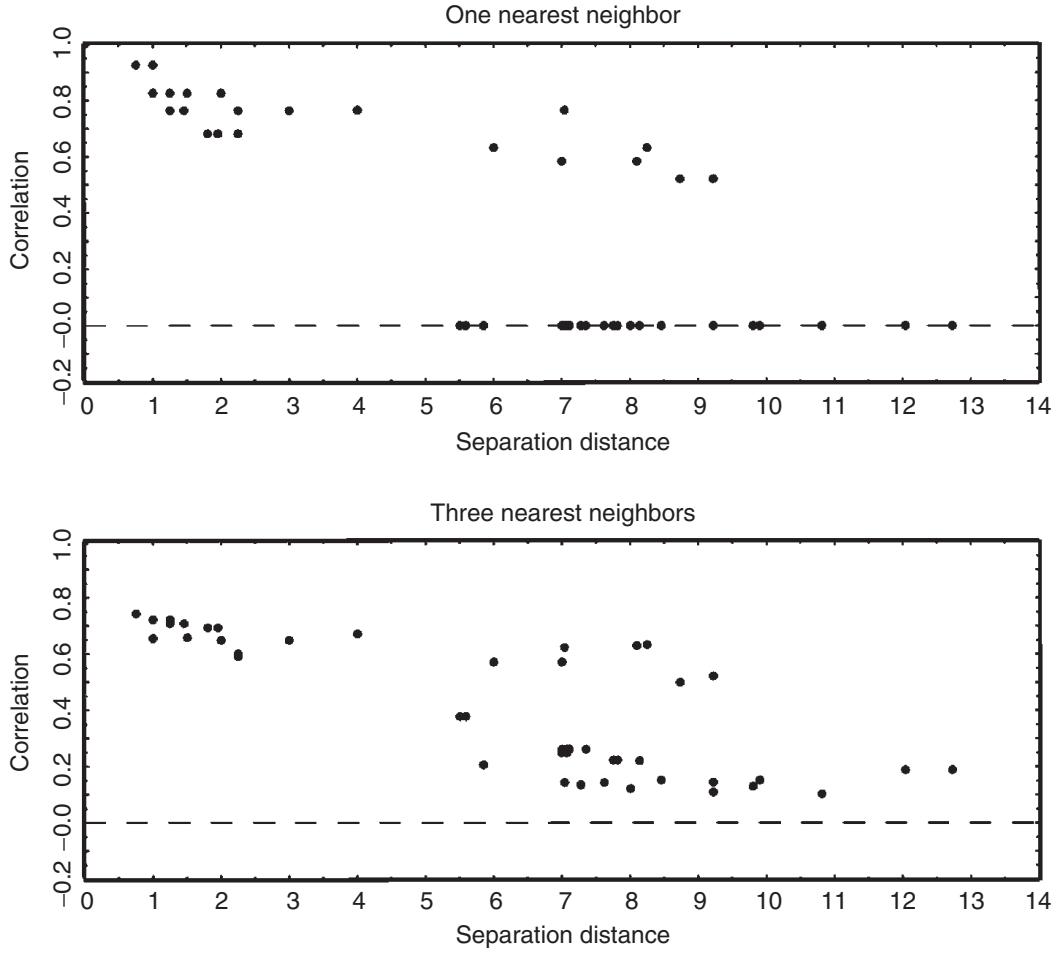


Figure 8.23 Nearest neighbor correlograms, $\lambda = 0.67$.

has a sizeable impact on the weight matrix and the associated correlation matrix. It is traditional for researchers to pick the number of neighbors *a priori*. These results show that this should be done with care.

8.5.5. Pace and Gilley's continuous version of nearest neighbors (P&G)

In this model, described in pace and Gilley (1998), the unstandardized weight matrix is given by:

$$W = \sum_{k=1}^{NN} \alpha^k N(k)$$

where $N(k)$ is an $N \times N$ matrix such that:

$$N(k)_{ij} = 1 \quad \text{if } j \text{ is } i\text{'s } k\text{th nearest neighbor}$$

$$= 0 \quad \text{otherwise,}$$

α is a parameter to be estimated, and NN is chosen by the researcher. This model was developed to finesse the fact that the number of neighbors is generally chosen by the researcher, rather than estimated. As the value of α increases, the influence of more distant neighbors increases. Thus, if the researcher does not know the number of neighbors, he can pick a number k^* (which is generally larger than the probable number of neighbors)

and then estimate α to find the optimal degree of influence.[1] In the example, *NN* is set to 5.

This weighting scheme falls into the continuous category, because the unstandardized weight matrix does not consist of ones and zeros, as in nearest neighbors. The standardized weight matrices are shown in Tables 8.10 and 8.11, for $\alpha = 0.1$ and $\alpha = 0.5$. These weight matrices differ from the nearest neighbor weight matrices, particularly as α becomes larger. For example, a comparison of Tables 8.7 and 8.10 shows that $\alpha = 0.1$ produces a weight matrix that is similar to that for one nearest neighbor. However, comparing Tables 8.9 and 8.11 shows that the $\alpha = 0.5$ case is quite different from three nearest neighbors. This is because, in the nearest neighbors weighting scheme, all of the neighbors are assigned equal weight. In the P&G weighting scheme, the first nearest neighbor always has the greatest weight. As the value of α increases, the more distant neighbors are given greater weight, but the weights are always less than for the first nearest neighbor. For example, in Table 8.9 (three nearest neighbors) W has 0.33 in the second, third and ninth elements of the first row, while in Table 8.11 ($\alpha = .3$) the corresponding elements are 0.52, 0.26, and 0.13.

Figure 8.24 shows the correlation matrices for $\alpha = 0.1$ and 0.5. Examination of this figure shows that $\alpha = 0.1$ corresponds well to one nearest neighbor. Not surprisingly however, $\alpha = 0.5$ differs from three nearest neighbors in that there is less bleeding of the clusters with P&G.

Table 8.10 Standardized weight matrix for $\alpha = 0.1$

0.00	0.90	0.09	0.00	0.00	0.00	0.00	0.00	0.01	0.00
0.90	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.01	0.00
0.90	0.09	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.09	0.01	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.09	0.90	0.01	0.00	0.00
0.00	0.00	0.00	0.00	0.90	0.00	0.09	0.00	0.01	0.00
0.00	0.00	0.00	0.00	0.90	0.09	0.00	0.01	0.00	0.00
0.00	0.00	0.00	0.00	0.90	0.00	0.09	0.00	0.01	0.00
0.00	0.00	0.00	0.00	0.90	0.09	0.00	0.01	0.00	0.00
0.00	0.00	0.00	0.00	0.01	0.00	0.09	0.90	0.00	0.00

Table 8.11 Standardized weight matrix for $\alpha = 0.5$

0.00	0.52	0.26	0.06	0.03	0.00	0.00	0.00	0.13	0.00
0.52	0.00	0.26	0.06	0.00	0.03	0.00	0.00	0.13	0.00
0.52	0.26	0.00	0.13	0.00	0.00	0.00	0.03	0.06	0.00
0.26	0.13	0.52	0.00	0.00	0.00	0.00	0.03	0.06	0.00
0.03	0.00	0.00	0.00	0.00	0.26	0.52	0.13	0.06	0.00
0.00	0.03	0.00	0.00	0.52	0.00	0.26	0.06	0.13	0.00
0.03	0.00	0.00	0.00	0.52	0.26	0.00	0.13	0.06	0.00
0.00	0.00	0.03	0.00	0.52	0.06	0.26	0.00	0.13	0.00
0.03	0.00	0.00	0.00	0.52	0.26	0.06	0.13	0.00	0.00
0.00	0.00	0.00	0.03	0.13	0.00	0.26	0.52	0.06	0.00

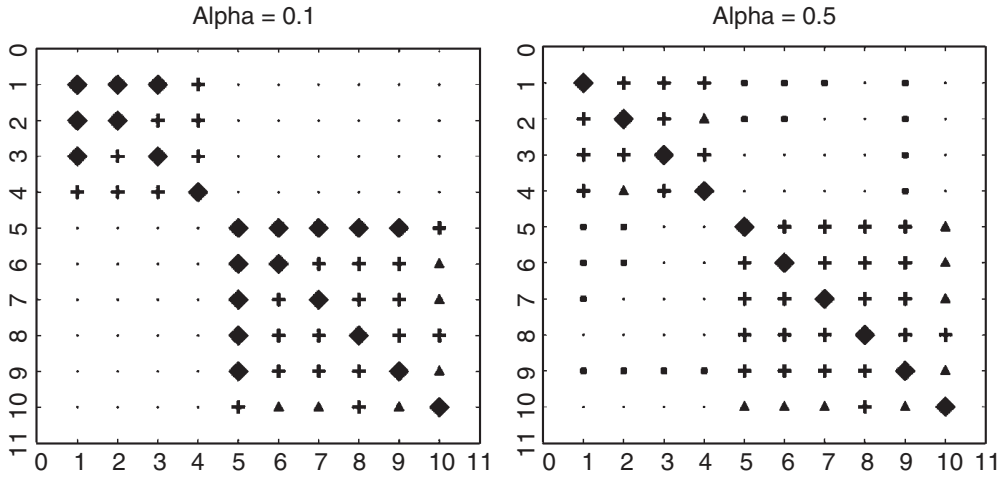


Figure 8.24 Correlation matrices for Pace and Gilley model.

Figure 8.25 shows the correlograms for the P&G model. These should be compared to the nearest neighbors correlograms in Figure 8.23. Not surprisingly, the first panels of these two figures agree quite closely, while the second panels do not. In the P&G model, the attenuation of the strong correlations with separation is more pronounced than in nearest neighbors. Additionally, the range of the strong correlations does not become as compressed when α increases as it does when the number of neighbors increases. Finally, comparing the two panels in Figure 8.25, the 'bleeding' of the clusters is shown by the zero correlations in the first panel ($\alpha = 0.1$) becoming positive in the second panel ($\alpha = 0.5$).

8.6. DISCUSSION

Pace's model is sufficiently different from nearest neighbors that I do not believe it should be considered as a replacement with an estimable parameter. Rather, I believe that it adds an additional feature to the nearest neighbors model. That is, it changes the weighting on the neighbors so that more distant (in terms of neighbors) neighbors have less weight. Further, the rate of decline of the weight is estimated. Thus it should be considered to be an important weighting scheme in its own right. The features of this weighting scheme may make sense in some situations. For example, if the data looks like cluster one, then it makes sense to weight the third neighbor less than the second. However, if the data looks like cluster two, it makes less sense. Note also that the researcher must choose the maximum number of neighbors, *NN*. As in all choices of parameters, the researcher must use his judgment as to what is best.

8.6.1. Limit models

Limit models use a weighting scheme such that:

$$W_{ij} = 1 \quad \text{if } D_{ij} \leq L$$

$$= 0 \quad \text{otherwise}$$

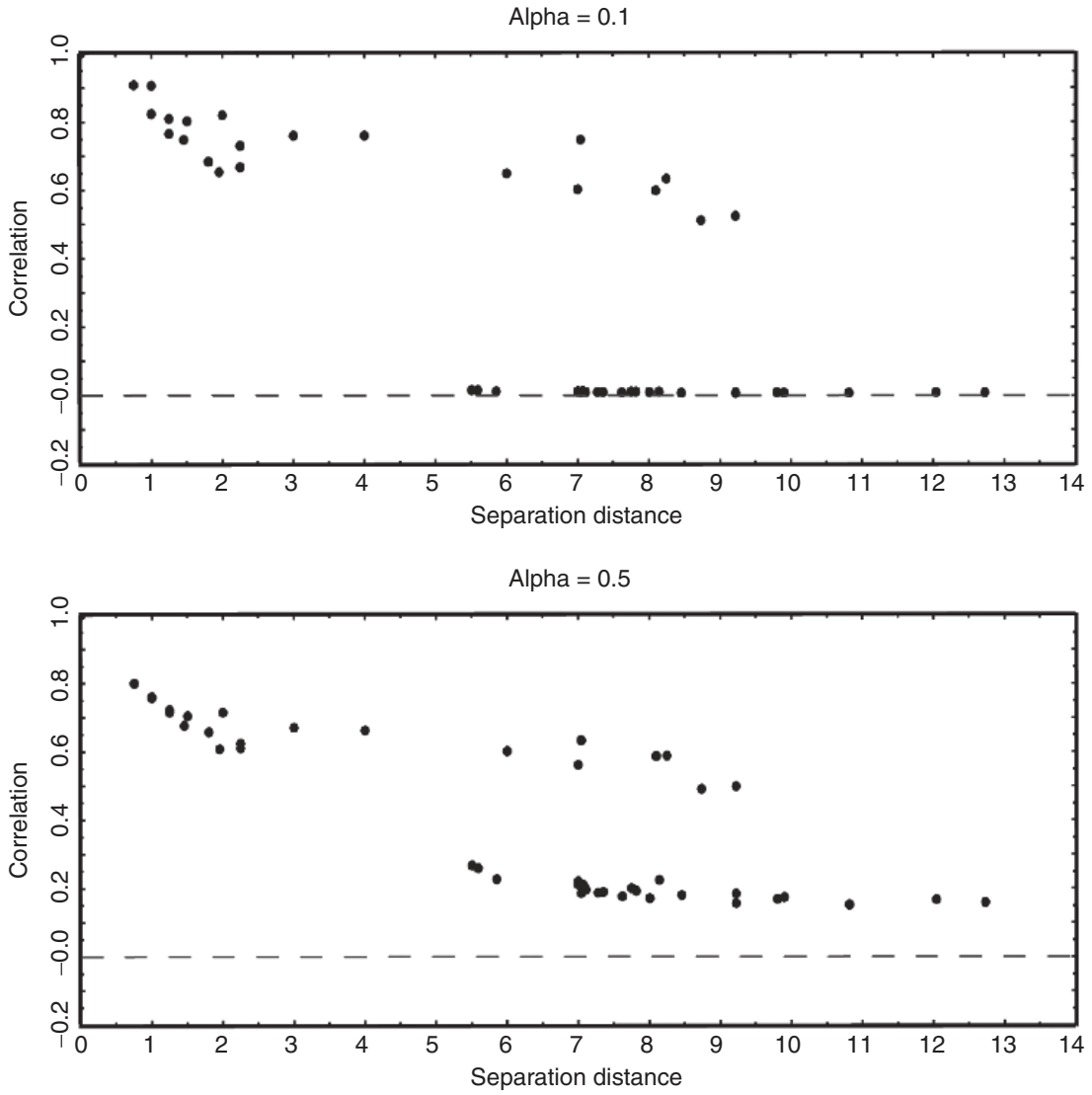


Figure 8.25 Correlograms for Pace and Gilley model.

where L is the distance limit. The parameter L is usually chosen by the researcher and its value can have a profound effect on both the weight and correlation matrices. The unstandardized version of W is symmetric because $D_{ij} = D_{ji}$. However, the standardized W is usually not symmetric because the number of points within the distance limit will vary by observation. This is a discrete weighting scheme because the unstandardized W consists entirely of ones and zeros.

Tables 8.12 and 8.13 show standardized weight matrices for distance limits of 1 and 3, respectively. When $L = 1$, W is very sparse, because there are very few pairs with separation distance less than one. The asymmetry is illustrated by the shaded cells in Table 8.12. Observation 5 has two other points located within 1 distance unit (points 6 and 7), and so these weights are standardized to 0.5. However, points 6 and 7 only have one 'neighbor' each (point 5), and so the weight for point 5 in

Table 8.12 Standardized weight matrix, distance limit = 1

0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 8.13 Standardized weight matrix, distance limit = 3

0.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.25	0.25	0.25	0.25	0.00
0.00	0.00	0.00	0.00	0.25	0.00	0.25	0.25	0.25	0.00
0.00	0.00	0.00	0.00	0.25	0.25	0.00	0.25	0.25	0.00
0.00	0.00	0.00	0.00	0.25	0.25	0.25	0.00	0.25	0.00
0.00	0.00	0.00	0.00	0.25	0.25	0.25	0.25	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

rows 6 and 7 is 1. Table 8.13 shows that, when the distance limit is increased to 3, the weight matrix becomes symmetric. This distance limit reveals the two clusters, putting observations 1 through 3 in Cluster 1 and 5 through 9 in Cluster 2. Points 4 and 10 have no 'neighbors'; these rows contain only zeros.

Figure 8.26 shows correlation matrices for the Limit Model, for $L = 1$ and $L = 3$. When $L = 1$, the correlation matrix is very sparse, and the correlations that are present are very high. Observations 1 and 5 are very dominant. The two clusters are apparent, but include too few observations. Expanding the distance limit to 3 reveals the two clusters more accurately, although observations 4 and 10 remain excluded from either. Although $L = 3$ seems to make the most sense for this data, the correlation matrix does not resemble those of the previously discussed weighting schemes. Also note that L is not required to be an integer.

Figure 8.27 shows the correlograms for $L = 1$ and $L = 3$. For the small distance limits shown here, there is no intermingling (with respect to distance) of correlated and uncorrelated points, as in nearest neighbors. Pairs are either correlated or not, and when they are, the correlation is high. The strictly positive correlations end at the distance limit. However, when the distance limit is larger (not shown), there is a range of separation distances in which positive and zero correlations are interspersed. This range occurs beyond the distance limit and shows the 'neighbors of neighbors' effect.

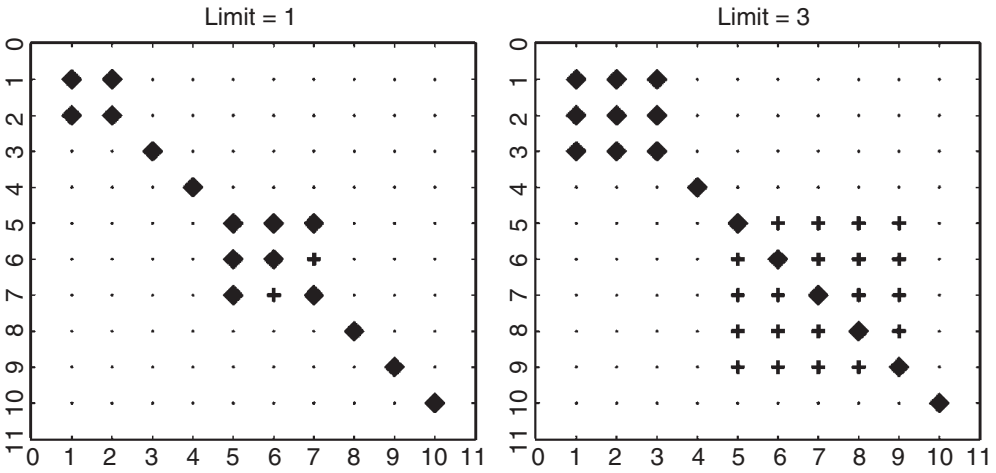


Figure 8.26 Correlation matrices for limit model.

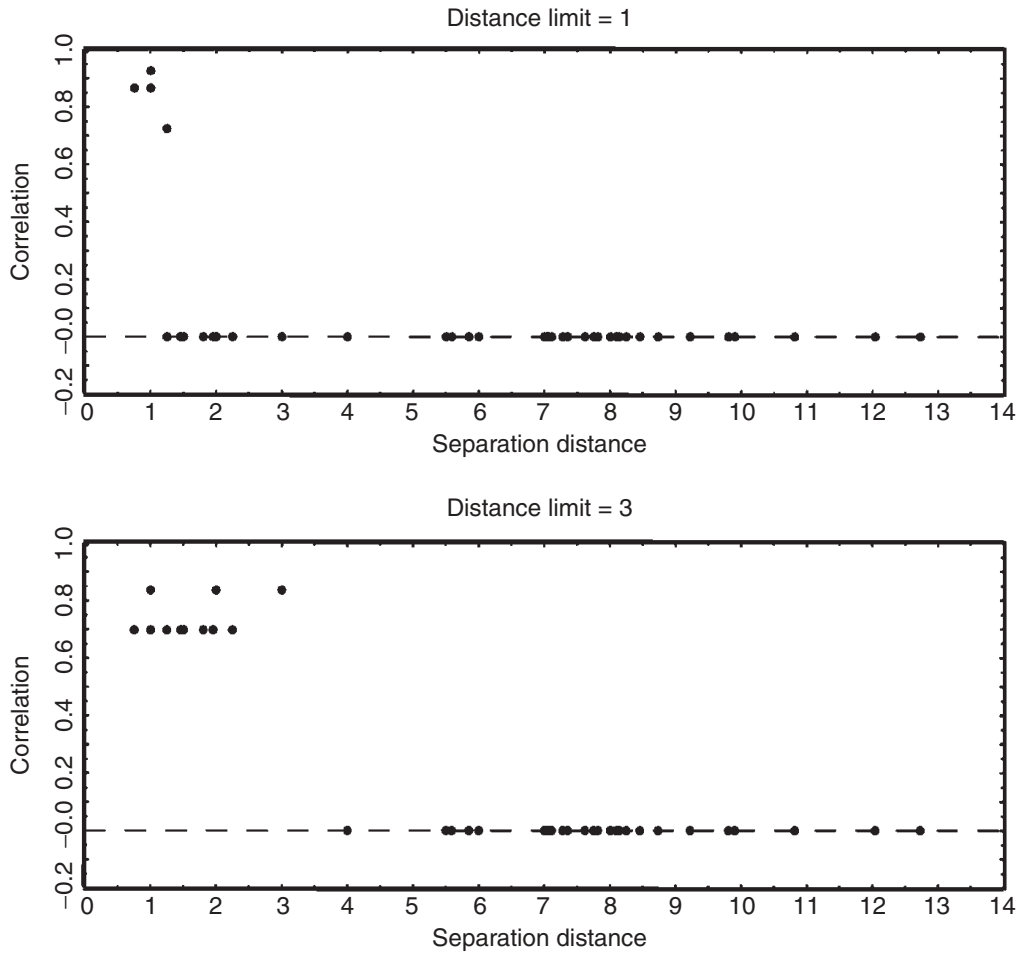


Figure 8.27 Correlograms for limit model.

8.6.2. Inverse distance

In this weighting scheme, the weights are inversely related to separation distance as shown below:

$$W_{ij} = \frac{1}{D_{ij}^{P}}$$

where the exponent P is a parameter that is usually set by the researcher. This weighting scheme falls into the continuous category because the unstandardized weights are between 1 and 0 (inclusive), rather than being restricted to 1 or 0.

Tables 8.14 and 8.15 show the standardized weight matrices for $P = 1$ and $P = 3$. When $P = 1$ (Table 8.14), the weights in the clusters are relatively low. For example, the weights associated with point 1 (the most central point in Cluster 1) are all less than 0.5, and the weights associated with point 5 are all less than 0.4. Additionally, the weights for pairs outside the clusters are relatively large, reaching values as high as 0.11.

When $P = 3$ the in-cluster weights are very large, while the out of cluster weights are close to zero. For example, the weights associated with point 1 are as high as 0.95 (since these weight matrices are standardized, all weights lie between 0 and 1). The weights associated with point 5 are as large as 0.72. Points 4 and 10 have relatively large weights.

Figure 8.28 shows the correlation matrices for $P = 1$ and $P = 3$. When $P = 0.5$

Table 8.14 Standardized weight matrix $P = 1$

0.00	0.40	0.20	0.07	0.06	0.06	0.05	0.06	0.07	0.03
0.44	0.00	0.15	0.06	0.06	0.06	0.06	0.06	0.08	0.03
0.27	0.18	0.00	0.13	0.07	0.07	0.07	0.08	0.09	0.05
0.14	0.12	0.20	0.00	0.09	0.08	0.08	0.10	0.10	0.09
0.03	0.03	0.03	0.02	0.00	0.22	0.30	0.18	0.15	0.03
0.04	0.04	0.04	0.03	0.29	0.00	0.23	0.13	0.16	0.03
0.03	0.03	0.03	0.03	0.34	0.21	0.00	0.18	0.11	0.03
0.05	0.04	0.05	0.04	0.26	0.14	0.22	0.00	0.16	0.05
0.06	0.06	0.06	0.04	0.23	0.19	0.15	0.17	0.00	0.04
0.09	0.08	0.10	0.11	0.12	0.11	0.13	0.15	0.12	0.00

Table 8.15 Standardized weight matrix $P = 3$

0.00	0.87	0.11	0.00	0.00	0.00	0.00	0.00	0.01	0.00
0.95	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.01	0.00
0.65	0.19	0.00	0.08	0.01	0.01	0.01	0.01	0.03	0.00
0.15	0.09	0.50	0.00	0.04	0.03	0.03	0.05	0.06	0.04
0.00	0.00	0.00	0.00	0.00	0.24	0.57	0.12	0.07	0.00
0.00	0.00	0.00	0.00	0.56	0.00	0.29	0.05	0.10	0.00
0.00	0.00	0.00	0.00	0.72	0.16	0.00	0.10	0.03	0.00
0.00	0.00	0.00	0.00	0.48	0.08	0.30	0.00	0.13	0.00
0.01	0.01	0.01	0.00	0.42	0.24	0.12	0.19	0.00	0.00
0.05	0.04	0.06	0.10	0.14	0.10	0.15	0.23	0.12	0.00

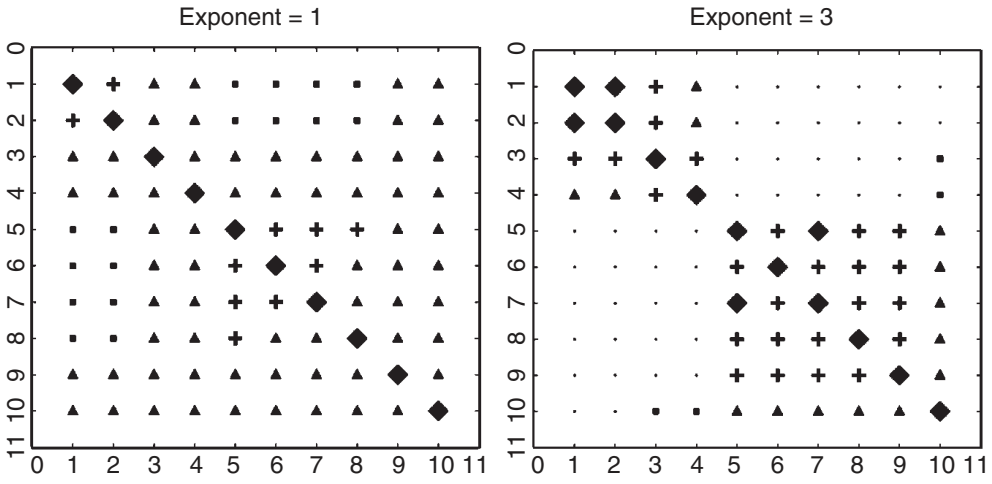


Figure 8.28 Correlation matrices for inverse distance model.

(not shown), all points are correlated with all other points, and the correlations are roughly the same. When $P = 1$, we begin to see some stronger correlations associated with points 1 and 5, but non-zero correlations still exist between all pairs. When $P = 3$, the clusters appear very clearly, and points 1 and 5 appear to be influential. Note that points 4 and 10 are always included in the clusters.

Figure 8.29 shows the correlograms for $P = 1$ and $P = 3$. Examination of this figure shows that the correlations decline monotonically when $P = 1$. At $P = 3$, an intermixing of large and small correlations occurs when separation distance is in the range of 5 to 9.

8.6.3. Negative exponential model

This is another continuous weighting scheme. Here the weights decline exponentially with separation distance.

$$W_{ij} = \exp(-D_{ij}/A)$$

where A is a parameter that is commonly chosen by the researcher.

Tables 8.16 and 8.17 show the standardized weight matrices for $A = 0.5$ and $A = 2$. When $A = 0.5$, the weights within the clusters are reasonably large, and the largest weights are associated with points 1 and 5. The weights for pairs outside of the clusters are all zero, except for point 10. When $A = 2$, the weights associated with points 1 and 5 get smaller, but there is an indeterminate effect on the weights on the other pairs in the cluster: some get smaller and some get larger. The weights on the pairs outside of the cluster become larger.

Figure 8.30 shows the correlation matrices for $A = 0.5$ and $A = 2$. When $A = 0.25$ (not shown), the clusters are clearly indicated, and correlations associated with points 1 and 5 are very large, indicating their centrality. When A is increased to 0.5, point 5 loses some of its centrality, remaining highly correlated only with point 7. When A is set to 1 (not shown), point 5 appears no different from the other points in Cluster 2, and the centrality of point 1 becomes weaker. Finally, when $A = 2$, all of the points become correlated, although the highest correlations remain in the clusters. Note that point 10 is always included in Cluster 2, regardless of the value of A.

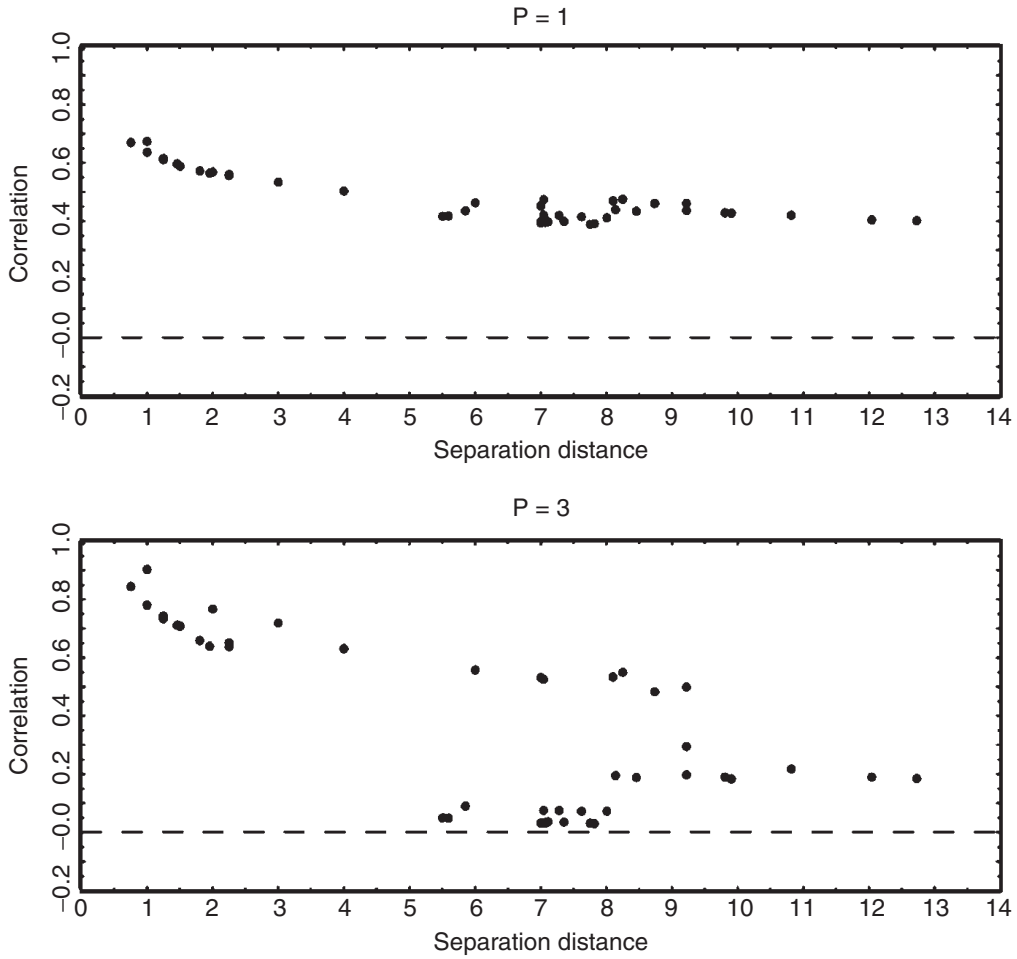


Figure 8.29 Correlograms for inverse distance model.

Table 8.16 Standardized weight matrix for negative exponential model, $A = 0.5$

0.00	0.88	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.98	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.87	0.12	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
0.02	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.28	0.46	0.17	0.10	0.00
0.00	0.00	0.00	0.00	0.53	0.00	0.32	0.04	0.11	0.00
0.00	0.00	0.00	0.00	0.60	0.22	0.00	0.15	0.03	0.00
0.00	0.00	0.00	0.00	0.49	0.07	0.32	0.00	0.12	0.00
0.00	0.00	0.00	0.00	0.46	0.25	0.10	0.19	0.00	0.00
0.00	0.00	0.00	0.01	0.07	0.01	0.10	0.79	0.03	0.00

Figure 8.31 shows correlograms for $A = 0.5$ and $A = 2$. When $A = 0.5$, the pattern is familiar: high correlations at small separation distances, a range between 5 and 9 where strong correlations are interspersed with zero correlations, and zero correlations at separation distances greater than 9. When $A = 2$, previously strong correlations are reduced somewhat, and the previously zero correlations become stronger.

Table 8.17 Standardized weight matrix for negative exponential model, $A = 2$

0.00	0.51	0.31	0.04	0.03	0.02	0.02	0.02	0.05	0.00
0.59	0.00	0.22	0.03	0.03	0.03	0.02	0.02	0.06	0.00
0.42	0.25	0.00	0.15	0.03	0.03	0.02	0.03	0.06	0.01
0.18	0.11	0.48	0.00	0.04	0.03	0.03	0.05	0.06	0.04
0.01	0.01	0.01	0.00	0.00	0.25	0.28	0.22	0.20	0.01
0.01	0.02	0.01	0.00	0.31	0.00	0.27	0.16	0.21	0.01
0.01	0.01	0.01	0.00	0.33	0.25	0.00	0.23	0.15	0.01
0.02	0.01	0.02	0.01	0.29	0.18	0.26	0.00	0.20	0.02
0.04	0.03	0.03	0.01	0.26	0.23	0.18	0.21	0.00	0.01
0.02	0.02	0.04	0.10	0.15	0.10	0.17	0.28	0.12	0.00

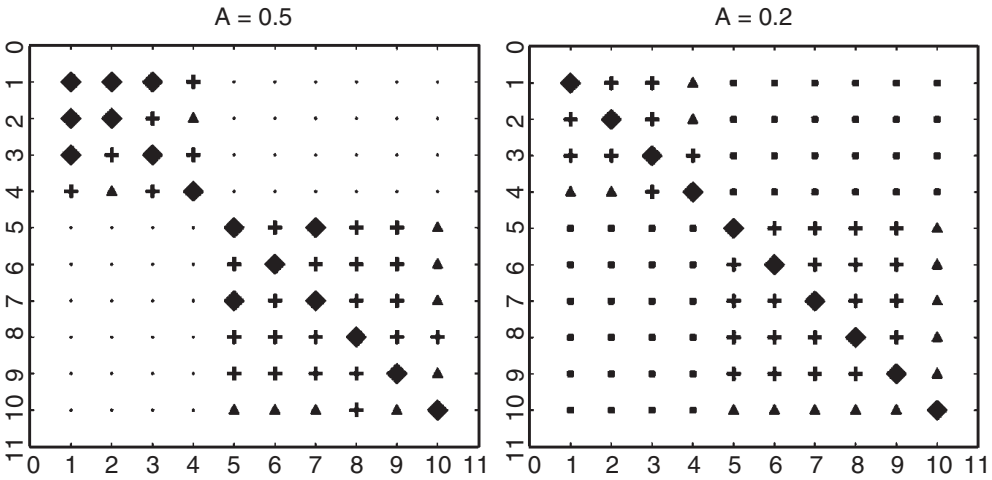


Figure 8.30 Correlation matrices for negative exponential model, selected values of *A*.

8.7. IRREGULARLY LOCATED AREAS

All of the weighting schemes described in the previous section can be used for areas, if they are applied to the centroids of the regions. Other weighting schemes have been suggested for areas. For example, Cliff and Ord (1981) suggest using weights based on centriod separation distance and the length of the shared boundary.

$$W_{ij} = \frac{\beta_{i(j)}^{b}}{D_{ij}^{a}}$$

where $\beta_{i(j)}$ is the proportion of the perimeter of area i that is shared with area j, and a and b are parameters. Dacey (1968) suggested taking the relative size of each area into consideration, and proposed the following weights:

$$W_{ij} = d_{ij}\alpha_i\beta_{i(j)}$$

where $d_{ij}$is one if the areas are contiguous and zero otherwise, and α_i is the fraction of the study area that is contained in area i. Many other weighting schemes have been

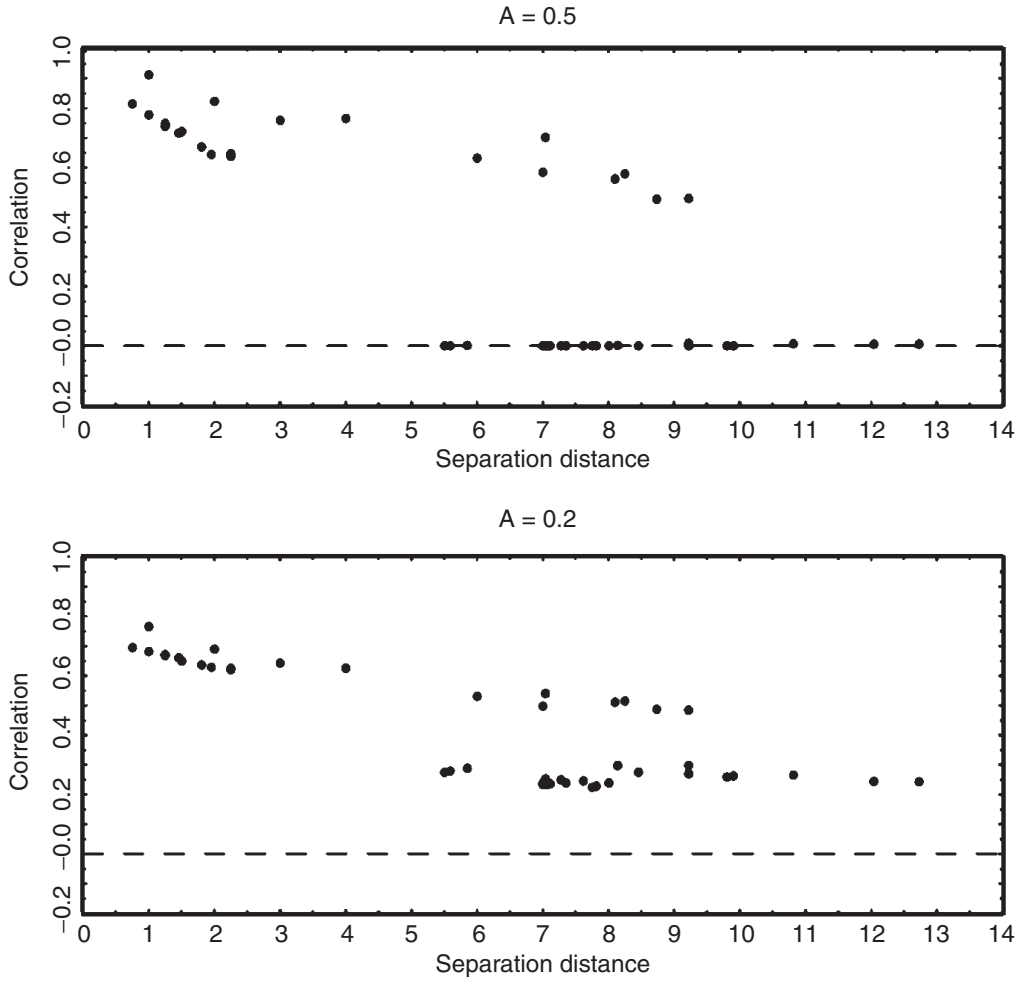


Figure 8.31 Correlograms for negative exponential model.

proposed. These will not be explored further in this chapter.

8.8. DISCUSSION

The weight matrix is a powerful tool for representing spatial relationships. There are many choices for the form that this matrix can take; only a few have been described in this chapter. The researcher will always have to specify a family of schemes (e.g., nearest neighbor, limit) and will often have to choose at least one parameter to complete the specification. Despite this, it is standard to treat the weight matrix as exogenous, which means that both the family and parameter value are known by the researcher. While this makes the estimation of other parameters easier, it is not very satisfying, since it implies that the researcher knows a great deal about the spatial interactions in the data. Furthermore, the remaining parameter estimates can be biased, since they will be conditional upon the specification of W. Given the impact of the choice of family and parameters upon the analysis, it is incumbent upon the researcher to choose carefully.

Estimation of W is appealing, although difficult. Maximum likelihood methods can be used, at the cost of assuming normality. Bhattacharjee and Jensen-Butler (2006) have

recently suggested an approach that is based on the eigenvalues and eigenvectors of the variance/covariance matrix estimated from a first stage OLS regression. Clearly, this is an area for future research. For further reading see Anselin (1988), Upton and Fingleton (1985), and Cliff and Ord (1973).

NOTE

1 It should be pointed out that it is probably as easy to estimate the number of neighbors as it is to estimate α.

REFERENCES

Anselin, L. (1988). *Spatial Econometrics: Method and Models*. Dordrecht: Kluwer.

Bhattacharjee, A. and Jensen-Butler, C. (2006). Estimation of spatial weights matrix in a spatial error model, with an application to diffusion in housing demand, *CRIEFF Discussion Paper 0519*. This paper can be downloaded from http://ideas.repec.org/p/san/crieff/0519.html

Cliff, A.D. and Ord, J.K. (1981). *Spatial Processes: Models and Applications*. London: Pion.

Cliff, A.D. and Ord, J.K. (1973). *Spatial Autocorrelation*. London: Pion

Dacey, M. (1968). "A Review of Measure of Contiguity for Two and K-Color Maps." In *Spatial Analysis: A Reader in Statistical Geography*, edited by B. Berry and D. Marble, pp. 479–95. Englewood Cliffs, N.J.: Prentiss-Hall.

Pace, K. and Gilley, O. (1998). Generalizing the OLS and grid estimator. *Real Estate Economics*, pp. 331–347.

Upton, G. and Fingleton, B. (1985). *Spatial Data Analysis by Example*. New York: Wiley.

9

Geostatistics and Spatial Interpolation

Peter M. Atkinson and Christopher D. Lloyd

9.1. INTRODUCTION

This chapter is concerned with geostatistics, a set of techniques for the analysis of spatial data (Journel and Huijbregts, 1978; Goovaerts, 1997; Chilès and Delfiner, 1999). Oliver and Webster (1990) and Burrough and McDonnell (1998) are two accessible introductions to geostatistics, the latter describing geostatistics within the context of geographical information systems. Geostatistics has its origins in mining but geostatistical approaches have been applied in many other disciplines including glaciology (Herzfeld and Holmlund, 1990), remote sensing (Curran and Atkinson, 1998) and archaeology (Lloyd and Atkinson, 2004). Geostatistics is characterized by the common dependence of its constituent techniques on the random function (RF) model, described below. Such techniques include those for spatial prediction, spatial simulation, regularization and spatial optimization. Commonly, the RF model is defined to be stationary in the sense that the parameters of the model are invariant through space. In this chapter, the focus of later discussion is on non-stationarity of parameters through space, in keeping with the local spatial analysis described in other chapters of this book.

A RF $Z(\mathbf{x})$ may be defined as a random variable (RV), that is, a stochastic process Z that varies as a function of location $\mathbf{x}$. The process of rolling a six-sided die is a commonly quoted example of a RV. The die can take any one of six possible outcomes (an integer between 1 and 6) where, for an unbiased die, each number has an equal chance of 1/6 of being rolled.

These possible outcomes define the discrete distribution function of the die. Rolling the die leads to a particular outcome, called a realization. For continuous variables, a continuous function (the probability density function, pdf, or cumulative distribution function, cdf) replaces the discrete definition of the distribution function. The cdf defines the probability of the outcome being less than a selected value (Goovaerts, 1997). See Isaaks and Srivastava (1989) for a discussion of RVs in a geostatistical context.

In defining a RF it is important to consider how the RV will be allowed to vary through space **x**. One simple possibility is to allocate to every position **x** in space its own cdf, with each independent of all other cdfs. A problem is that this model requires a large number of parameters; one set (e.g., mean and variance of the Gaussian model) for each possible location. Moreover, such a possibility is unlikely to be realistic in practice; we know that places close together tend to have similar characteristics. Therefore, this model is too loosely controlled and does not make use of our practical knowledge of spatially varying phenomena. For these reasons, we place some restrictions on the RF model. The most common set of restrictions are referred to as stationarity constraints, meaning that particular parameters are invariant with **x**. In the strictest sense, the mean and variance parameters can be held constant for all locations **x**. However, under this model each point is identical, independent distributed (iid), meaning that spatial inference is severely limited (we now have too tight a control over the possibilities).

In geostatistics, it is common to define a stationary mean parameter. Various alternative models have been proposed in which the mean is allowed to vary through space. Such a non-stationary mean parameter is generally referred to as a trend (see Goovaerts, 1997, and section 9.4.1 below). For the present purpose, a stationary mean provides a basic starting point. A second important restriction, which is not as restrictive as defining a stationary variance, is to define a stationary spatial covariance function (representing second-order stationarity) or semi-variogram (representing intrinsic stationarity, a weaker form of stationarity). Although much of the computation in geostatistics is based on the spatial covariance, the equations are often written in terms of semi-variograms and, thus, we shall focus on the semi-variogram from this point onwards.

The semi-variogram defines the relations between points and, thus, facilitates spatial statistical inference. It is usually estimated from empirical data as a plot of half the average squared difference between pairs of values (the semivariance) against the vector separation or lag. Then a mathematical model is commonly fitted to the empirical semi-variogram plot for use in geostatistical operations. Various methods may be employed in the fitting, although weighted least squares is a common basic starting point. Several important considerations should be taken into account during model fitting (see McBratney and Webster, 1986). Once the parameters are estimated (either with or without the uncertainty of estimation accounted for) the RF is defined and geostatistical operations can proceed. Variogram estimation and model fitting are described in section 9.2.

The mean and semi-variogram are, thus, the parameters that define the RF model, and that need to be estimated, effectively replacing the mean and variance of the RV model. It should be pointed out that the variogram may itself be comprised of several further parameters. For example, the spherical model is an example of a transitive variogram model (i.e., for which a positive finite maximum value is defined).

The spherical model has two parameters; the sill c and the non-linear parameter a usually referred to as the range. The sill defines the maximum value of semivariance while the range defines the lag at which the sill is reached.

Geostatistical operations include spatial prediction, spatial simulation, regularization and spatial optimization. In spatial prediction or kriging, the objective is to predict the value of $z(\mathbf{x}_0)$ at some unobserved location $\mathbf{x}_0$ given a sample of data $z(\mathbf{x}_i)$, $i = 1, 2, \ldots, n$ usually defined on point supports (the space on which each observation is defined) or quasi-points. The RF model helps because it is useful to base the prediction of $z(\mathbf{x}_0)$ on a model that captures our knowledge of the underlying processes or form. In environmental science (in the broadest sense) process knowledge is often limited and the RF model provides a useful stochastic framework that builds on some general principles.

The RF model is useful for several reasons, but prime among them are:

1 the dependence of the prediction $z(\mathbf{x}_0)$ on the data $z(\mathbf{x}_i)$, $i = 1, 2, \ldots, n$ is estimated by the semi-variogram. In a general sense, the closer $z(\mathbf{x}_0)$ and a given data point the more similar the two values are likely to be. The semi-variogram quantifies this spatial dependence. Critically, this means in a linear weighting of proximate data to be used in spatial prediction the weights can be determined automatically through linear algebra. This process is referred to as kriging.

2 In kriging, the relations between the sample data themselves are accounted for so that, at a given separation, a cluster of data points will contribute less to the prediction than a dispersed set (Journel and Huijbregts, 1978).

3 The cdf of the predicted value (i.e., the set of possible values from which one realization is drawn) can be conditioned on the sample data. In particular, the variance of the conditional cdf (ccdf) is likely to be less than that of the original cdf. In general terms, this means that the range of possible values for the unknown value is restricted to be close to the neighbouring data by an amount determined by the spatial proximity of the prediction location to the neighbours. Such information can be used to extend the process of spatial prediction (in which the mean of the posterior or conditional cdf is drawn) to spatial simulation (in which a value is drawn from the ccdf randomly).

Geostatistics, as described above, has been used widely to characterize spatial variation (using the semi-variogram or other function) in relatively small data sets and to predict unobserved values using kriging informed by the modelled semi-variogram. In such circumstances, the decision to adopt a stationary model of the mean and semi-variogram makes sense. In fact, it is necessary for statistical inference. However, very large spatially-extensive and spatially-detailed data sets are increasingly readily available. Examples include digital elevation data and image-based data sets provided primarily through remote sensing (Atkinson, 2005). Researchers and practitioners are increasingly overwhelmed by the magnitudes of the datasets available for analysis. This has led to a realization that the biggest problem facing spatial analysts today is one of data richness rather than data sparcity. In these circumstances, a stationary RF model is not only inappropriate, but wasteful of data. More suitable solutions can be found by allowing the previously stationary parameters to vary across the region of interest (Atkinson, 2001).

The present chapter provides an introduction to linear geostatistics, but with a particular focus on models that include non-stationarity parameters, particularly of (a) the mean and (b) the semi-variogram. The next

section describes the process of fitting the RF model parameterized by a spatial covariance or semi-variogram, while section 9.3 describes geostatistical prediction (kriging). Section 9.4 considers non-stationary models and section 9.5 discusses a range of issues related to the use of geostatistics within GIS.

9.2. CHARACTERIZING SPATIAL VARIATION

9.2.1. Estimating the experimental semi-variogram

Much of the effort and time associated with geostatistical analysis is expended in analysis of the spatial structure of a variable. One simple way of examining spatial structure is through estimating the semi-variogram cloud. The semi-variogram cloud is a plot of the semivariances for paired data against the distances separating the paired data points in a given direction. The semivariance is half the squared difference between values at two locations, and can be thought of as a measure of dissimilarity. Thus, the semi-variogram cloud shows how dissimilar paired data points are as a function of their separation distance and direction (termed spatial lag, $\mathbf{h}$). If data are spatially structured then pairs separated by small lags will tend to be less dissimilar than pairs separated by large lags.

A core idea in geostatistics is that the spatial structure in a variable should be characterized and used for spatial prediction and simulation. The objective of geostatistical prediction is to find optimal weights to assign to observations located around the prediction location. If information is available on how dissimilar two observations are likely to be for a given lag then this information can be used to determine these weights. The most commonly used approach is based on the estimated semi-variogram. The experimental semi-variogram is estimated by calculating the squared differences between all the available paired observations and obtaining half the average for all observations separated by a given lag (or within a lag tolerance where the observations are not on a regular grid). So, while the semi-variogram cloud provides semivariances as a function of a set of actual lags the experimental semi-variogram provides only a set of average semivariances at a set of discrete lags. Examination of the semi-variogram cloud provides a means of identifying heterogeneities in spatial variation within a variable (Webster and Oliver, 2000) that are obscured through the summation over lags that occurs with the experimental semi-variogram. Therefore, examination of the semi-variogram cloud is a sensible step prior to estimation of the experimental semi-variogram.

The experimental semi-variogram, $\hat{\gamma}(\mathbf{h})$, can be estimated from $p(\mathbf{h})$ paired observations, $z(\mathbf{x}_\alpha)$, $z(\mathbf{x}_\alpha + \mathbf{h})$, $\alpha = 1, 2, \ldots, p(\mathbf{h})$ using:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2\,p(\mathbf{h})} \sum_{\alpha=1}^{p(\mathbf{h})} \left\{ z(\mathbf{x}_\alpha) - z(\mathbf{x}_\alpha + \mathbf{h}) \right\}^2 \tag{9.1}$$

The semi-variogram can be estimated for different directions to enable the identification of directional variation (termed anisotropy). Where a variable is preferentially sampled in areas with large or small values of the property of interest, the histogram will be unrepresentative and often a declustering algorithm is necessary to correct this. For example, values in areas or cells with more data may be given smaller weights than values in sparsely sampled areas (Deutsch and Journel, 1998). Preferential sampling

of a variable also impacts on the form of the experimental semi-variogram. Richmond (2002) shows that clustering can, in some cases, alter drastically the form of the semi-variogram. Two methods of declustering for weighting paired data in estimation of the experimental semi-variogram are given by Richmond (2002).

In the presence of large-scale, low-frequency variation (e.g., that would be fitted well by a trend model), the form of the semi-variogram will be affected. If the semi-variogram increases more rapidly than a quadratic polynomial for large lags then a RF which is non-stationary in the mean should be adopted (Armstrong, 1998). This topic is explored in greater depth in section 9.4.1.

9.2.2. Fitting a semi-variogram model

A mathematical model may be fitted to the experimental semi-variogram and the coefficients of this model can be used for a range of geostatistical operations such as spatial prediction (kriging) and conditional simulation. A model is usually selected from one of a set of so-called authorized models. McBratney and Webster (1986) provide a review of some of the most widely used authorized models. There are two principal classes of semi-variogram model. Transitive (bounded) models have a sill (finite variance), and indicate a second-order stationary process. Unbounded models do not reach an upper bound; they are intrinsically stationary only (McBratney and Webster 1986). Figure 9.1 shows the parameters of a bounded semi-variogram model (the spherical model as defined below). The nugget effect, c_0, represents unresolved variation (a mixture of spatial variation at a scale finer than the sample spacing and measurement error). The sill, c, represents the spatially correlated variation. The total sill, $c_0 + c$, is the *a priori* variance. The range, a, represents the scale of spatial variation (Atkinson and Tate, 2000). For example, if a measured property varies markedly over small distances then the property can be said to exhibit short range spatial variation.

Some of the most commonly used authorized models are detailed below. The nugget effect model, defined above, is given by:

$$\gamma(h) = \begin{cases} 0 & \text{for } h = 0 \\ c_0 & \text{for } |h| > 0. \end{cases} \tag{9.2}$$

Three of the most frequently used bounded models are the spherical model, the exponential model and the Gaussian model and these are defined in turn. The spherical model is perhaps the most widely used semi-variogram model. Its form corresponds closely with what is often observed in many real world studies; almost linear growth in semivariance with separation and then stabilization (Armstrong, 1998). It is given by:

$$\gamma(h) = \begin{cases} c \cdot [1.5(h/a) - 0.5(h/a)^3] & \text{if } h \leq a \\ c & \text{if } h > a \end{cases} \tag{9.3}$$

where c is the sill of the spherical model and a is the non-linear parameter, known as the range.

The exponential model is given by:

$$\gamma(h) = c \cdot \left[1 - \exp\left(-\frac{h}{d}\right)\right] \tag{9.4}$$

where d is the non-linear distance parameter. The exponential model reaches the sill asymptotically and the practical range is $3d$

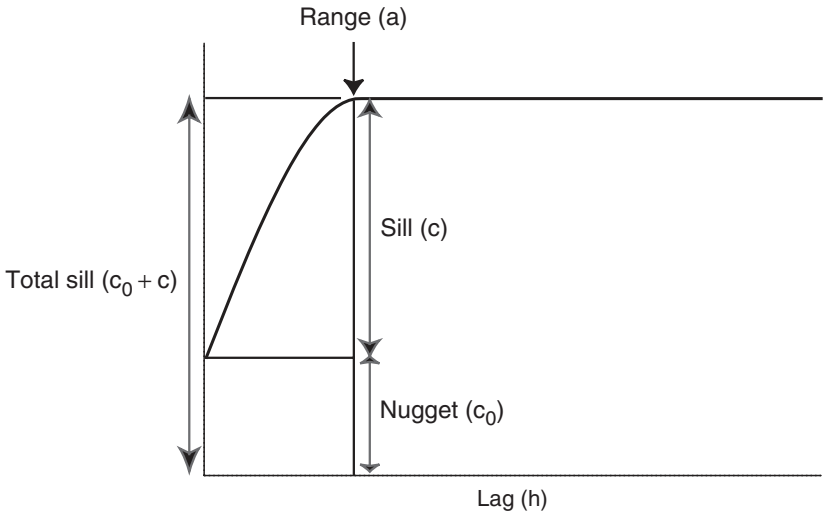


Figure 9.1 Bounded semi-variogram.

(i.e., the separation at which approximately 95% of the sill is reached).

The Gaussian model is given by:

$$\gamma(h) = c \cdot \left[1 - \exp\left(-\frac{h^2}{d^2}\right)\right]. \qquad (9.5)$$

The Gaussian model does not reach a sill at a finite separation and the practical range is $a\sqrt{3}$ (Journel and Huijbregts, 1978). Semi-variograms with parabolic behaviour at the origin, as represented by the Gaussian model here, are indicative of very regular spatial variation (Journel and Huijbregts, 1978). Authorized models may be used in positive linear combination where a single model is insufficient to represent well the form of the semi-variogram.

Where the semi-variogram appears to increase indefinitely with separation the most widely used model is the power model:

$$\gamma(h) = m \cdot h^{\omega} \qquad (9.6)$$

where ω is a power $0 < \omega < 2$ with a positive slope, m (Deutsch and Journel, 1998). The linear model is a special case of the power model.

One of the advantages of kriging is that it is often fairly straightforward to model anisotropic structure using the semi-variogram. Two primary forms of anisotropy have been outlined in the geostatistical literature. If the sills for all directions are not significantly different and the same structural components (for example, spherical or Gaussian) are used then anisotropy can be accounted for by a linear transformation of the co-ordinates: this is called geometric or affine anisotropy (Webster and Oliver, 1990). Where the sill changes with direction but the range is similar for all directions the anisotropy is called zonal (Isaaks and Srivastava, 1989). However, the modelling of zonal anisotropy is much more problematic than the modelling of geometric anisotropy. In practice, a mixture of geometric and zonal anisotropy has been found to be common (Isaaks and Srivastava, 1989).

There are various approaches for fitting models to semi-variograms. Some geostatisticians

prefer fitting semi-variogram models 'by eye' on the grounds that it enables one to use personal experience and to account for features or variation that may be difficult to quantify (Christakos, 1984; Journel and Huijbregts, 1978). Weighted least squares (WLS) has been proposed as a suitable means of fitting models to semi-variograms (Cressie, 1985; Pardo-Igúzquiza, 1999) and the approach has been used by many geostatisticians. The technique is preferred to unweighted ordinary least squares (OLS) as in WLS the weights can be used to reflect the uncertainty in the individual semivariance estimates or the desire to fit at certain lags more accurately than at others. For example, the weights are often selected to be proportional to the number of pairs at each lag (Cressie, 1985), such that lags with many pairs have greater influence in the fitting of the model. The use of generalized least squares (GLS) has also been demonstrated in a geostatistical context (Cressie, 1985; McBratney and Webster, 1986). Use of maximum likelihood (ML) estimation (McBratney and Webster, 1986) has become widespread amongst geostatisticians and has been used for WLS. The goodness of fit of models to the semi-variogram, and of the relative improvement or otherwise in using different numbers of parameters, may be compared through the examination of the sum of squares of the residuals or through the use of the Akaike Information Criterion (McBratney and Webster, 1986; Webster and McBratney, 1989).

Figure 9.2 shows an experimental semi-variogram estimated from precipitation data acquired in Great Britain in January 1999. The semi-variogram was estimated using the Gstat software (Pebesma and Wesseling, 1998). The data are described by Lloyd (2002, 2005). The semi-variogram was fitted with a nugget and two spherical components. Authorized models are often used in combination in this way to model nested spatial structures. In Figure 9.3, the directional semi-variogram, estimated from the same data, is shown. It indicates that the scale of spatial variation is similar in all directions while the magnitude of the variation (the semivariance) is clearly different for different directions.

9.3. SPATIAL PREDICTION AND SIMULATION

9.3.1. Ordinary kriging

There are many varieties of kriging. Its simplest form is called simple kriging (SK). To use SK it is necessary to know the mean of the property of interest and this must be modelled as constant across the region of interest. In practice, this model is often unsuitable. The most widely used variant of kriging, ordinary kriging (OK), allows the mean to vary spatially: the mean is modelled as constant within each prediction neighbourhood only. For each point to be predicted a new neighbourhood is defined and so effectively the mean is allowed to vary locally.

OK predictions are weighted averages of the n available data. The OK weights define the best linear unbiased predictor (BLUP). The OK prediction, $\hat{z}_{\mathrm{OK}}(\mathbf{x}_0)$, is defined as:

$$\hat{z}_{\mathrm{OK}}(\mathbf{x}_0) = \sum_{\alpha=1}^{n} \lambda_\alpha^{\mathrm{OK}} z(\mathbf{x}_\alpha) \qquad (9.7)$$

with the constraint that the weights, $\lambda_\alpha^{\mathrm{OK}}$, sum to 1 to ensure an unbiased prediction:

$$\sum_{\alpha=1}^{n} \lambda_\alpha^{\mathrm{OK}} = 1. \qquad (9.8)$$

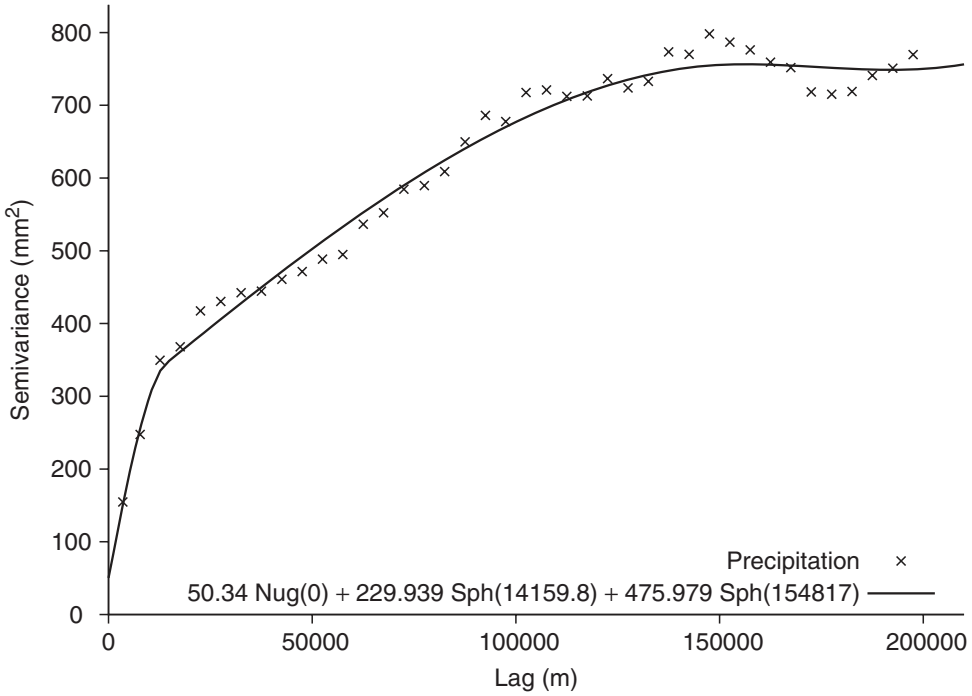


Figure 9.2 Omnidirectional semi-variogram of precipitation.

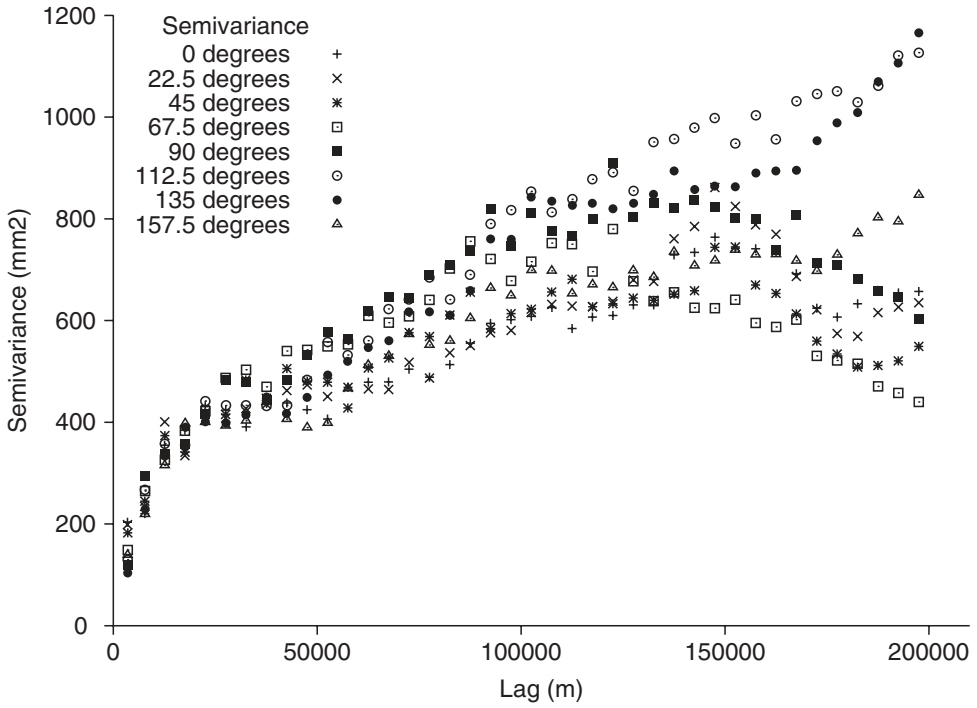


Figure 9.3 Directional semi-variogram of precipitation.

So, the objective of the kriging system is to find appropriate weights by which the available observations will be multiplied before summing them to obtain the predicted value. These weights are determined using the coefficients of a model fitted to the semi-variogram (or another function such as the covariance function).

The kriging prediction error must have an expected value of 0:

$$E\{\hat{Z}_{\mathrm{OK}}(\mathbf{x}_0) - Z(\mathbf{x}_0)\} = 0. \qquad (9.9)$$

The kriging (or prediction) variance, σ^2_{OK}, is expressed as:

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{OK}}(\mathbf{x}_0) &= E[\{\hat{Z}_{\mathrm{OK}}(\mathbf{x}_0) - Z(\mathbf{x}_0)\}^2] \\ &= 2\sum_{\alpha=1}^{n} \lambda_\alpha^{\mathrm{OK}}\, \gamma(\mathbf{x}_\alpha - \mathbf{x}_0) \\ &\quad - \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_\alpha^{\mathrm{OK}}\, \lambda_\beta^{\mathrm{OK}}\, \gamma(\mathbf{x}_\alpha - \mathbf{x}_\beta).\end{aligned} \qquad (9.10)$$

That is, we seek the values of $\lambda_1, \ldots, \lambda_n$ (the weights) that minimize this expression with the constraint that the weights sum to one (equation (9.8)). This minimization is achieved through Lagrange multipliers. The conditions for the minimization are given by the OK system comprising $n + 1$ equations and $n + 1$ unknowns:

$$\begin{cases} \sum_{\beta=1}^{n} \lambda_\beta^{\mathrm{OK}}\, \gamma(\mathbf{x}_\alpha - \mathbf{x}_\beta) + \psi_{\mathrm{OK}} = \gamma(\mathbf{x}_\alpha - \mathbf{x}_0) \\ \qquad \alpha = 1, \ldots, n \\ \sum_{\beta=1}^{n} \lambda_\beta^{\mathrm{OK}} = 1 \end{cases} \qquad (9.11)$$

where ψ_{OK} is a Lagrange muliplier. Knowing ψ_{OK}, the kriging variance, an estimator of the prediction variance of OK, can be given as:

$$\hat{\sigma}^2_{\mathrm{OK}} = \sum_{\alpha=1}^{n} \lambda_\alpha^{\mathrm{OK}}\, \gamma(\mathbf{x}_\alpha - \mathbf{x}_0) + \psi_{\mathrm{OK}}. \qquad (9.12)$$

The kriging variance is a measure of confidence in predictions and is a function of the form of the semi-variogram, the sample configuration and the sample support (Journel and Huijbregts, 1978). The kriging variance is not conditional on the data values locally and this has led some researchers to use alternative approaches such as conditional simulation (discussed in the next section) to build models of spatial uncertainty (Goovaerts, 1997).

There are two varieties of OK: punctual OK and block OK. With punctual OK the predictions cover the same area (the support, $\mathbf{v}$) as the observations. In block OK, the predictions are made to a larger support than the observations. With punctual OK the data are honoured. That is, they are retained in the output map. Block OK predictions are averages over areas (that is, the support has increased). Thus, at $\mathbf{x}_0$ the prediction is not the same as an observation and does not need to honour it.

The choice of semi-variogram model affects the kriging weights and, therefore, the predictions. However, if the form of two models is similar at the origin of the semi-variogram then the two sets of results may be similar (Armstrong, 1998). The choice of nugget effect may have marked implications for both the predictions and the kriging variance. As the nugget effect is increased, the predictions become closer to the global average (Isaaks and Srivastava, 1989).

A map of precipitation in Britain in January 1999 generated using OK is shown

in Figure 9.4. It was generated using the semi-variogram model given in Figure 9.2 and the 16 nearest neighbours to each grid cell were used in the prediction process. The map is very smooth in appearance; this is a common feature of maps derived using OK.

9.3.2. Cokriging

Where a secondary variable (or variable) is available that is cross-correlated with the primary variable both variables may be used simultaneously in prediction using cokriging. To apply cokriging, the semi-variograms (that is, auto semi-variograms) of both variables and the cross semi-variogram (describing the spatial dependence between the two variables) are required. The operation of cokriging is based on the linear model of coregionalization (see Webster and Oliver, 2000). For cokriging to be beneficial, the secondary variable should be cheaper to obtain or more readily available to make the most of the technique. If the variables are clearly linearly related then cokriging may estimate more accurately than, for example, OK.

9.3.3. Conditional simulation

Kriging predictions are weighted moving averages of the available sample data. Kriging is, therefore, a smoothing interpolator. Conditional simulation (also called stochastic imaging) is not subject to the smoothing associated with kriging (conceptually, the variation lost by kriging due to smoothing is added back) as predictions are drawn from equally probable joint realizations of the RVs which make up a RF model (Deutsch and Journel, 1998). That is, simulated values are not the expected values (i.e., the mean) but are values drawn randomly from the conditional cdf: a function of the available observations and the modelled spatial variation (Dungan, 1999). The simulation is considered 'conditional' if the simulated values honour the observations at their locations (Deutsch and Journel, 1998). As noted above, simulated realizations represent a possible reality whereas kriging does not. Simulation allows the generation of many different possible realizations that may be used as a guide to potential errors in the construction of a map (Journel, 1996) and multiple realizations encapsulate the uncertainty in spatial prediction. Arguably, the most widely used form of conditional simulation is sequential Gaussian simulation (SGS). With sequential simulation, simulated values are conditional on the original data and previously simulated values (Deutsch and Journel, 1998). In SGS the ccdfs are all assumed to be Gaussian. SGS is discussed in detail in several texts (for example, Goovaerts, 1997; Deutsch and Journel, 1998; Chilès and Delfiner, 1999; Deutsch, 2002).

9.4. NON-STATIONARY MODELS

This section discusses non-stationarity in the mean and the semi-variogram. Approaches for dealing with non-stationarity in the mean are well developed and are the subject of section 9.4.1. There is a variety of methods for estimating the local semi-variogram where the spatial structure in the property of interest varies from place to place. However, such approaches are less widely used than methods that allow for non-stationarity in the mean. Some approaches for estimating the non-stationary semi-variogram are discussed in section 9.4.2.

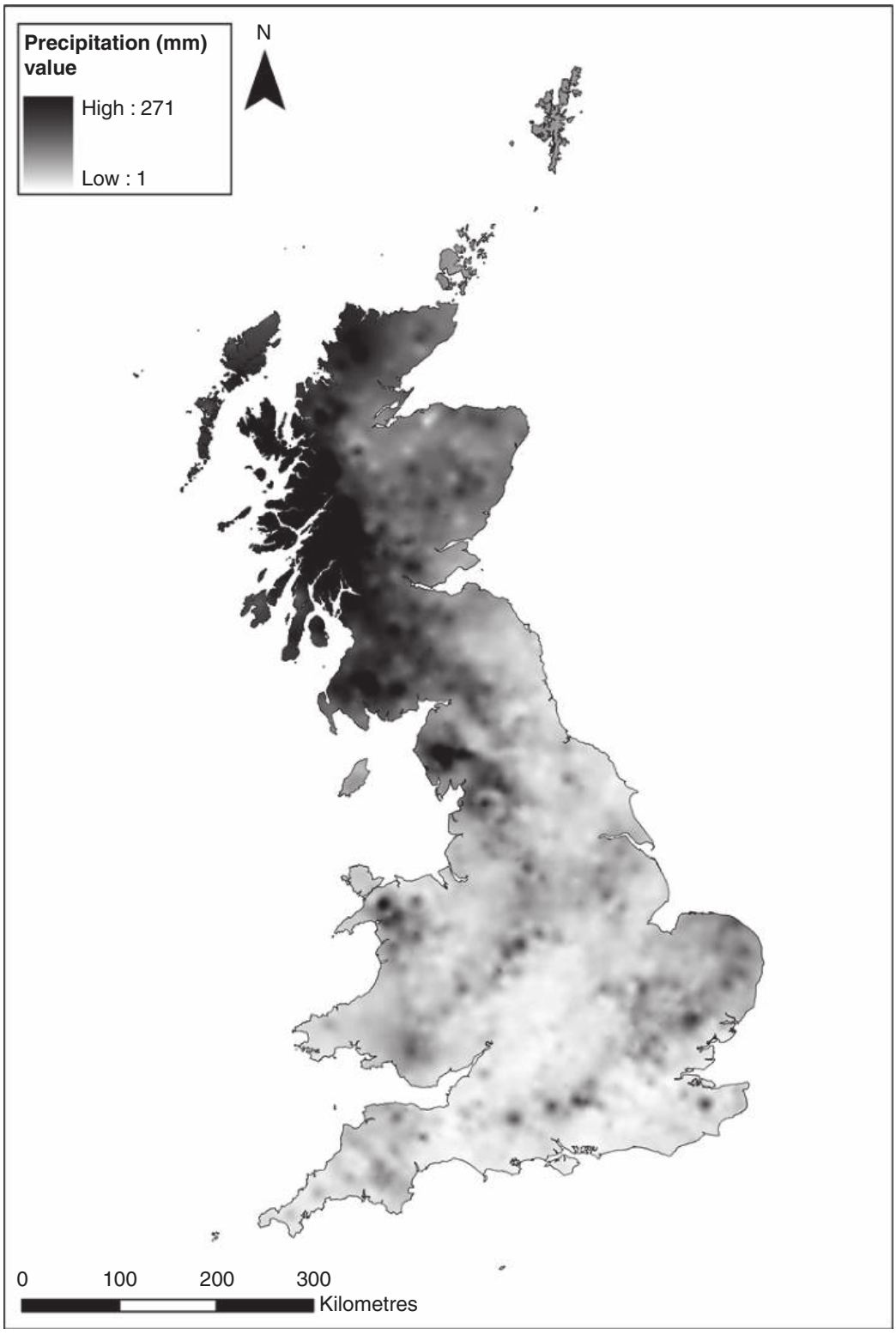


Figure 9.4 OK derived map of precipitation.

9.4.1. Non-stationary mean: fitting a trend and kriging with a trend model

OK is robust but, in some cases, an even more general form of kriging may be appropriate. In cases where the mean of the variable changes markedly over small distances a non-stationary model of the mean may provide more accurate spatial prediction. While the mean varies from place to place with OK it does not vary within the search window. Several approaches exist that provide a non-stationary mean.

One approach is to fit a global polynomial trend model and estimate the semi-variogram of the residuals. SK can then be used to make predictions after which the trend can be added back to the predicted values. Figure 9.5 shows the omnidirectional semi-variogram of raw precipitation values and of residuals from a first-order and a second-order polynomial trend. In this case, the form of each of the semi-variograms is similar although the variance decreases as a higher-order trend is removed. Another approach to obtaining the trend-free semi-variogram is to estimate the semi-variogram for several directions and retain the semi-variogram for the direction that has least evidence of trend, that is, for which the variance is smallest. Figure 9.6 shows the semi-variogram of precipitation for the direction with the smallest variance.

The most widely used approach to prediction where the mean is non-stationary is called kriging with a trend model (KT; sometimes termed universal kriging). In KT, the mean is modelled using a polynomial. The principal problem with KT is that the underlying trend-free semi-variogram must

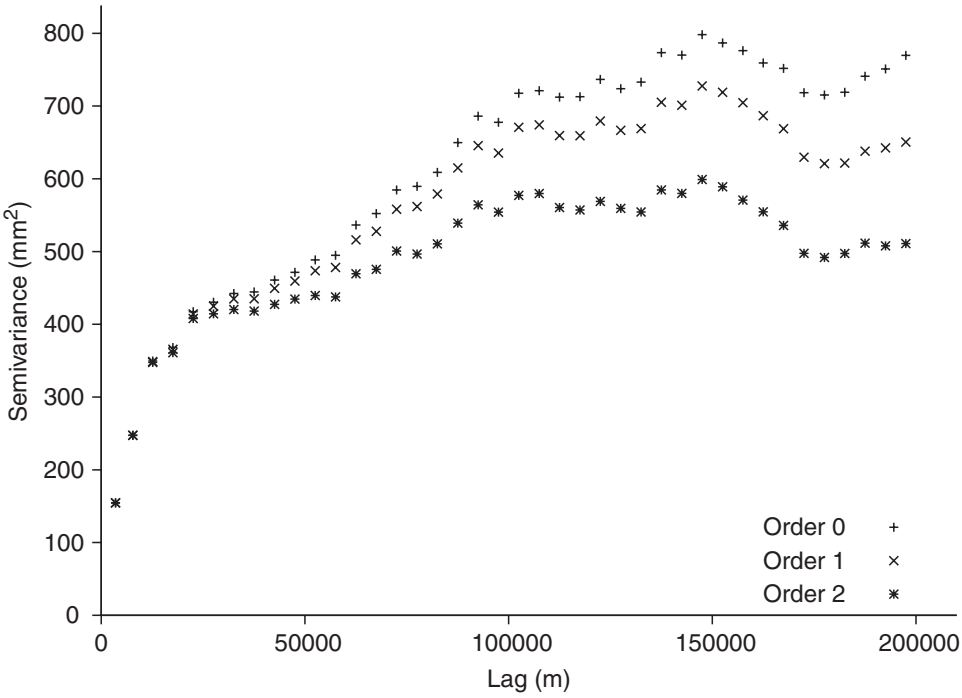


Figure 9.5 Semi-variogram of precipitation: raw data (order 0) and residuals from a polynomial trend of order 1 and 2.

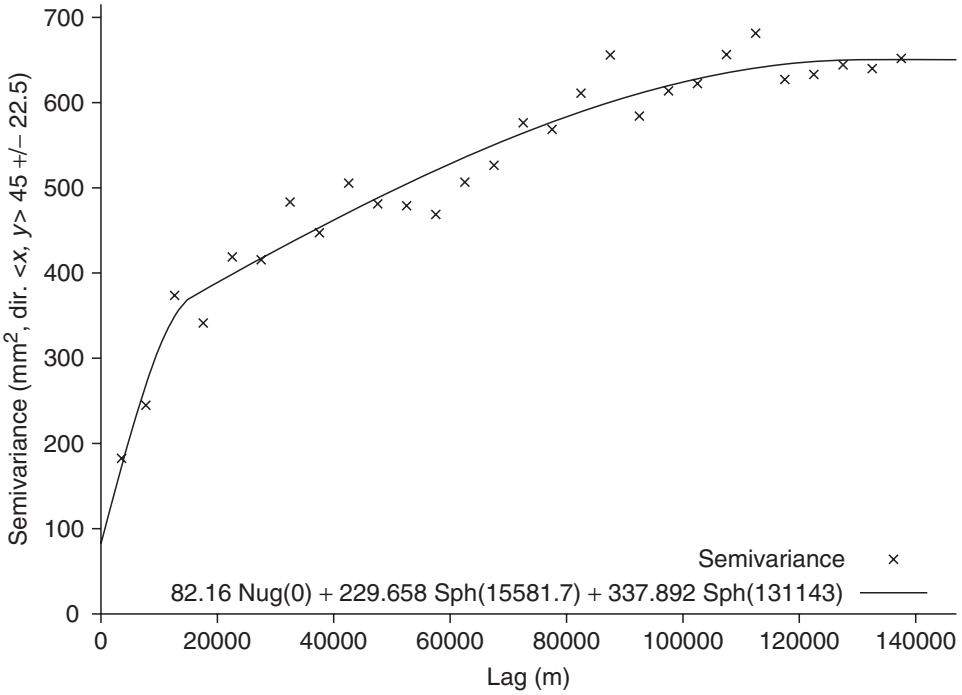


Figure 9.6 Semi-variogram of precipitation: direction with the smallest variance.

be estimated yet the local trend (or drift) is estimated as a part of the KT procedure which itself requires the semi-variogram. Various approaches for estimating the trend-free semi-variogram are described in the literature and two approaches are summarized above. Figure 9.7 shows the KT predictions made using 16 nearest neighbours with the semi-variogram model given in Figure 9.6; the semi-variogram for the direction with the least evidence of trend. An alternative approach is Intrinsic Random Functions of Order k kriging whereby the generalized covariance is used in place of the semi-variogram (Chilès and Delfiner, 1999).

Making use of secondary variables: KED and SKlm

As well as estimating the form of the trend from the variable of interest, there are various approaches that make use of secondary variables that describe the shape of the mean in the primary variable. If some variable is available that is linearly related to the primary variable and varies smoothly (i.e., there are no marked local changes in values) it could be used to inform spatial prediction of values of the primary variable. Two such approaches are described below.

With SK, the mean is assumed to be constant (there is no systematic change in the mean of the property across the region of study) and known. If the mean is not constant, but we can estimate the mean at locations in the domain of interest, then this locally varying mean can be used to inform prediction. That is, the local mean can be estimated prior to kriging. The locally-varying mean can be estimated in various different ways. One approach, termed simple kriging with locally varying means (SKlm), is to use regression to estimate the value of the primary

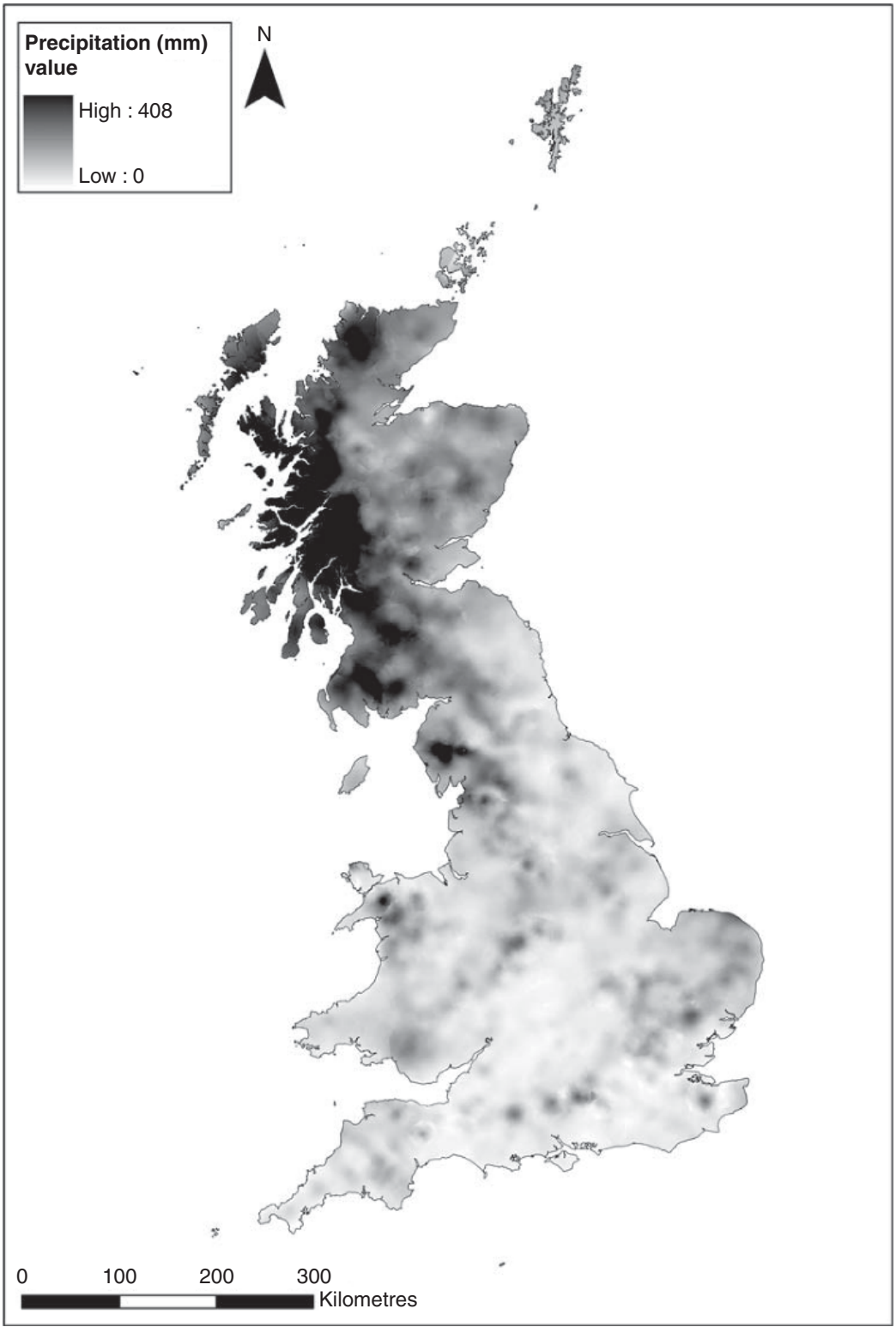


Figure 9.7 KT derived map of precipitation.

variable at (a) all observation locations and (b) all locations where SKlm predictions will be made. The semi-variogram is then estimated using the residuals from the regression predictions at the data locations. SKlm is conducted using the residuals and the trend is added back after the prediction process is complete (an example is given by Lloyd, 2005).

An alternative approach is kriging with an external drift model (KED). In KED, the secondary data act as a shape function (the external trend) and the function describes the average shape of the primary variable (Wackernagel, 2003). The local mean of the primary variable is derived as a part of the kriging procedure using the secondary information and SK is carried out on the residuals from the local mean. So, the approach differs from SKlm in that the local mean is estimated as part of the kriging procedure and not before it, as is the case with SKlm (Goovaerts, 1997). Lloyd (2002, 2005) illustrates the use of KED in mapping monthly precipitation whereby elevation is used as the external trend.

As noted above, a major problem with KT and KED is that the underlying (trend-free) semi-variogram is assumed known. That is, if the mean changes from place to place the semi-variogram estimated from the raw data will be biased, so it is necessary to remove the local mean and estimate the semi-variogram of the residuals. Since the trend (that is, local mean) is estimated as a part of the KED (and KT) system, which requires the semi-variogram model coefficients as inputs, we are faced with a circular problem. A potential solution is to infer the trend-free semi-variogram from paired data that are largely unaffected by any trend (Goovaerts, 1997; Wackernagel, 2003). Hudson and Wackernagel (1994), in an application concerned with mapping mean monthly temperature in Scotland, achieved this by estimating directional semi-variograms and retaining the semi-variogram for the direction that showed least evidence of trend. That is, temperature values systematically increase or decrease in one direction (there is a trend in the values), but values of temperature are more constant in the perpendicular direction. In such cases, the concern is to characterize spatial variation in the direction for which values of temperature are homogeneous. Hudson and Wackernagel (1994) assumed that the trend-free semi-variogram was isotropic and the semi-variogram for the direction selected was used for kriging.

9.4.2. Non-stationary semi-variogram

In cases where the semi-variogram does not represent well spatial variation across the whole of the region of interest some approach may be necessary to account for the change in spatial variation locally. In the geostatistical literature, there are several approaches presented for estimation of non-stationary semi-variograms. These vary from approaches that estimate and model automatically the semi-variogram in a moving window (this approach is discussed below) to approaches that transform the data so that the transformed data have a stationary semi-variogram. Reviews of some methods are provided by Sampson *et al*. (2001) and Schabenberger and Gotway (2005).

The estimation and automated modelling of local semi-variograms for kriging is one published approach that accounts for non-stationarity in the semi-variogram (Haas, 1990). This approach is employed here. The WLS semi-variogram model fitting routine presented by Pardo-Igúzquiza (1999) was used to fit models to semi-variograms

estimated in a moving window. Fortran 77 code was written to visit each observation in the precipitation dataset and estimate the semi-variogram using the n nearest neighbours to each observation. The routine of Pardo-Igúzquiza (1999) was then used to fit a model to each semi-variogram automatically with the result that there were 3037 (equal to the number of observations in the precipitation dataset) sets of semi-variogram model coefficients. The WLS routine allows the fitting of several different models, but in this case a spherical model was fitted to all of the semi-variograms. No nugget effect was fitted as it proved problematic to fit a nugget effect while at the same time obtaining a feasible range parameter.

In the example presented, the semi-variograms were estimated using the 1000 nearest neighbours to each observation. The variogram bin size was 5000 m and the number of bins was 14. In Figure 9.8, the values of the spherical model sills are mapped. There is a clear trend in values from the south (small semivariances) to the north (large semivariances) of Britain. This corresponds with expectations: the magnitude of variability in precipitation is greater in the north and west than in the south and east of Britain. In Figure 9.9 the ranges are shown. As for the sills, there is spatial variation. In the south of Britain the range values tend to be large while in the north they tend to be smaller. This suggests that precipitation amount varies less over short distances in the south than it does in the north of Britain.

It is clear that the spatial structure of precipitation in Britain varies spatially and, as such, a global semi-variogram model does not represent variability across Britain well. Use of locally-estimated and modelled semi-variograms may increase the accuracy of predictions using kriging.

9.5. DISCUSSION

9.5.1. Automatic fitting of variogram models

Fitting semi-variogram models automatically is not straightforward. In Figure 9.9, five local semi-variograms are selected and illustrated. In most of the selected cases, the model appears to fit the experimental semi-variogram well. However, in one case (the second semi-variogram from the bottom) the form of the semi-variogram is not well represented by a (single) spherical structure. This problem could be at least partially resolved by fitting several models and selecting the best fitting model. However, as the complexity of the model fitting process increases further problems can arise with automatic fitting. Generally, we have found that the use of simple constraints to guide the fitting (e.g., nugget variance is constrained to lie within a sensible range of between zero and some positive value less than half of the total sill, for very smooth variation) leads to acceptable results in the vast majority of cases.

9.5.2. Non-stationary semi-variograms and kriging

It is easy to see how the local semi-variograms estimated in section 9.4 could be used in kriging. The parameters of the local semi-variogram are inserted into the local kriging equations instead of the global parameters (Haas, 1990). There is little restriction on the variant of kriging to which this non-stationary set of semi-variogram parameters can be applied. For example, in recent years, local semi-variogram parameters have been used in local space-time kriging (Gething *et al.*, 2007) and local semi-variogram and cross semi-variogram parameters have

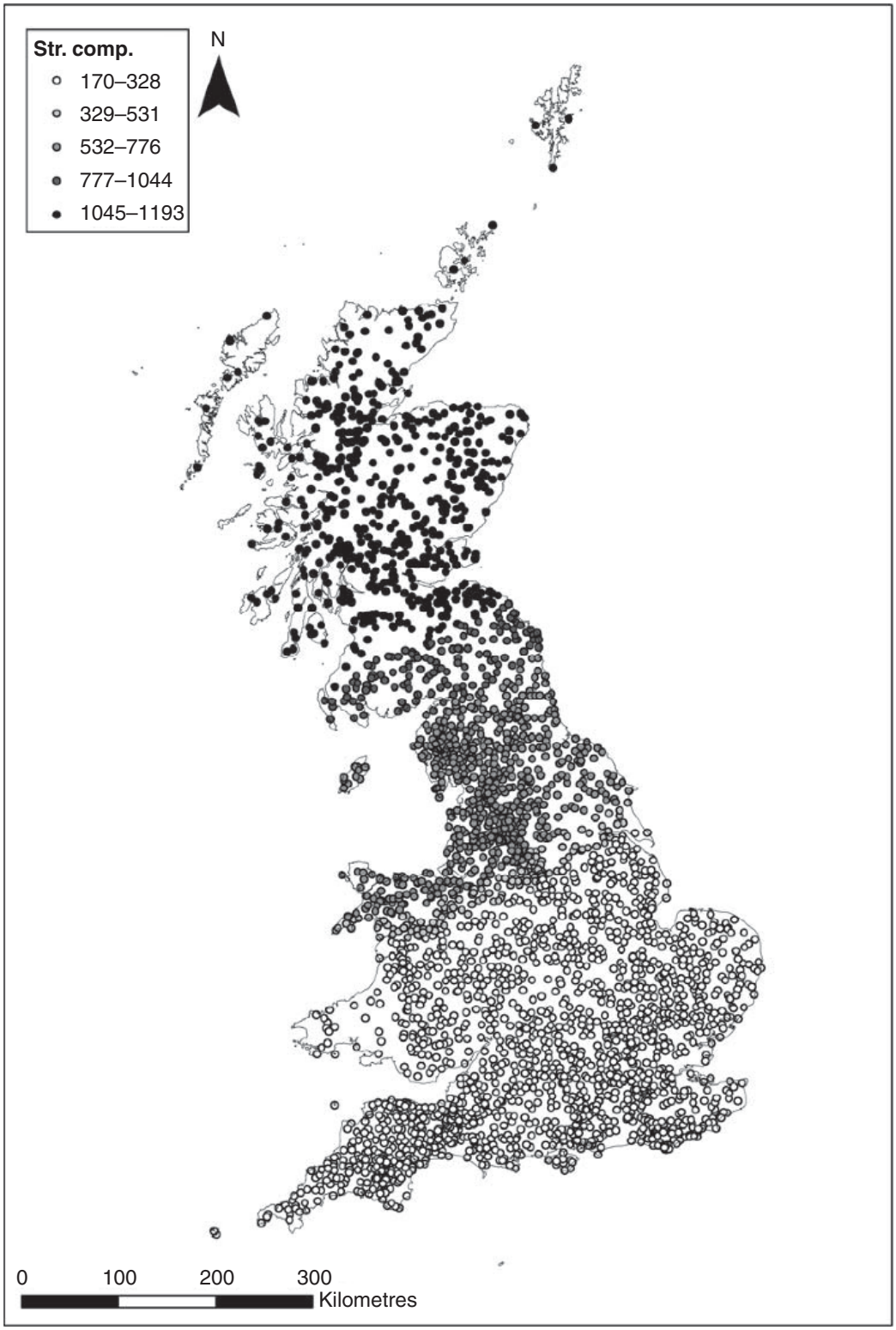


Figure 9.8 Structured component of spherical model for a moving window.

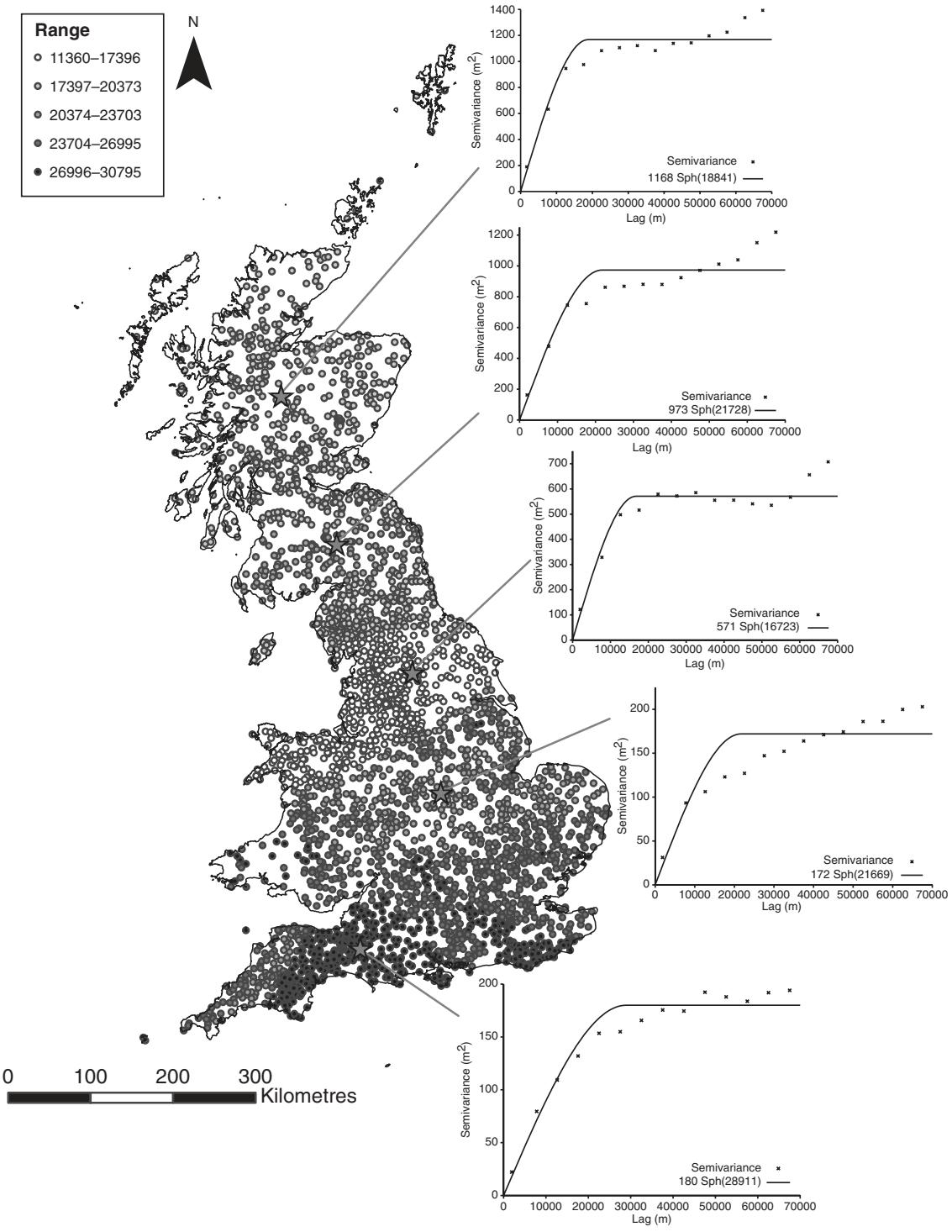


Figure 9.9 Range of spherical model for a moving window, showing five selected semi-variograms with automatically fitted models.

been used in local downscaling cokriging (Pardo-Igúzquiza and Atkinson, 2007).

9.5.3. The objective of non-stationary modelling

Several other chapters of this book have been concerned with geographically weighted regression (GWR). The non-stationary approaches presented in this chapter differ from GWR in their objective. For GWR the objective is to explore the spatially varying parameters of a local regression model; spatial variation in the estimated parameters is the primary interest. For the non-stationary mean and semi-variogram modelling presented in this chapter the objective is spatial prediction or some other geostatistical operation. Thus, non-stationary modelling will be useful where it leads to an increase in the precision of prediction and where it leads to an increase in the precision of the estimation of the prediction variance.

While the objective is spatial prediction, it is often informative to map the non-stationary parameters (in the sense of GWR). For GWR, the coefficients inform on local relations between variables. For geostatistics, the non-stationary mean and (especially) semi-variogram parameters inform on the nature of local spatial structure. For example, the local sill c is very much related to the magnitude of variation locally. The local sill parameter is related mathematically to the local variance (LV), which itself has been used repeatedly as a texture measure in describing remotely sensed images (e.g., Bocher and McCloy, 2006). The local range parameter is related to the scale of spatial variation locally. The local range has also been mapped and used as a texture measure in the classification of remotely sensed images (e.g., Ramstein and Raffy, 1989; Atkinson and Lewis, 2000). Recently, Lloyd *et al*. (2005) have used the local range to show that the choice of optimum spatial resolution for a given scene itself varies locally. In the multivariate case, local variation in the parameters of the linear model of co-regionalization contains more information than the parameters mapped through GWR. The latter omits information on the spatial correlation in each variable, as well as the cross-correlation between variables.

One of the reasons that local modelling is so important for remotely sensed images is that remotely sensed scenes rarely lend themselves to description using the RF model directly. Often, scenes are comprised of several objects arranged on a background (e.g., buildings in a rural area) or comprised of a mosaic of objects (e.g., an agricultural scene). In such circumstances, it is unreasonable to expect the RF model parameterized with a global semi-variogram function to capture the full range of variability in the image locally. Non-stationary variogram modelling achieved by fitting within a moving window goes some way to addressing this problem, but probably not far enough. It would be preferable to define the objects of interest and then fit the RF model locally within the boundaries of those objects. For example, Berberoglu *et al*. (2000) and Lloyd *et al*. (2004) estimated semi-variograms on a per-field basis; semi-variograms were estimated using values within pre-defined boundaries. The semivariances were then used as inputs, along with spectral values, to maximum likelihood and artificial neural network (ANN) classifiers.

9.5.4. When is local, local enough?

The size of neighbourhood within which the local variogram is estimated, whether defined in terms of a search radius or the

nearest number of data points, represents a compromise between two competing factors. The first is the desire to achieve sufficient data points (i.e., sufficiently large neighbourhood) to reduce the uncertainty of variogram estimation to a tolerable level. McBratney and Webster (1986) and Webster and Oliver (1992) provide excellent discussions of the number of data required for reliable estimation of the variogram. The second is the desire to reduce the neighbourhood such as to localize sufficiently the variogram parameters. With regard to the latter point, it should be remembered that since the objective is precise spatial prediction, what is actually required is to represent accurately the local variogram within the window used for local kriging. So an extremely localized variogram may be counter-productive. Ultimately, a balance between these factors should be achieved, potentially through calibration of the window size, although this possibility is often too expensive computationally.

9.6. FUTURE TRENDS IN GEOSTATISTICS

The availability of extensive data sets which cover large areas and have a variety of supports poses problems for conventional geostatistics, as this chapter indicates. Much research is being conducted to develop solutions to the kinds of problems that have arisen. Gotway and Young (2002) review a variety of approaches for area to point interpolation while Kyriakidis (2004) outlines one possible framework in the univariate (kriging) case and Pardo-Igúzquiza and Atkinson (2007) a possible solution in the multivariate (cokriging) case. There are various nonstationary geostatistical models, as discussed at some length in this chapter (see section 9.4), and such approaches overcome the problem of nonstationarity of the mean and variogram which is likely to be encountered if the region of concern is large.

Perhaps the biggest change in focus in the application of geostatistics in the last 20 years has been a shift from prediction (kriging) based analyses to those based on conditional simulation (see section 9.3.3). Simulation allows the generation of many equally-probable realizations and the exploration of spatial uncertainty in the property of interest. In cases where extreme values are of interest kriging is problematic because of its smoothing properties. In such cases, conditional simulation is more appropriate (Goovaerts, 1997).

Another research focus has been on the use and development of model-based geostatistics (Diggle and Ribeiro, 2006). The term was coined by Diggle *et al*. (1998) who introduced a body of approaches that is applicable where Gaussian distributional assumptions, and therefore classical geostatistics, are inappropriate. A Bayesian approach is presented that the authors argue enables uncertainty in the prediction of model parameters to be accounted for properly.

The advances in geostatistical methodology that have been made are limited in their application if extensive expert knowledge is required to apply such models. In the last decade, the range of software packages with extensive geostatistical functionality has grown markedly. Functions for estimating variograms and for kriging and simulation are now commonplace in GIS software. Undoubtedly, with widespread access to often very sophisticated methods misuse and misunderstanding are apparent (Atkinson, 2005). However, an increasingly well educated user base will hopefully contribute to more effective use of spatial data in all application areas.

9.7. SUMMARY

Geostatistics represents a set of tools for the analysis of spatial data. This set is characterized by its shared dependence on the RF model. Central to the RF is the notion of parameter stationarity. For many data sets in mining engineering and petroleum geology the decision to adopt a stationary model is a necessity due to sparcity of data. For many geographical data sets such as are provided by remote sensing (e.g., LiDAR elevation data) it is sensible to relax the constraint of stationarity and estimate the parameters of the RF model locally. This chapter has reviewed geostatistics with a particular focus on non-stationary approaches. Readers are now directed to Chilès and Delfiner (1999) which is widely regarded as a standard reference on the subject.

ACKNOWLEDGEMENTS

The authors thank the British Atmospheric Data Centre (BADC) for providing access to the United Kingdom Meteorological Office (UKMO) Land Surface Observation Stations Data used in the case study.

REFERENCES

Atkinson, P.M. (2001). Geographical information science: GeoComputation and non-stationarity. *Progress in Physical Geography*, **25**: 111–122.

Atkinson, P.M. (2005). Spatial prediction and surface modelling. *Geographical Analysis*, **36**: 113–123.

Atkinson, P.M. and Lewis, P. (2000). Geostatistical classification for remote sensing: an introduction. *Computers and Geosciences*, **26**: 361–371.

Atkinson, P.M. and Tate, N.J. (2000). Spatial scale problems and geostatistical solutions: a review. *Professional Geographer*, **52**: 607–623.

Armstrong, M. (1998). *Basic Linear Geostatistics*. Berlin: Springer.

Berberoglu, S., Lloyd, C.D., Atkinson, P.M. and Curran, P.J. (2000). The integration of spectral and textural information using neural networks for land cover mapping in the Mediterranean. *Computers and Geosciences*, **26**: 385–396.

Bocher, P.K. and McCloy, K.R. (2006). The fundamentals of average local variance – part I: detecting regular patterns. *IEEE Transactions on Image Processing*, **15**: 300–310.

Burrough, P.A. and McDonnell, R.A. (1998). *Principles of Geographical Information Systems*. Oxford: Oxford University Press.

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Uncertainty*. New York: Wiley.

Christakos, G. (1984). On the problem of permissible covariance and semi-variogram models. *Water Resources Research*, **20**: 251–265.

Cressie, N.A.C. (1985). Fitting semi-variogram models by weighted least squares. *Mathematical Geology*, **17**: 563–586.

Curran, P.J. and Atkinson, P.M. (1998). Geostatistics and remote sensing. *Progress in Physical Geography*, **22**: 61–78.

Deutsch, C.V. (2002). *Geostatistical Reservoir Modelling*. New York: Oxford University Press.

Deutsch, C.V. and Journel, A.G. (1998). *GSLIB: Geostatistical Software and User's Guide*, 2nd edn. New York: Oxford University Press.

Diggle, P.J. and Ribeiro, P.J. (2006). *Model-based Geostatistics*. New York: Springer.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**: 299–350.

Dungan, J.L. (1999). Conditional simulation. In: A. Stein, F. van der Meer and B. Gorte (eds), *Spatial Statistics for Remote Sensing*, pp. 135–152. Dordrecht: Kluwer Academic Publishers.

Gething, P.W., Atkinson, P.M., Noor, A.M., Gikandi, P.W., Hay, S.I. and Nixon, M.S. (2007) A local space-time kriging approach applied to a national outpatient malaria dataset. *Computers and Geosciences*, **33**: 1337–1350.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.

Gotway, C.A. and Young, J.J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, **97**: 632–648.

Haas, T.C. (1990). Lognormal and moving window methods of estimating acid deposition. *Journal of the American Statistical Association*, **85**: 950–963.

Herzfeld, U.C. and Holmlund, P. (1990). Geostatistics in glaciology: implications of a study of Scharffenbergbotnen, Dronning Maud Land, East Antarctica. *Annals of Glaciology*, **14**: 107–110.

Hudson, G. and Wackernagel, H. (1994). Mapping temperature using kriging with external drift: theory and an example from Scotland. *International Journal of Climatology*, **14**: 77–91.

Isaaks, E.H. and Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics*. New York: Oxford University Press.

Journel, A.G. (1996). Modelling uncertainty and spatial dependence: stochastic imaging. *International Journal of Geographical Information Systems*, **10**: 517–522.

Journel, A.G. and Huijbregts, C.J. (1978). *Mining Geostatistics*. London: Academic Press.

Kyriakidis, P.C. (2004). A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, **36**: 259–289.

Lloyd, C.D. (2002). Increasing the accuracy of predictions of monthly precipitation in Great Britain using kriging with an external drift. In: Foody, G.M. and Atkinson, P.M. (eds), *Uncertainty in Remote Sensing and GIS*, pp. 243–267. Chichester: John Wiley and Sons.

Lloyd, C.D. (2005). Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain. *Journal of Hydrology*, **308**: 128–150.

Lloyd, C.D. and Atkinson P.M. (2004). Archaeology and geostatistics. *Journal of Archaeological Science*, **31**: 151–165.

Lloyd, C.D., Atkinson, P.M. and Aplin, P. (2005). Characterising local spatial variation in land cover imagery using geostatistical functions and the discrete wavelet transform. In: Renard, P., Demougeot-Renard, H. and Froidevaux, R. (eds), *Geostatistics for Environmental Applications: Proceedings of the Fifth European Conference on Geostatistics for Environmental Applications*. pp. 391–402. Berlin: Springer.

Lloyd, C.D., Berberoglu, S., Curran P.J. and Atkinson P.M. (2004). Per-field mapping of Mediterranean land cover: A comparison of texture measures. *International Journal of Remote Sensing*, **15**: 3943–3965.

McBratney, A.B. and Webster, R. (1986). Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. *Journal of Soil Science*, **37**: 617–639.

Oliver, M.A. and Webster, R. (1990). Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, **4**: 313–332.

Pardo-Igúzquiza, E. (1999). VARFIT: a Fortran-77 program for fitting semi-variogram models by weighted least squares. *Computers and Geosciences*, **25**: 251–261.

Pardo-Igúzquiza, E. and Atkinson, P.M. (2007). Automatic modelling of variograms and cross-variograms in downscaling cokriging by numerical convolution-deconvolution. *Computers and Geosciences*, **33** 1273–1284.

Pebesma, E.J. and Wesseling, C.G. (1998). Gstat, a program for geostatistical modelling, prediction and simulation. *Computers and Geosciences*, **24**: 17–31.

Ramstein, G. and Raffy, M. (1989). Analysis of the structure of radiometric remotely-sensed images. *International Journal of Remote Sensing*, **10**: 1049–1073.

Richmond, A. (2002). Two-point declustering for weighting data pairs in experimental semi-variogram calculations. *Computers and Geosciences*, **28**: 231–241.

Sampson, P.D., Damien, D. and Guttorp, P. (2001). Advances in modelling and inference for environmental processes with non-stationary spatial covariance. In: Monestiez, P., Allard, D. and Froidevaux, R. (eds), *GeoENV III: Geostatistics for Environmental Applications*, pp. 17–32. Dordrecht: Kluwer Academic Publishers.

Schabenberger, O. and Gotway, C.A. (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton: Chapman and Hall/CRC.

Wackernagel, H. (2003). *Multivariate Geostatistics. An Introduction with Applications*, 3rd edn. Berlin: Springer.

Webster, R. and Oliver, M.A. (1990). *Statistical Methods in Soil and Land Resource Survey*. Oxford University Press: Oxford.

Webster, R. and Oliver, M.A. (1992). Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, **43**: 177–192.

Webster, R. and Oliver, M.A. (2000). *Geostatistics for Environmental Scientists*. John Wiley and Sons: Chichester.

Webster, R. and McBratney, A.B. (1989). On the Akaike information criterion for choosing models for semi-variograms of soil properties. *Journal of Soil Science*, **40**: 493–496.

10

Spatial Sampling

Eric Delmelle

10.1. INTRODUCTION

When trying to make inferences about a phenomenon, we are forced to collect a limited number of samples instead of trying to acquire information at every possible location (see, e.g., Cochran, 1963; Dalton *et al*., 1975; Hedayat and Sinha, 1991; and Thompson 2002 for various summaries). A full inventory would yield a clear picture of the variability of the variable of interest, although this process is very time-consuming and expensive. Haining (2003) underlines that the cost of acquiring information on each individual may rule out a complete census. Sparse sampling on the other hand is cheap, but misses important features. However, there are instances where the level of precision may be the major motivation of the sampling process, especially when sampling remains relatively inexpensive. As a rule of thumb, it is generally desirable to have a higher concentration of samples where exhaustive and accurate information is needed, keeping in mind that the number of samples should always be as representative as possible of the entire population (Berry and Baker, 1968).

When surveying a phenomenon characterized by spatial variation, it is necessary to find optimal sample locations in the study area D. This problem is referred to spatial or two-dimensional sampling and has been applied to many disciplines such as mining, soil pollution, environmental monitoring, telecommunications, ecology, geology, and geography, to cite a few. Specific studies on spatial sampling can be found in Ripley (1981), Haining (2003), Cressie (1991), Stehman and Overton (1996) and Muller (1998). Spatial and non-spatial

sampling strategies share common characteristics:

1 the size m of the set of samples;

2 the selection of a sample design, limited by the available budget;

3 an estimator (e.g., the mean) for the population characteristic; and

4 an estimation of the sampling variance to compute confidence intervals.

Following Haining, spatial sampling challenges can be divided into three different categories. The first pertains to problems concerned with estimating some non-spatial characteristics of a spatial population; for example, the average income of households in a state. The second category deals with problems where the spatial variation of a variable needs to be known, in the form of a map, or as a summary measure that highlights scales of variation. The third category includes problems where the objective is to obtain observations that are independent of each other, allowing classical statistical procedures to assist in classifying data.

10.1.1. Spatial structure

A common objective in both spatial and non-spatial approaches is to design a sampling configuration that minimizes the variance associated with the estimation. In this regard, the location of the samples is very critical and depends heavily on the structure of the variable. In non-spatial problems, it may be crucial to stratify the sampling scheme according to important underlying covariates. This holds for spatial phenomena as well. Unfortunately, this variation is often unknown, and an objective is to design an optimal sampling arrangement, to obtain a maximum amount of information. If we undersample in some areas, the spatial variability will not be captured. Oversampling on the other hand can result in redundant data. Consequently, both the location and quantity of the samples is important. This chapter is concerned primarily with the second category of sampling challenges, i.e., capturing the spatial structure of the primary variable.

10.1.2. Structure of the chapter

In this chapter, spatial sampling configurations are reviewed along with their benefits and drawbacks. Second, the influence of geostatistics on sampling schemes is discussed. Sampling schemes can be designed to capture the spatial variation of the variable of interest. Two common objectives therein are the estimation of the covariogram and the minimization of the kriging variance. Third, methods of adaptive sampling and second-phase sampling are presented. Such methods are of a nonlinear nature, and appropriate optimization techniques are necessary to solve such problems. Finally, salient sampling problems such as sampling in the presence of multivariate information, and the use of heuristics are discussed.

10.2. SPATIAL SAMPLING CONFIGURATIONS

This section reviews significant sampling schemes for the purpose of two-dimensional sampling. In the following subsections I will assume that a limited number of samples m is collected within a study area denoted D. The variable of interest Z is sampled on m supports, generating

observations $\{z(\mathbf{s}_i) \mid _i = 1, 2, \ldots m\}$. For ease of illustration, a square study area is used.

10.2.1. Major spatial sampling designs

Random sampling

A *simple random sampling* scheme consists of choosing randomly a set of m sample points in D, where each location in D has an equal probability of being sampled (Ripley, 1981). The selection of a unit does not influence the selection of any other one (King, 1969). Figure 10.1(a) illustrates the random configuration. This type of design is also called *uniform random sampling* since each point is chosen independently uniformly within D. Practically, two random numbers K_i and K_i' are drawn from the interval [0, 1]. Then the point $\mathbf{s}_i$, defined by the pair $\{x_i, y_i\}$ is selected such that:

$$x_i = K_i L, \quad y_i = K_i' L, \tag{10.1}$$

where L denotes the length of the study area D (Aubry, 2000). The process is repeated m-times. According to Griffith and Amrhein (1997), the distribution of the points may not be representative of the underlying geographic surface, because for most samples drawn, some areas will be oversampled while other will be undersampled. The advantages of this design however reside in its operational simplicity, and its capacity to generate a wide variety of distances among pairs of points in D.

Systematic sampling

The population of interest is divided into m intervals of equal size. The first element is randomly or purposively chosen within the first interval, starting at the origin. Depending on the location of the first sample, the remaining $m - 1$ elements are aligned regularly by the size of the interval Δ. If the first sample is chosen at random, the resulting scheme is called *systematic random sampling*. When the first sample point is not chosen at random, the resulting configuration is called *regular systematic sampling*. A *centric systematic sampling* occurs when the first point is chosen in the center of the first interval. The resulting scheme is a checkerboard configuration. The most common regular geometric configurations are the equilateral triangular grid, the rectangular (square) grid, and the hexagonal one (Cressie, 1991). Practically, consider the case where D is divided into a set of small, square cells of size $\Delta = L/\sqrt{m}$. A first point $\mathbf{s}_1 = \{x_1, y_1\}$ is selected within the first cell in the bottom left of D. The coordinates of $\mathbf{s}_1$ are subsequently used to determine the following point $\mathbf{s}_i = \{x_i, y_i\}$ (Aubry, 2000):

$$x_i = x_1 + (i-1)\Delta, \quad y_i = y_1 + (j-1)\Delta$$
$$\forall i,\ j = 1, \ldots, \sqrt{m}. \tag{10.2}$$

To locate sample points along the x- and y-directions, it is imperative to have a desired number of samples m for which $\sqrt{m}$ must be an integer value. The benefits of a systematic approach reside in a good spreading of observations across D, guaranteeing a representative sampling coverage. Additionally, the spreading of the observations prevents sample clustering and redundancy. This design however presents two inconveniences:

1 the distribution of distances between points of D is not sampled adequately because many pairs of points are separated by the same distance; and

2 there is a danger that the spatial process shows evidence of recurring periodicities that will remain uncaptured, because the systematic

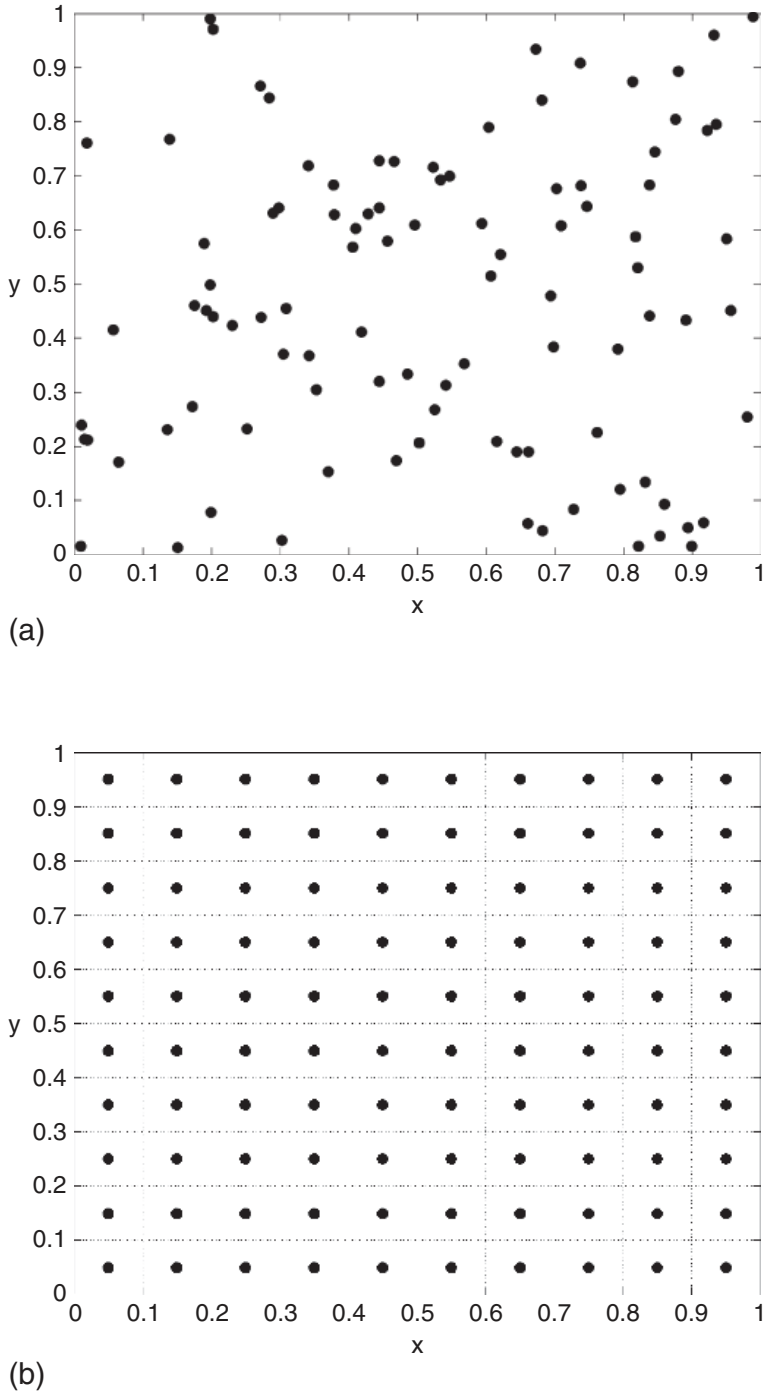
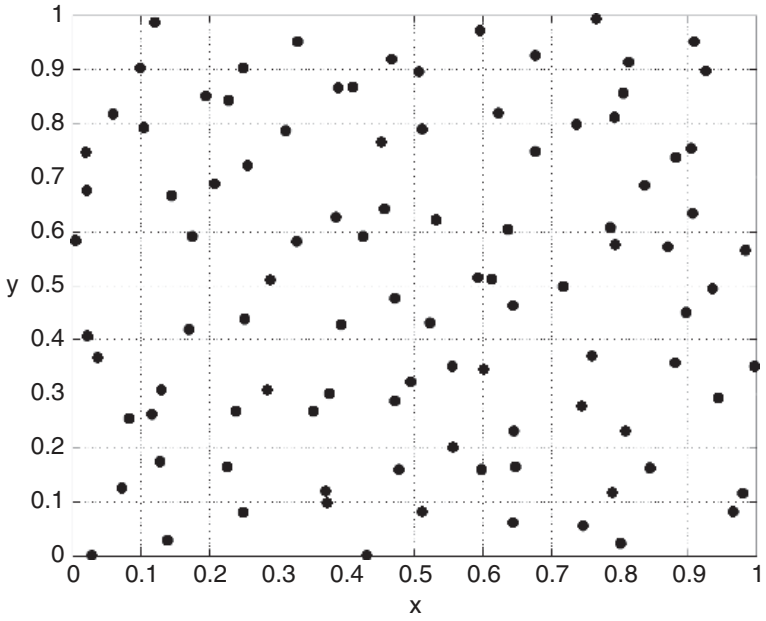
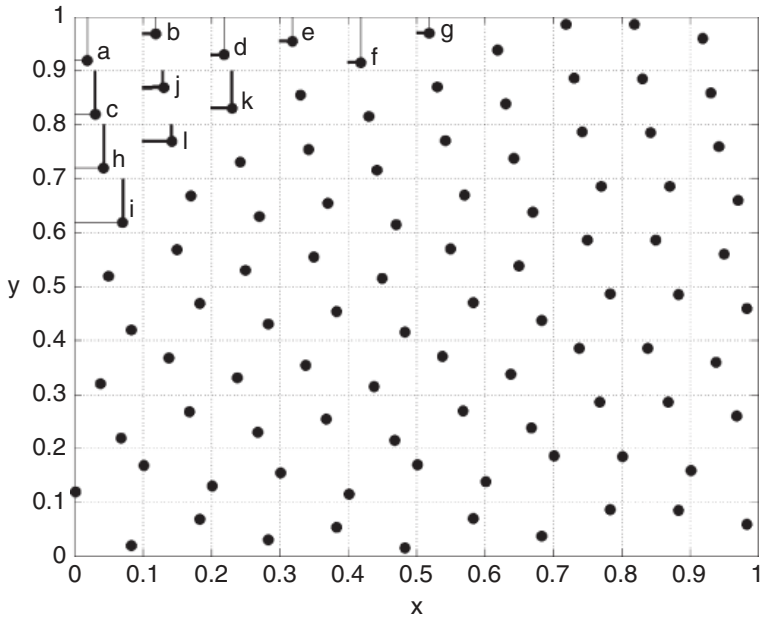


Figure 10.1 From left to right, top to bottom: random, centric systematic, systematic random, and systematic unaligned sampling schemes. Sampling size $m = 100$.



(c)



(d)

design coincides in frequency with a regular pattern in the landscape (Griffith and Amrhein, 1997; Overton and Stehman, 1993).

The second drawback can be lessened considerably by use of a *systematic random* method that combines systematic and random procedures (Dalton *et al.*, 1975). One sample point is randomly selected within each cell. However, sample density needs to be high enough to have some clustering of observations or the spatial relationship

between observations cannot be built. From Figure 10.1(c), some patches of D remain undersampled, while others regions show evidence of clustered observations. A *systematic unaligned* scheme prevents this problem from occurring by imposing a stronger restriction on the random allocation of observations (King, 1969).

Stratified sampling

According to Haining (2003), there are cases when local-area estimates are to be examined, causing stratification to be built into the sampling strategy. In *stratified sampling*, the survey area (or D) is partitioned into non-overlapping strata.[1] For each stratum, a set of samples is collected, where the sum of the samples over all strata must equal m. The knowledge of the underlying process is a determining factor in defining the shape and size of each stratum. Some subregions of D may exhibit stronger spatial variation, ultimately affecting the configuration of each stratum (Cressie, 1991). Smaller strata are preferred in non-homogeneous subregions. When points within each stratum are chosen randomly, the resulting design is named *stratified random sampling*. In Figure 10.2(a), six strata are sampled in proportion to their size. For instance, stratum A represents 30% of D, therefore if $m = 100$, 30 sample points will be allocated within A. Figure 10.2(b) illustrates the allocation of one sample per stratum (*in casu* the centroid), undersampling larger strata.

10.2.2. Efficiency of spatial sampling designs

The sampling efficiency is defined as the inverse of the sampling variance. According to Aubry (2000), the most efficient design leads to the most accurate estimation. Consider the estimation of the global mean z_D:

$$z_D = \frac{1}{[D]} \int_D z(\mathbf{s})\, \mathrm{d}\mathbf{s}. \tag{10.3}$$

It is desirable, from a statistical standpoint to select a configuration that minimizes the prediction error of z_D for a given estimator, for instance the arithmetic mean:

$$\bar{z} = \frac{1}{m} \sum_{i=1}^{m} z(\mathbf{s}_i). \tag{10.4}$$

Efficiency is calculated for all possible realizations of the variable Z by $Var_\xi \left[Z'_D - Z_D \right]$ using σ_k^2, which is the geostatistical prediction error, defined later. In terms of the sampling variance, *stratified random sampling* is at least always equally or more accurate than *random sampling*; its relative efficiency is a monotone increasing function of sample size.

Spatial autocorrelation

Ideally, the density of sample points should increase in locations exhibiting greater spatial variability. Values of closely spaced samples will show strong similarities and it may be redundant to oversample in those areas. The spatial autocorrelation function summarizes the similarity of the values of the variable of interest at different sample locations, as a function of their distance (Gatrell, 1979; Griffith, 1987). Moran's I (Moran, 1948, 1950) is a measure of the degree of spatial autocorrelation among data points:

$$I = \frac{m}{\mathbf{1'W1}} \frac{\sum_{i,j} w(\mathbf{s}_{ij})(z(\mathbf{s}_i) - \bar{z})(z(\mathbf{s}_j) - \bar{z})}{\sum_i (z(\mathbf{s}_i) - \bar{z})^2} \tag{10.5}$$

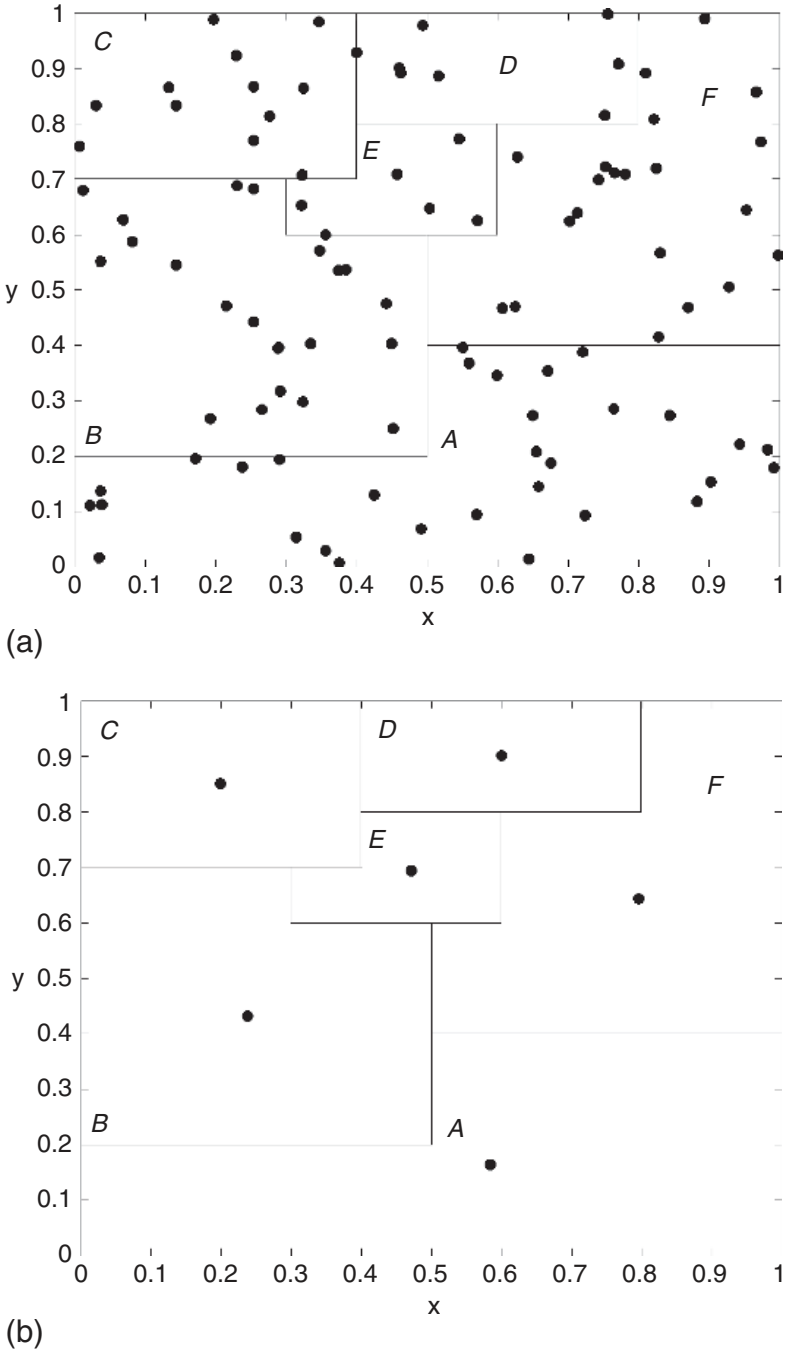


Figure 10.2 Stratified sampling designs with six strata of different sizes ($m = 6$ on the right figure and $m = 100$ to the left).

with $\mathbf{W}$ defined as a weight matrix $w(\mathbf{s}_{ij})$, m is the number of observations, the mean of the sampled values is denoted by $\bar{z}$ and $z(\mathbf{s}_i)$ is the measured attribute value at location $\mathbf{s}_i$. The weight $w(\mathbf{s}_{ij})$ is a measure of spatial proximity between points $\mathbf{s}_i$and $\mathbf{s}_j$; for example:

$$w(\mathbf{s}_{ij}) = \exp(-\beta \mathrm{d}(\mathbf{s}_{ij})^2) \qquad (10.6)$$

where $d(\mathbf{s}_{ij})^2$ is the squared distance between location $\mathbf{s}_i$ and point $\mathbf{s}_j$. Moran's I is not implicitly constrained within the interval $[-1, +1]$. Spatial autocorrelation generally decreases as the distance between sample points increases. A positive autocorrelation occurs when values taken at nearby samples are more alike than samples collected farther away. When the autocorrelation is a linearly decreasing function of distance, *stratified random sampling* has a smaller variance than a *systematic design* (Quenouille, 1949). If the decrease in autocorrelation is not linear, yet concave upwards, systematic sampling is more accurate than stratified random sampling, and a centered systematic design, where each point falls exactly in the middle of each interval, is more efficient than a random systematic sampling configuration (Madow, 1953; Zubrzycki, 1958; Dalenius *et al*., 1960; Bellhouse, 1977; Iachan, 1985).

10.2.3. Other sampling designs

Nested or hierarchical sampling

Nested or hierarchical sampling designs require the study area D to be partitioned randomly into sample units (or blocks) creating the first level in the hierarchy, and this is then further subdivided into sample units nested within level 1, and so forth (Haining, 2003). These units can be systematically or irregularly arranged. As the process progresses, the distances between observations decreases (Corsten and Stein, 1994). One advantage of a nested sampling design is that it allows for multiple scale analysis and supports quadrat analysis. Spatially nested sampling designs may work well for geographic phenomenon that are naturally clustered and for exploring multiple scale effects. Hierarchical sampling is also possible at the discrete level. In such cases, it is desirable to first select randomly one or more counties in a state. Then within these counties we might sample a number of quadrats, or say, townships and finally, within the latter, randomly select some farmsteads (King, 1969).

In the multivariate case, dependent and independent variables are hierarchically organized and are thus not collected at the same sampling frequency (Haining, 2003). The primary variable may exhibit rapid change in spatial structure while the secondary variables are much more homogeneous. A hierarchical sampling design captures such variation by collecting one variable at points nested within larger sampling units so that it can be collected more intensively than another variable.

Clustered sampling

This type of sampling consists of the random selection of groups of sites where sites are spatially close 'within' groups (Cressie, 1991). Clusters of observations are drawn independently with equal probability. In the first stage, when the population is grouped into clusters, the clusters are first sampled (Haining, 2003). Either all of the observations in the clusters, or only a random selection from it, are included. Cluster sampling is essentially useful in the discrete case, when a complete list of the members of a population cannot be obtained, yet a complete list of *groups* (i.e., clusters) of the variable is available. The method is also useful in reducing sampling cost.

10.3. SAMPLING RANDOM FIELDS USING GEOSTATISTICS

Most classical statistical sampling methods make no use of the spatial information provided by nearby samples. Geostatistics describes the spatial continuity that is an essential feature of many natural phenomena.

It can be seen as a collection of statistical methods, describing the spatial autocorrelation among sample data. In geostatistics, multidimensional random fields are formalized and modeled as stochastic processes (see, e.g., Matérn, 1960; Whittle, 1963). In other words, the variable of interest is modeled as a random process that can take a series of outcome values, according to some probability distribution (Goovaerts, 1997). Kriging is an interpolation technique that estimates the value of the primary variable at unsampled locations (usually on a set G of grid points $\{\mathbf{s}_g \mid g = 1, 2, \ldots, G\}$, while minimizing the prediction error. Using data values of Z, an empirical semivariogram $\hat{\gamma}(h)$ summarizing the variance of values separated by a particular distance lag (h) is defined:

$$\hat{\gamma}(h) = \frac{1}{2\mathrm{d}(h)} \sum_{|s_i - s_j| = h} \big(z(\mathbf{s}_i) - z(\mathbf{s}_j)\big)^2 \tag{10.7}$$

where $d(h)$ is the number of pairs of points for a given lag value, and $z(\mathbf{s}_i)$ is the measured attribute value at location $\mathbf{s}_i$. The semivariogram is characterized by a nugget effect a, and a sill σ^2 where $\hat{\gamma}(h)$ levels out. The nugget effect is the spatial dependence at micro scales, caused by measurement errors at distances smaller than the possible sampling distances (Cressie, 1991). Once the lag distance exceeds a value r, called the range, there is no spatial dependence between the sample sites. The variogram function $\hat{\gamma}(h)$ becomes constant at a value called the sill, σ^2. A model $\gamma(h)$ is fitted to the experimental variogram (e.g., an exponential model). With the presence of a nugget effect a:

$$\gamma(h) = a + (\sigma^2 - a)(1 - e^{-3h/r}). \tag{10.8}$$

The corresponding covariogram $C(h)$ that summarizes the covariance between any two points is:

$$C(h) = C(0) - \gamma(h) = \sigma^2 - \gamma(h). \tag{10.9}$$

The interpolated, kriged value at a location $\mathbf{s}$ in D is a weighted mean of surrounding values; each value is weighted according to the covariogram model:

$$\hat{z}(s) = \sum_{i=1}^{I} w_i(\mathbf{s}) z(\mathbf{s}_i) \tag{10.10}$$

where I is the set of neighboring points that are used to estimate the interpolated value at location $\mathbf{s}$, and $w_i(\mathbf{s})$ is the weight associated with each surrounding point. The optimization of spatial sampling in a geostatistical context first requires the estimation of a model to express the spatial dependence at different pairs of distances. This is summarized in the covariogram function. Secondly, such a model is then used for optimal interpolation of the variable under study (Van Groenigen, 1997).

10.3.1. Optimal geometric designs for covariogram estimation

To compute the most representative covariogram and to capture the main features of spatial variability, a good spreading of sample points across the study area is necessary (Van Groenigen *et al.*, 1999). In that context, *systematic sampling* (Figure 10.1(b)) performs well. However, such a sampling design does not guarantee a wide range of separating distances (which is

necessary to estimate the covariogram), because:

1 distances are not evenly distributed; and

2 there are few pairs of points at very small distances to estimate the nugget effect.

A *systematic random* or *systematic unaligned* sample will generate a greater variety of distance pairs. Another solution consists of designing a sampling arrangement where a subset of the m observations are evenly spread across the study area D and the remaining points are somewhat more clustered (Figure 10.3), to capture the covariance at very small distances.

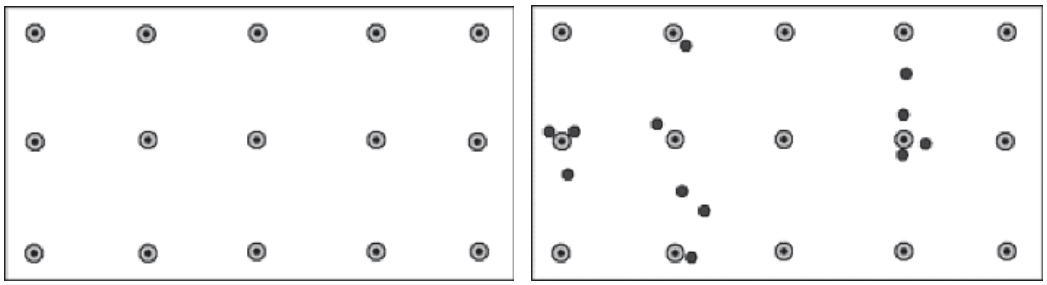
Sample size and sample configuration issues

Optimizing the sampling configuration to estimate the parameters of the covariogram is not an easy task. Webster and Oliver (1993) suggested that a total of at least $m = 150$ samples over the study area is necessary. Moreover, the reliability of the covariogram is partly dependent on the number of pairs of points available within each distance class. In this context, the Warrick/Myers (WM) criterion tries to reproduce an *a priori* defined ideal distribution of pairs of points for estimating the covariogram. The procedure allows one to account for the variation in distance. Following Van Groenigen (1997), the WM-criterion is defined as:

$$J_{w/m}(S) = \mathbf{a}\sum_{i=1}^{K}\mathbf{w}_i(\xi_i^* - \xi_i)^2 + \mathbf{b}\sum_{i=1}^{K}\sigma(m_i) \tag{10.11}$$

$$\sum_{i=1}^{K}\xi_i^* = \frac{m(m-1)}{2} \tag{10.12}$$

where i denotes a given lag class of the covariogram, K represents the total number of classes, and the parameters **a**, **b**, and $\mathbf{w}_i$ are user-defined weights. The term ξ_i^* is a prespecified number of point-pairs for the ith class, ξ_i is the actual number of distances within that class, and $\sigma(m_i)$ is the standard deviation from the median of the distance lag class (Warrick and Myers, 1987). Equation (10.12) expresses the total number of possible distance pairs, given the number of samples. So for instance, when $m = 4$, six pairs of points are generated.



(a) (b)

Figure 10.3 A systematic sampling scheme of $m = 36$ points in D is improved by the introduction of $n = 12$ additional samples (•) clustered among the initial samples.

Presence of anisotropy

Anisotropy (as opposed to isotropy) is a property of a natural process, where the autocorrelation among points changes with distance and direction between two locations. In other words, spatial variability is direction-dependent. Spatial variables may exhibit linear continuity, such as in estimating riparian habitat along rivers, aeolian deposits, and soil permeability along prevailing wind directions. We talk about an isotropic process however when there is no effect of direction in the spatial autocorrelation of the primary variable. It is generally desirable to augment the sampling frequency in the angle of minimum continuity, since the spatial gradient of variation is maximum in that direction.

Impact of the nugget effect

Bogaert and Russo (1999) made an attempt to understand how the covariogram parameters are influenced by the choice of particular sampling locations. Their objective was to limit the variability of the covariogram estimator. When the covariogram has no nugget effect, the benefits of the optimization procedure are somewhat diminished. In the presence of a nugget effect, a *random sampling* configuration will score poorly, because of the limited information offered by random sampling for small distances.

Using nested designs

A nested design allows good estimation of the nugget effect at the origin. However, *nested sampling* configurations produce inaccurate estimation of the covariogram in comparison to *random* and *systematic sampling*. This occurs due to the rather limited area covered by the sampling scheme, yielding a high observation density in subregions of the area, and a low observation density for other parts of the area. This in turn generates only a few distances for which covariogram values are available. Nested sampling designs are especially unsuitable when the observations collected according to such a design are used subsequently to estimate values at unvisited locations (Corsten and Stein, 1994).

10.3.2. Optimal designs to minimize the kriging variance

Kriging provides not only a least-squares estimate of the attribute but also an error variance (Isaaks and Srivastava, 1989), quantifying the prediction uncertainty at a particular location in space. This uncertainty is minimal, or zero when there is no nugget effect, at existing sampling points and increases with the distance to the nearest samples. A major objective consists of designing a sampling configuration to minimize this uncertainty over the study area. This can be achieved when the covariogram, representing the spatial structure of the variable, is known *a priori* or has been estimated. In this regard, optimal sampling strategies have been suggested to reduce the prediction error associated with the interpolation process (Pettitt and McBratney, 1993; Van Groenigen *et al.*, 1999). Equation (10.13) formulates the kriging variance at a location $\mathbf{s}$, where $\mathbf{C}_M^{-1}$ is the inverse of the covariance matrix $\mathbf{C}_M$ based on the covariogram function (Bailey and Gatrell, 1995). M denotes the set of initial samples and has cardinality m. The term $\mathbf{c}$ is a column vector and $\mathbf{c}^T$ the corresponding row vector, as given in Equation (10.15):

$$\sigma_k^2(\mathbf{s}) = \sigma^2 - \mathbf{c}^T(\mathbf{s}) \cdot \mathbf{C}_M^{-1} \cdot \mathbf{c}(\mathbf{s}) \qquad (10.13)$$

$$C_M = \begin{bmatrix} \sigma^2 & C_{1,2} & \dots & C_{1,m} \\ C_{2,1} & \sigma^2 & \cdots & C_{2,m} \\ \vdots & \vdots & \ddots & \\ C_{m,1} & C_{m,2} & \cdots & \sigma^2 \end{bmatrix} \tag{10.14}$$

$$\mathbf{c} = \begin{bmatrix} \sigma^2 \\ C_{2,1} \\ \vdots \\ C_{m,1} \end{bmatrix}, \quad \mathbf{c}^T = \begin{bmatrix} \sigma^2 & C_{1,2} & \dots & C_{1,m} \end{bmatrix}. \tag{10.15}$$

The total kriging variance *TKV* is obtained by integrating Equation (10.13) over *D*:

$$TKV = \int_D \sigma_k^2(\mathbf{s})\mathrm{d}\mathbf{s}. \tag{10.16}$$

Computationally, it is easier to discretize *D* and sum the kriging variance over all grid points $\mathbf{s}_g$. The average kriging variance *AKV* over the study area is defined as:

$$AKV = \sum_{g \in G} \sigma_k^2(\mathbf{s}_g). \tag{10.17}$$

The only requirement to calculate the kriging variance is to have an initial covariogram and the locations of the *m* initial sample points. It then depends solely on the spatial dependence and configuration of the observations (Cressie, 1991).

Illustration

Since continuous sampling is not feasible, it is necessary to discretize the area into a set of potential points. Seeking the best sampling procedure becomes a combinatorial problem. Figure 10.4 illustrates the kriging variance associated with random sampling and systematic random sampling from an exponential model. Darker areas denote a higher interpolation uncertainty, which is increasing away from existing points. The estimation error is low at visited points.

Distance-based criteria

It is possible to design sampling configurations considering explicitly the spatial correlation of the variable (Arbia, 1994). What would you do if you were in a dark room with candles? You would probably light the first candle at a random location or in the middle of the room. Then you would find it convenient to light the second candle somewhere further away from the first. How far away will depend on the luminosity of the first candle. The stronger the light, the further it can be located from the first candle. You would then light the third candle far away from the two first ones. Such an approach – known as ***D****epending Areal* ***U****nits* ***S****equential* ***T****echnique* (DUST) – is an infill sampling algorithm, and very suitable to locate points to minimize the kriging variance over *D*. Another method, known as the ***M****inimization of the* ***M****ean of the* ***S****hortest* ***D****istances* (MMSD) requires all sampling points spread evenly over the study area, ensuring that unvisited locations are never far from a sampling point. Both MMSD and DUST methods assume:

1 prior knowledge of the spatial structure of the variable; and

2 a stationary variable – an assumption violated in practice.

Both criteria are purely deterministic, resulting in spreading pairs of points evenly across the study area, similar to the systematic configuration. Van Groenigen (1997) notes that the area *D* is a continuous, infinite plane.

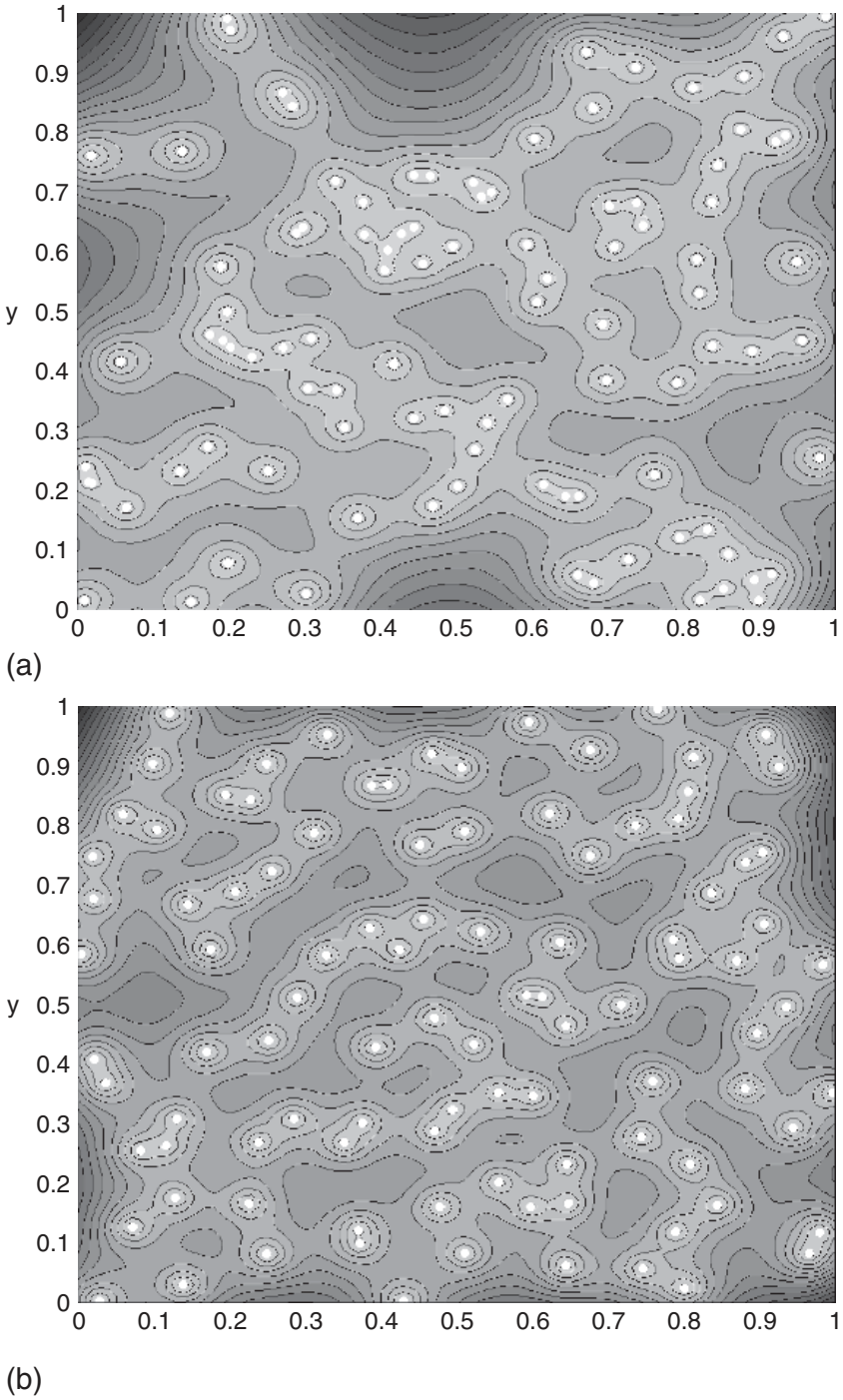


Figure 10.4 The kriging variance of a systematic random pattern (right figure) reduces the value of Equation 17 by 20% from a random pattern. Sample patterns are similar to those in Figure 10.1.

In reality, it is not physically possible to sample everywhere; the presence of spatial barriers such as roads, buildings or mountains restricts the sampling process and limits the number and location of potential points.

Impact of the nugget effect

What is the influence of the nugget effect and sampling densities on the final sampling configuration? As the ratio *nugget*/*sill* increases, a different sampling configuration is reached, placing more observations near the boundaries of the study area, because of the high variance at short distances. In that case, more samples are needed to obtain the same level of objective function (equation (10.17)) over D (Burgess *et al*., 1981). When the nugget effect is maximum ($\approx$ sill), the covariogram is pure noise, and the resulting optimal sampling scheme is purely random, because no spatial correlation is present. At maximum sampling density, the estimation variance can never be less than the nugget effect. When the variance among pairs of points at very small distances ($\approx$ nugget effect) is very high, a hexagonal design will perform best.

Presence of anisotropy

Which type of sampling design performs better in reducing the maximum kriging variance, when anisotropy is present? When the process is isotropic, a systematic equilateral triangle design will keep the variance to a minimum, because it reduces the farthest distance from initial sample points to points that are not visited. A square grid performs well, especially in the case of isotropy (McBratney and Webster, 1981; McBratney *et al*., 1981). When anisotropy is present on the other hand, a square grid pattern is preferred to a hexagonal arrangement, although the improvement is marginal (Olea, 1984).

Choice of a covariogram fitting model

Does the choice of a covariogram fitting model affect the value of equation (10.17)? According to Van Groenigen (2000), an exponential model generates a point-symmetric sampling configuration that is identical to a linear model. However, the use of a Gaussian model tends to locate sample points very close to the boundary of D. This is explained by the large kriging weights assigned to small distance values (parabolic behavior at the origin).

10.3.3. Sampling reduction

Sampling density reduction of an existing spatial network is a problem related to sampling designs and is relevant in many regions of the world where funding for environmental monitoring is decreasing. The process entails lowering the number of samples to reach an effective level of accuracy. Technically, it consists of selecting existing samples from the original data set that will, in combination with a spatial interpolation algorithm, produce the best possible estimate of the variable of interest, in comparison with the results obtained if all sample points were used (Olea, 1984). Usually, it is assumed that the residuals come from a stationary process, and that the covariogram is linearly decreasing, with no nugget effect, and that the process is isotropic. In a study aimed at predicting soil water content, Ferreyra *et al*., (2002) developed a similar sampling density reduction method, from 57 observations to 10 observations. With an optimal arrangement of 10 samples, over 70% of the predicted water content had an error within $\pm 10\%$, showing that a similar

level of confidence is reached with a limited number of samples.

10.4. SECOND-PHASE AND ADAPTIVE SAMPLING

When there is a need or desire to gather more information (i.e., additional samples) about the variable of interest, we talk about adaptive and second-phase sampling, depending on the study objective. In the following subsections, both techniques are discussed.

10.4.1. Adaptive sampling

Adaptive sampling finds its roots in the concept of *progressive sampling* (Makarovic, 1973). It provides an objective and automatic method for sampling, for example, terrain of varying complexity when sampling altitude variation. As illustrated in Figure 10.5, progressive sampling involves a series of successive runs, beginning with a coarse sampling grid and then proceeding to grids of higher densities. The grid density is doubled on each successive sampling run and the points to be sampled are determined by a computer analysis of the data obtained on the preceding run. The analysis proceeds as follows: a square patch of nine points on the coarsest grid is selected and the height differences between each adjacent pair of points along the rows and columns are computed. The second differences are then calculated. The latter carries information on the terrain curvature. If the estimated curvature exceeds a certain threshold, it becomes necessary on the next run to increase the sampling density and sample points at the next level of grid density.

A similar study was carried out by Ayeni (1982) to determine the optimum number and spacing of terrain elevation data points to produce a Digital Elevation Model (DEM). The importance of evaluating the adequate number of data points as well as the appropriate sampling distribution of such points, that in turn constitute a good match to characterize a given terrain. Determining a sufficient number of points is not straightforward, since it depends on terrain roughness in relation to the size of the area occupied by the terrain. The ideas suggested in progressive sampling were later carried over to the field of *adaptive sampling* (see Thompson and Seber, 1996). A major difference with conventional designs lies in the selection of additional samples in adaptive designs, because the location of a new sample will depend upon the value of the points observed in the field. In other words, the procedure for selecting additional samples depends on the outcome of the variable of interest, as observed during the survey of an initial sampling phase. The addition of a new sample improves confidence in the sampling distribution. Adaptive sampling is very efficient in the context of soil contamination (Cox, 1999). How should a risk manager decide where to re-sample in order to maximize information on contamination? In this particular context it is generally recommended to sample in locations above a particular threshold and draw a fixed number of additional samples around them until subsequent measurement values are below a pre-specified contamination threshold. Figure 10.6 illustrates the procedure for adaptive cluster sampling, where sample points represent measurement locations of hypothetical contamination rates. On the left, contamination rates have been measured at seven locations. A geographic location is said to be at risk (and needs remediation) when its value is above 0.7 or at 70% of the contamination threshold. Call a property fathomed if samples have been taken from its immediate neighbors. A common choice is to define new neighbors

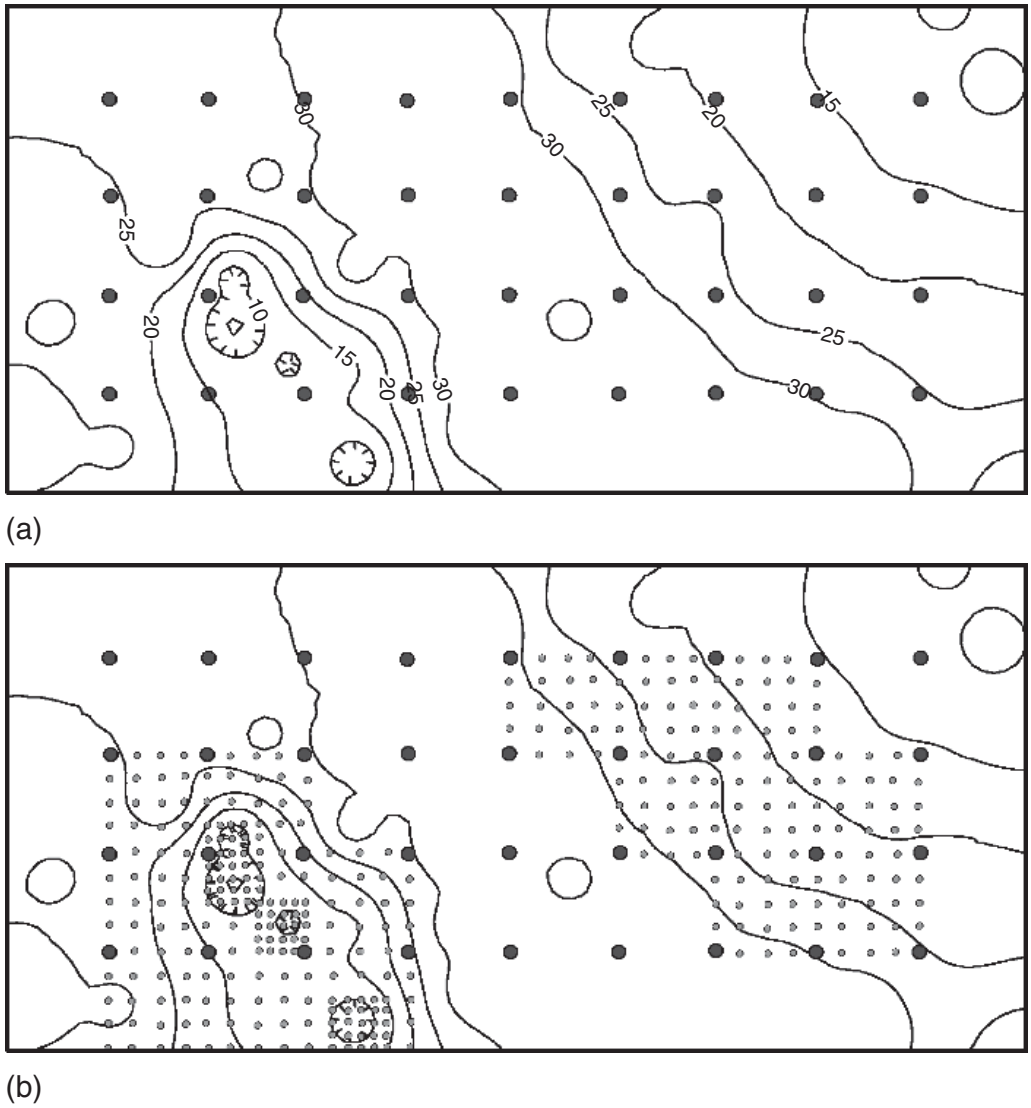
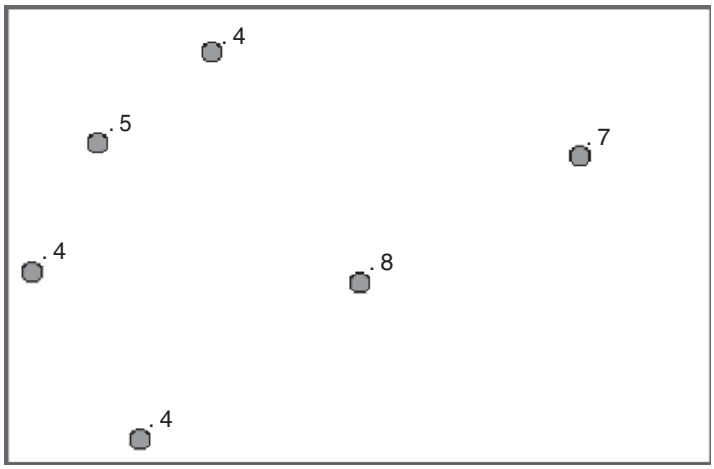


Figure 10.5 Initial systematic sampling of altitude is performed over the study in the top figure. When strong variation in elevation is encountered, the sampling density is increased until desirable threshold is met.

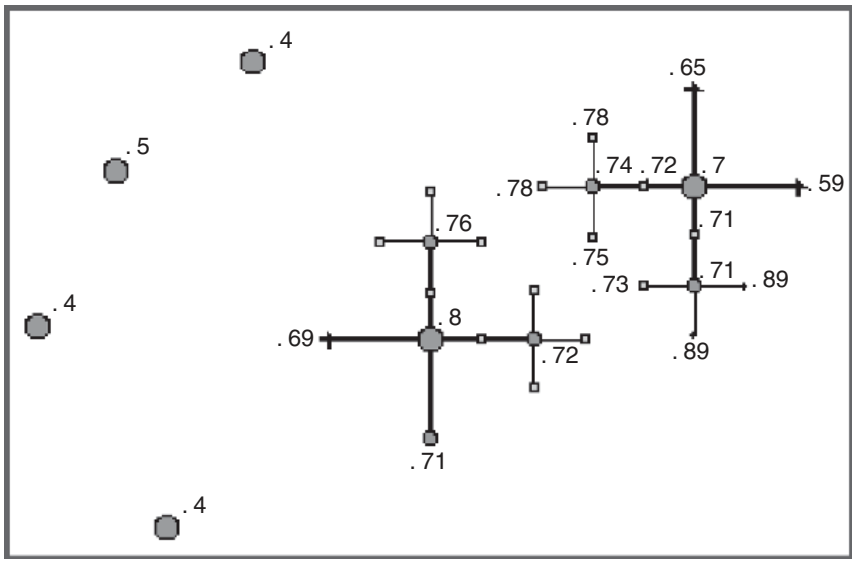
of a contaminated zone to the North, South, East, and West: fathom each property on the list by sampling and remove it from the risk list when it has been fathomed. In other words, the procedure re-samples four neighboring locations of a contaminated site. Once a site shows a contamination rate under the threshold value, it is fathomed. Otherwise, the procedure continues until a trigger condition is satisfied (e.g., a maximum number of additional samples is reached). This approach has some limitations however, because there is little rationale in taking additional samples in areas where we know that the probability of exceeding a particular threshold is maximal.

10.4.2. Second-phase sampling

In second-phase spatial sampling, a set M of m initial measurements has been collected, and a covariogram $C(h)$ has been calculated. In the second-phase, the scientist



(a)



(b)

Figure 10.6 The cluster adaptive sampling procedure, illustrated in the context of toxic waste remediation. A site is fathomed (+) when its toxicity rate does not exceed the contamination value.

augments the set of observations, guided by the covariogram. The objective function aims to collect new samples to reduce the kriging variance or uncertainty by as much as possible. Equation (10.18) formulates the change in kriging variance $\Delta\sigma_k^2$ over all grid points $\mathbf{s}_g$, when a set N of size n containing new sample points is added to our initial sample set M. The change $\Delta\sigma_k^2$ is the difference between the kriging variance calculated with initial sample points and the kriging variance of the augmented set $M \cup N$ containing $[m + n]$ samples:

$$\Delta\sigma_k^2 = \left[TKV^{\text{old}} - TKV^{\text{new}}\right] = \frac{1}{G}\left[\sum_{g\in G}\sigma_{k,\text{old}^2}(\mathbf{s}_g) - \sum_{g\in G}\sigma_{k,\text{new}}^2(\mathbf{s}_g)\right] \tag{10.18}$$

$$\sigma^2_{k,\text{old}}(\mathbf{s}_g) = \sigma^2 - \underbrace{c(\mathbf{s}_g)}_{[1,m]} \cdot \underbrace{\mathbf{C}^{-1}}_{[m]} \cdot \underbrace{\mathbf{c}^T(\mathbf{s}_g)}_{[m,1]} \quad (10.19)$$

$$\sigma^2_{k,\text{new}}(\mathbf{s}_g) = \sigma^2 - \underbrace{c(\mathbf{s}_g)}_{[1,m+n]} \cdot \underbrace{\mathbf{C}^{-1}}_{[m+n]} \cdot \underbrace{\mathbf{c}^T(\mathbf{s}_g)}_{[m+n,1]}. \quad (10.20)$$

The objective function (equation (10.21)) is to find the optimal set S^* containing $m + n$ points that will maximize this change in kriging variance (Christakos and Olea, 1992; Van Groenigen *et al.*, 1999), where S is a specific sampling scheme:

$$\underbrace{MAX}_{\{\mathbf{s}_{m+1},\ \ldots,\mathbf{s}_{m+n}\}} \quad J(S) = \frac{1}{G}\sum_{g\in G} \Delta\sigma^2_k(\mathbf{s}_g; S). \quad (10.21)$$

For simplicity, the continuous region D is usually approximated by a finite set P of p points (Cressie, 1991). The set of new points is selected from the set of potential points P. Hence, there is a total of $\binom{p}{n}$ possible sampling combinations and it is too time-consuming to find the optimal set using combinatorics. Figure 10.7 illustrates the case where 50 sample points have been collected in the first stage, leading to an exponential covariogram, with the sequential addition of $n = 10$ new points and an improvement in the objective function of nearly 20%.

Weighting the kriging variance?

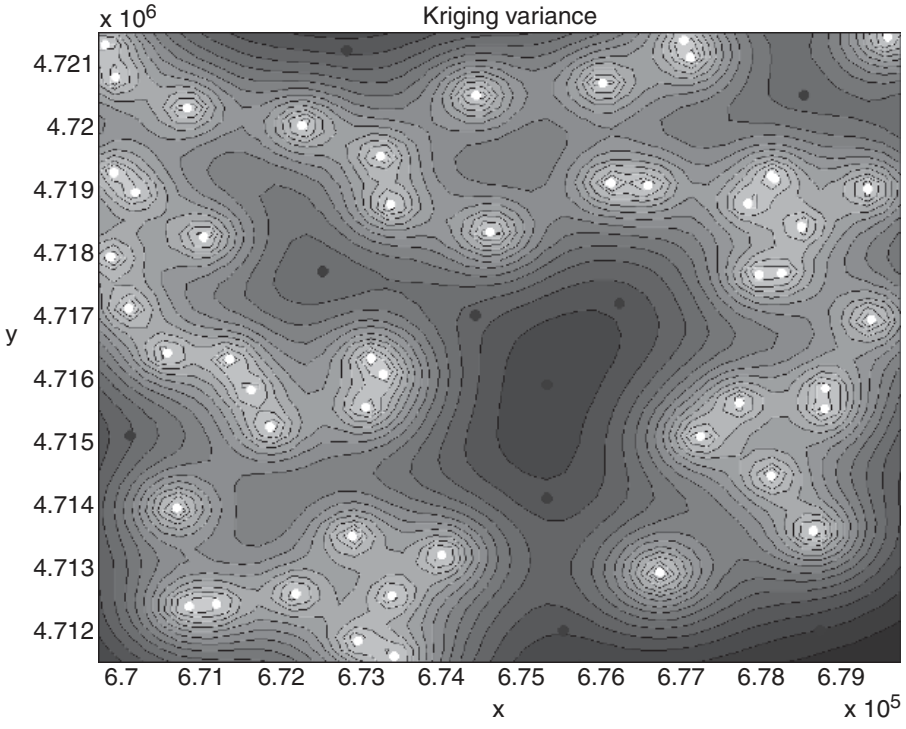
The use of a weighting function $w(\bullet)$ for the kriging variance was originally suggested by Cressie (1991) and has been applied by Van Groenigen *et al.*, (2000), Rogerson *et al.*, (2004), and Delmelle (2005). The importance of a location to be sampled is represented by a weight $w(\mathbf{s})$. The objective is to find the optimal sampling scheme S^* containing $m+n$ points that will maximize the change in weighted kriging variance. From equation (10.21):

$$\underbrace{MAX}_{\{\mathbf{s}_{m+1},\ldots,\mathbf{s}_{m+n}\}} \quad J(S) = \frac{1}{G}\sum_{g\in G} w(\mathbf{s}_g)\Delta\sigma^2_k(\mathbf{s}_g; S). \quad (10.22)$$

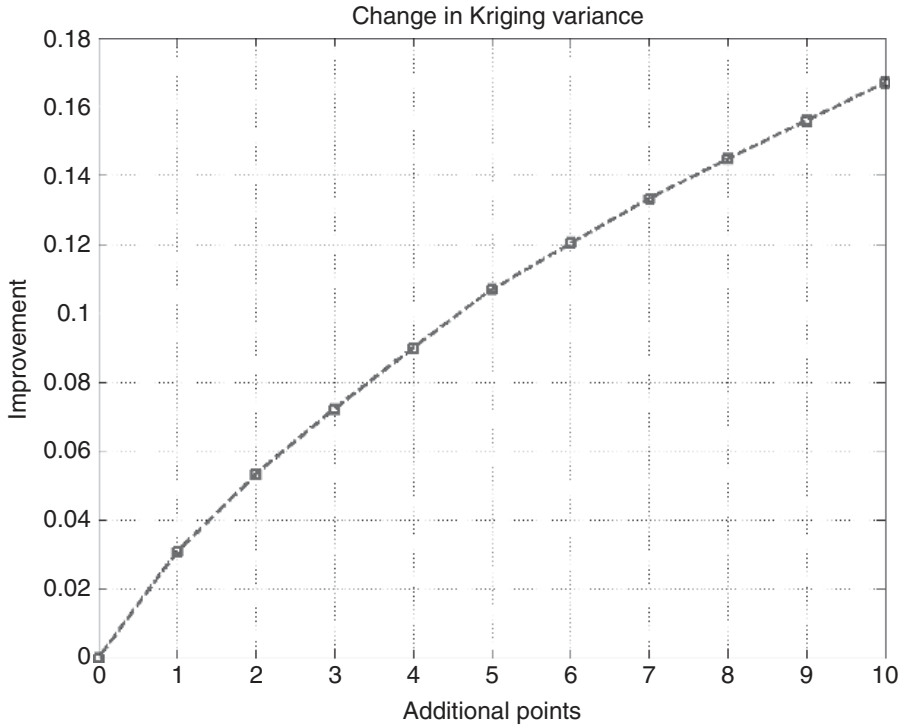
In an effort to detect contaminated zones in the Rotterdam harbor, Van Groenigen *et al.*, (2000) introduced the ***W**eighted **M**eans of **S**hortest **D**istance* (WMSD) criterion, offering a flexible way of using prior knowledge on the variable under study. However, the weights do not reflect the spatial structure of the variable, but rather the scientist's perception of the risks of contamination. In the first sampling phase, sampling weights are assigned to sub-areas based on their risks for contamination. In the second phase however, a greater weight is assigned to locations expected to exhibit a higher priority for remediation. Four weighting factors are considered with weights $w = 1$, 1.5, 2, and 3, leading to more intensive sampling where the weight is higher. In a more recent study, Rogerson *et al.*, (2004) have developed a second-phase sampling technique, allowing re-sampling in areas where there is some uncertainty associated with a variable of interest, and hence not in areas where the probability of an event occurring is near 0 or 1. A greedy algorithm was proposed to locate the points that would maximize the change in weighted kriging variance.

Shortcomings of the use of the kriging variance

Many authors have advocated the use of the kriging variance as a measure of uncertainty. It is unfortunately misused as a measure of reliability of the kriging estimate, as noted by several authors (Deutsch and



(a)



(b)

Figure 10.7 An initial sampling network of $m = 50$ points (in white) has been augmented with the addition of $n = 10$ new samples (in blue). The figure to the right displays the improvement.

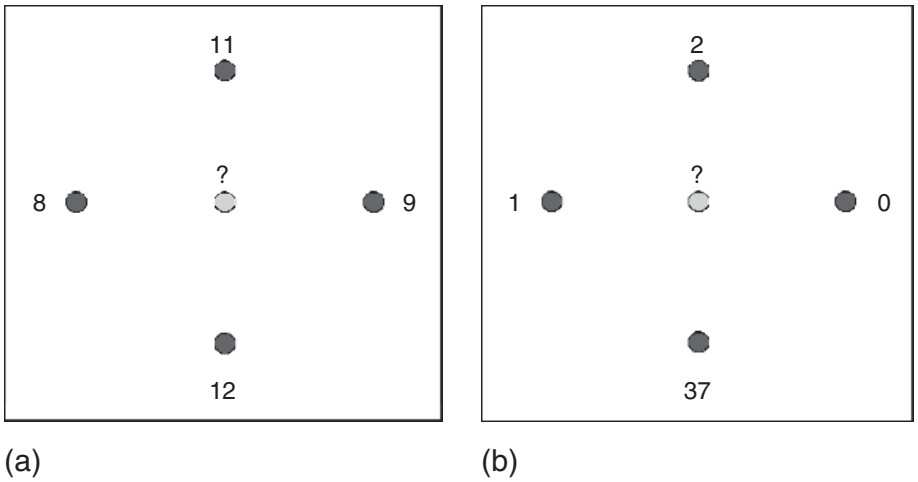


Figure 10.8 Example of two-dimensional non-stationarity. Dark points are used as data values to interpolate the center point (light gray). After Armstrong (1994).

Journel, 1997; Armstrong, 1994). It is solely a function of the sample pattern, sample density, the numbers of samples and their covariance structure. The kriging variance assumes that the errors are independent of each other. This means that the process is stationary, an assumption usually violated in practice. Stationarity entails that the variation of the primary variable between two points remains similar at different locations in space, as long their separation distance remains unchanged. Figure 10.8 illustrates non-stationarity in two dimensions (Armstrong, 1994). The objective in this particular example is to interpolate the value of the inner grid point, highlighted with a question mark. The interpolation depends on the values of the four surrounding points. Two scenarios are presented. The scenario in *b* shows three very similar values and an extreme one. The scenario in *a* however shows four values in a very narrow range. Assuming the spatial structure is similar in both cases, and since the configuration of the data points is the same, the kriging variances are identical. However, we have more confidence in the scenario on the left since there is less variation among the neighbors. This illustrates that the prediction error is not suitable for setting up confidence intervals and should not be used as an optimization criterion for additional sampling strategies.

10.5. CURRENT RESEARCH DIRECTIONS

10.5.1. Incorporating multivariate information

Sample data can be very difficult to collect, and very expensive, especially in monitoring air or soil pollution for instance (Haining, 2003). Secondary data can be a valuable asset if they are available over an entire study area and combined within the primary variable (Hengl *et al.*, 2003). Secondary spatial data sources include maps, socioeconomic, and demographic census data, but also data generated by public sources (local and regional). This is very valuable and there has been a dramatic

growth in the availability of secondary data associated with DEMs and satellites (for environmental data). Such secondary data is easily integrated within a GIS framework (Haining, 2003). In multi-phase sampling, for instance, research has been confined to the use of covariates in determining the locations of initial measurements, whereby sample concentration is increased where covariates exhibit substantial spatial variation (Makarovic, 1973). Ideally, secondary variables should be used to reduce the sampling effort in areas where their local contribution in predicting the primary variable is maximum (Delmelle, 2005). If a set of covariates predicts accurately the data value where no initial sample has been collected yet, there is little incentive to perform sampling at that location. On the other hand, when covariates perform poorly in estimating the primary variable, additional samples may be necessary. The general issue pertains to quantifying the spatial contribution given by covariates.

10.5.2. Weighting the kriging variance appropriately

Some current research has looked at ways to weight the kriging variance. Intuitively, one would like to sample at unvisited locations, far from existing ones. This is accomplished using the kriging variance as a sampling criterion. However, the spatial variability of the primary variable is not accounted for. It is recommended to weight the kriging variance where the gradient of the primary variable is maximum, because there is a rapid change at that location in the variable (Delmelle, 2005). It is also desirable to reduce sampling effort by using information provided by auxiliary variables, when available.

10.5.3. The use of heuristics in sampling optimization

In second-phase sampling, the set N of additional samples will be chosen from a set P of candidate sampling locations. This set is relatively large in practice, and hence the number of possible solutions forbids an exhaustive search for the optimum (Michalewicz and Fogel, 2000). A total enumeration of all potential solutions is not possible, because of the combinatorial explosion. (Goldberg, 1989; Grötschel and Lovàsz, 1995). The search for an approximate solution for complex problems is conducted using a suitable heuristic method H. The use of a heuristic is necessary to assist in the identification of an optimal sample set S^* (or near optimal set $S^+) \subset P$. The heuristic controls a process that intends to solve this optimization problem. The set S^* is optimal for the objective function J defined in equation (10.22). The efficiency of a heuristic depends on its capacity to give as often as possible a solution S^+ close to S^* (Grötschel and Lovàsz, 1995). In second-phase sampling, there are two different ways of supplementing an initial set. Either n points are selected at one time and added to the initial set or one point at a time is added n-times to the initial set. The former is defined as simultaneous addition and the latter is known as sequential addition and is suboptimal. Note that a hybrid approach that would combine both techniques is possible as well. In spatial sampling, limited research has been devoted to comparing the benefits and drawbacks of these heuristics. The greedy (or myopic) algorithm has been used by Aspie and Barnes (1990), Christakos and Olea (1992) and Rogerson *et al*., (2004). Simulated annealing has been applied to spatial sampling problems in Ferri and Piccioni (1992), Van Groenigen and Stein (1998), and Pardo-Igúzquiza (1998).

10.5.4. Spatio-temporal sampling issues

Spatial sampling optimization as discussed in this chapter is based on the assumption of stationarity of the variable itself over time ($\approx$ no temporal variation). Variables such as rainfall, temperature, and snowfall vary over time and it is not possible to take a second set of samples to improve the prediction of these variables without affecting the stability of the model. Work in this context has been carried out by Lajaunie *et al*., (1999).

NOTE

1 Note that a *systematic sampling scheme* is a special case of a *stratified design* in that the strata are all squares of equal size.

REFERENCES

Arbia, G. (1994). Selection techniques in sampling spatial units. *Quaderni di Statistica e Matematica Applicata Alle Scienze Economico-Sociali*, **16**: 81–91.

Armstrong, M. (1994). Is research in mining geostats as dead as a dodo? In: Dimitrakopoulos R. (ed.). *Geostatistics for the Next Century*, pp. 303–312. Dordrecht: Kluwer Academic Publisher.

Aubry, P. (2000). Le Traitement des Variables Régionalisées en Ecologie: Apports de la Géomatique et de la Géostatistique}. Thèse de doctorat. Université Claude Bernard – Lyon 1.

Aspie, D. and Barnes, R.J. (1990). Infill-sampling design and the cost of classification errors. *Mathematical Geology*, **22**: 915–932.

Ayeni, O. (1982). Optimum sampling for digital terrain models: A trend towards automation. *Photogrammetric Engineering and Remote Sensing*, **48**: 1687–1694.

Bailey, T.C. and Gatrell, A.C. (1995). *Interactive Spatial Data Analysis*. Longman. 413p.

Bellhouse, D.R. (1977). Optimal designs for sampling in two dimensions. *Biometrika*, **64**: 605–611.

Berry, B.J.L. and Baker, A.M. (1968). Geographic sampling. In: Berry B.J.L. and Marble D.F. (eds), *Spatial Analysis: a Reader in Statistical Geography*; pp. 91–100. Englewood Cliffs, N.J.: Prentice-Hall.

Bogaert, P. and Russo, D. (1999). Optimal sampling design for the estimation of the variogram based on a least squares approach. *Water Resources Research*, **35**(4): 1275–1289.

Burgess, T.M., Webster, R. and McBratney, A.B. (1981). Optimal interpolation and isarithmic mapping of soil properties: IV. Sampling strategy. *Journal of Soil Science*, **32**: 643–659.

Christakos, G. and Olea, R.A. (1992). Sampling design for spatially distributed hydrogeologic and environmental processes. *Advanced Water Resources*, **15**: 219–237.

Cochran, W.G. (1963). *Sampling Techniques*. Second Edition. New York: Wiley. 413p.

Corsten, L.C.A. and Stein, A. (1994). Nested sampling for estimating spatial semivariograms compared to other designs. *Applied Stochastic Models and Data Analysis*, **10**: 103–122.

Cox, L.A. (1999). Adaptive spatial sampling of contaminated soil. *Risk Analysis*, **19**: 1059–1069.

Cressie, N. (1991). *Statistics for Spatial Data*. New York: Wiley. 900p.

Dalenius, T., Hajek, J. and Zubrzycki, S. (1960). On plane sampling and related geometrical problems. *Proceedings of the Fourth Berkeley Symposium*, **1**: 125–150.

Dalton, R., Garlick, J., Minshull, R. and Robinson, A. (1975). *Sampling Techniques in Geography*. London: Georges Philip and Son Limited. 95p.

Delmelle, E.M. (2005). *Optimization of Second-Phase Spatial Sampling Using Auxiliary Information*. Ph.D. Dissertation, Department of Geography, SUNY at Buffalo.

Deutsch, C.V. and Journel, A.G. (1997) *Gslib: Geostatistical Software Library and User's Guide*. 2nd edition, 369p. Oxford University Press.

Ferreya, R.A., Apezteguía, H.P., Sereno, R. and Jones, J.W. (2002). Reduction of soil water sampling density using scaled semivariograms and simulated annealing. *Geoderma*, **110**: 265–289.

Ferri, M. and Piccioni, M. (1992). Optimal selection of statistical units. Computational Statistics and Data Analysis, **13**: 47–61.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press. 483p.

Gatrell, A.C. (1979). Autocorrelation in spaces. *Environmental and Planning A*, **11**: 507–516.

Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley.

Griffith, D. (1987). *Spatial Autocorrelation: A Primer*. Washington, DC: AAG.

Griffith, D. and Amrhein, C. (1997). *Multivariate Statistical Analysis for Geographers*. New Jersey: Prentice Hall. 345p.

Grötschel, M. and Lovàsz, L. (1995). Combinatorial optimization. In: Graham R.L., Grötschel and Lovàsz (eds), *Handbook of Combinatorics*, Vol. 2; pp. 1541–1579. Amsterdam, The Netherlands: Elsevier.

Haining, R.P. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge University Press. 452p.

Hedayat, A.S. and Sinha, B.K. (1991). *Design and Inference in Finite Population Sampling*. New York: Wiley. 377p.

Hengl, T., Rossiter, D.G. and Stein, A. (2003). Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Australian Journal of Soil Research*, **41**: 1403–1422.

Iachan, R. (1985). Plane sampling. *Statistics and Probability Letters*, **3**: 151–159.

Isaaks, E.H. and Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics*. New York: Oxford University Press. 561p.

King, L.J. (1969). *Statistical Analysis in Geography*. 288p.

Lajaunie, C., Wackernagel, H., Thiéry, L. and Grzebyk, M. (1999). Sampling multiphase noise exposure time series. In: Soares A., Gomez-Hernandez J. and R. Froidevaux (eds), *GeoENV II – Geostatistics for Environmental Applications*; pp. 101–112. Dordrecht: Kluwer Academic Publishers.

Madow, W.G. (1953). On the theory of systematic sampling. III. Comparison of centered and random start systematic sampling. *Annals of Mathematical Statistics*, **24**: 101–106.

Makarovic, B. (1973). Progressive sampling for digital terrain models. *ITC Journal*, **15**: 397–416.

Matérn, B. (1960). *Spatial variation*. Berlin: Springer-Verlag. 151p.

McBratney, A.B. and Webster, R. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables: II. Program and examples. *Computers and Geosciences*, **7**: 331–334.

McBratney, A.B., Webster, R. and Burgess, T.M. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables: I. Theory and method. *Computers and Geosciences*, **7**: 335–365.

McBratney, A.B. and Webster, R. (1986). Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. *Journal of Soil Science*, **37**: 617–639.

Michalewicz, Z. and Fogel, D. (2000). *How to Solve It: Modern Heuristics*. Berlin: Springer. 467p.

Moran, P.A.P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, **10**: 245–251.

Moran, P.A.P. (1950). Notes on continuous phenomena. *Biometrika*, **37**: 17–23.

Muller, W. (1998). *Collecting Spatial Data: Optimal Design of Experiments for Random Fields*. Heidelberg: Physica-Verlag.

Olea, R.A. (1984). Sampling design optimization for spatial functions. *Mathematical Geology*, **16**: 369–392.

Oliver, M.A. and Webster, R. (1986). Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. *Geographical Analysis*, **18**: 227–242.

Overton, W.S. and Stehman, S.V. (1993). Properties of designs for sampling continuous spatial resources from a triangular grid. *Communications in Statistics – Theory and Methods*, **21**: 2641–2660.

Pardo-Igúzquiza, E. (1998). Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing. *Journal of Hydrology*, **210**: 206–220.

Pettitt, A.N. and McBratney, A.B. (1993). Sampling designs for estimating spatial variance components. *Applied Statistics*, **42**: 185–209.

Quenouille, M.H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics*, **20**: 355–375.

Rogerson, P.A., Delmelle, E.M., Batta, R., Akella, M.R., Blatt, A. and Wilson, G. (2004). Optimal sampling design for variables with varying spatial importance. *Geographical Analysis*, **36**: 177–194.

Ripley, B.D. (1981). *Spatial statistics*. New York: Wiley. 252p.

Stehman, S.V. and Overton, S.W. (1996). Spatial sampling. In: Arlinghaus, S. (ed.) *Practical Handbook of Spatial Statistics*; pp. 31–64. Boca Raton, FL: CRC Press.

Thompson, S.K. and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: Wiley. 288p.

Van Groenigen, J.W. (1997). Spatial simulated annealing for optimizing sampling – different optimization criteria compared. In: Soares, A., Gómez-Hernández, J. and Froidevaux, R. (eds). *GeoENV I – Geostatistics for Environmental Applications*. Dordrecht: Kluwer Academic Publishers.

Van Groenigen, J.W. (2000). The influence of variogram parameters on optimal sampling schemes for mapping by kriging. *Geoderma*, **97**: 223–236.

Van Groenigen, J.W. and Stein, A. (1998). Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, **27**: 1078–1086.

Van Groenigen, J.W., Siderius, W. and Stein, A. (1999). Constrained optimisation of soil sampling for minimization of the kriging variance. *Geoderma*, **87**: 239–259.

Van Groenigen, J.W., Pieters, G. and Stein, A. (2000). Optimizing spatial sampling for multivariate contamination in urban areas. *Environmetrics*, **11**: 227–244.

Warrick, A.W. and Myers, D.E. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research*, **23**: 496–500.

Webster, R. and Oliver, M.A. (1993). How large a sample is needed to estimate the regional variogram adequately. In: Soares, A. (ed.), *Geostatistics Tróia '92*; pp. 155–166. Dordrecht: Kluwer Academic Publishers.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, **40**: 974–994.

Zubrzycki, S. (1958). Remarks on random, stratified and systematic sampling in a plane. *Colloquium Mathematicum*, **6**: 251–264.

11

Statistical Inference for Geographical Processes

Chris Brunsdon

It is often necessary to make informed statements about something that cannot be observed or verified directly. It is equally useful to assess how reliable these statements are likely to be. A great deal of research is based on the collection of data, both qualitative and quantitative in order to make such statements. For this reason, inference in science is a fundamental topic, and the development of theories of *statistical inference* should be seen as a cornerstone of any field of study claiming to be based on scientific method. Indeed, the American Association for the Advancement of Science (AAAS) listed the development of the chi-squared test as one of the twenty key scientific developments of the twentieth century.[1]

In general, the success of the statistical hypothesis testing methodology is reflected in the vast number of publications in which some form of statistical test appears, and in the wide range of software packages (spreadsheets, statistical packages and others) in which code for carrying out such techniques appears.

However, despite this clear recognition of the importance of statistical inference, many commercial GIS packages claiming to offer 'spatial analysis' facilities have no procedures for this. The reasons for this are complex, but one thing to note is that it was the chi-squared test, and not statistical inference in general that was cited by the AAAS as a key development. Chi-squared tests are relatively simple computationally, and make a number of assumptions about the simplicity of the underlying processes about which inferences are to be made. In particular, they assume that each observation is probabilistically independent, and drawn

from the same distribution. For spatial data this is unlikely to be the case – recall Tobler's law stating that nearby things are likely to be more related than distant things. In addition, the distributions of observations may well depend on their geographical location. This violates the 'drawn from the same distribution' assumption. Thus, although tools of inference are just as important for geographical data as for any other kind of data, there are potential problems when 'borrowing' standard statistical methods and applying them to spatial phenomena. The aim of this chapter is to consider some fundamental ideas about inference, and then to discuss some of the difficulties of applying these ideas on to spatial processes – and hopefully offer a few constructive suggestions. It is also important to note that although for some areas a degree of consensus has been reached, the subject of statistical inference is not without its controversies – see Fotheringham and Brunsdon (2004) for example, and in particular there are unresolved issues in inference applied to geographical data.

11.1. BASIC CONCEPTS OF STATISTICAL INFERENCE

To begin it is important to identify – and distinguish between – some key concepts of statistical inference. These are:

- ***The inferential framework.*** This is essentially the model of how inferences are made. Examples of these are *Bayesian inference* (Bayes, 1763) and *classical inference*. Each model provides a characteristic set of general principles underpinning how some kind of decision related to a model (or set of models) can be made, given a set of observations.

- ***The process model.*** This is a model, with a number of unknown parameters, describing the process that generated the observations. This will take a mathematical form, describing the probability distribution of the observations. The mathematical model can be very specific, so that only a small number of parameters are unknown – or quite broad – so that for example a mathematical function of the general form $f(x, y)$ is not known.

- ***The inferential task.*** The task that the analyst wishes to perform having obtained his or her data. Typical tasks will be testing whether a hypothesis about a given model is true, estimating the value of a parameter in a given model, or deciding which model out of a set of candidates is the most appropriate.

- ***The computational approach.*** Having chosen a process model, the inferential framework should determine what mathematical procedure is necessary to carry out the inferential task. In many cases, the procedure is the relatively simple application of a simple formula (for example a chi-squared test). However, sometimes it is not. In such cases alternative strategies are needed. Sometimes they involve numerical solution of equations or optimizations. In other cases Monte Carlo simulation-based approaches are used, where characteristics of statistical distributions are determined by simulating variables drawn from those distributions. The strategy used to carry out the task is what will be termed the 'computational approach' here.

Probably the most fundamental of these concepts is the inferential framework. This is also the most invariant across different kinds of statistical applications – even if geographers have special process models or computational approaches, or inferential tasks, most of the time they are still appealing to the same fundamental principles when they draw inferences from their data. For example, one frequently sees geographers declare parameters in models to be 'significantly different from zero', or quote confidence

intervals. When they do so, they are making use of two key ideas from classical inference[2] which may be applied to geographical and non-geographical problems alike.

The most geographically specific of the concepts is the process model. As stated earlier, many inferential tests are based on the assumption that observations are independent of one another – in many geographical processes (such as those influencing house prices) this is clearly not the case. In some cases, the geographical model is a generalization of a simpler aspatial model – perhaps the situation where geography plays no role is a special case where some parameter equals zero. In these situations, one highly intuitive inferential task is to determine whether this parameter *does* equal zero. In other cases, the task is to estimate the parameters (and find confidence intervals) that appear in both spatial and aspatial cases of the models (for example regression coefficients). In these cases, the spatial part of the model is essentially a nuisance, making the inferential task related to another aspect of the model more difficult.

The previous examples are relatively simple from a geographical viewpoint, but more sophisticated geographical inferential tasks can be undertaken. In particular, the tasks above are related to what Openshaw (1984) terms 'whole-map statistics'. That is, they consider single parameters (or sets of parameters) that define the nature of spatial interaction at all locations, but supply no information about any specific locations. To the geographer, or GIS user, it is often more important to identify *which* locations are in some way different or anomalous. Arguably, this is a uniquely geographical inferential task. Although this inferential task can be approached with standard inferential frameworks, some careful thought is required.

Thus, to address the issue of statistical inference for geographical data one must consider the nature of statistical inference *in general*, the particular nature of statistical inference when spatial processes are considered and the way in which these two are related. This provides a broad framework for the chapter. First, a (very) brief overview of the key statistical inferential frameworks will be outlined. Next, spatial process models and related inferential tasks will be considered, together with a discussion of how the inferential approaches may be applied in this context. Finally, a set of suggested computational approaches will be considered.

11.2. AN OVERVIEW OF FORMAL INFERENTIAL FRAMEWORKS

The two most commonly encountered inferential frameworks are Classical and Bayesian. Suppose we assume a model M with some unobserved parameters θ, and some data x. Two kinds of tasks commonly encountered are:

1 Given M and x, to infer whether some statement about θ is likely to be true.

2 Given M and x, to estimate the value of θ or some function of θ, $f(\theta)$.

Although both methods can address both types of question, they do so in quite different ways.

11.2.1. Classical inference

The classical framework is most commonly used, and will be defined first. The classical framework generally addresses two kinds of inferential tasks. The first task is dealt with using the *significance* test.

Hypothesis testing

The statement about θ mentioned above is termed the *null hypothesis*. Next a *test statistic* is defined. Of interest here is the distribution of the test statistic if the null hypothesis is true. The *significance* (or *p*-value) of the test statistic is the probability of obtaining a value at least as extreme as the observed value of the test statistic if the null hypothesis is true. When the significance is very low, this suggests that the null hypothesis is unlikely to be true. To perform an $\alpha\%$ significance test one calculates the value of the test statistic with a significance of α – this is called the *critical value*. Typical values of α are 0.05 and 0.01. If the observed value is more extreme than the critical value, then the null hypothesis is rejected. Note that adopting the above procedure has a probability of α of rejecting the null hypothesis when it is actually true.

This may seem rather abstract without an example. One commonly used technique based on these principles is the two-sample *t*-test. Here $\theta = (\mu_1, \mu_2)$ where μ_1 and μ_2 are means of two normally distributed samples having the same variance σ^2. The null hypothesis here is that $\mu_1 = \mu_2$. Here the test statistic is the well-known *t*-statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (11.1)$$

where x_1 and x_2 are the sample means from the two samples, n_1 and n_2 are the respective sample sizes, and s^2 is defined by:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad (11.2)$$

where s_1^2 and s_2^2 are the respective sample variances for the two samples. A significance test is often performed by looking up the critical value of t from a set of tables, computing the observed t for the two samples and comparing this to the critical value. In this case, the t statistic has $\nu = n_1 + n_2 - 2$ degrees of freedom.

The above outlines the procedure of a significance test, one of the two inferential tasks performed using classical inference. Of course, such inference is probabilistic – one cannot be certain if we reject the null hypothesis that it really is untrue. However, we do know what the probability of incorrectly rejecting the null hypothesis is. This kind of error is referred to as the *type I error*. Another form of error results when we incorrectly accept the null hypothesis – this is called a *type II error*. It is generally harder to compute the probability of committing a type II error – usually denoted as $1 - \beta$. The relationship between α and β is given in Table 11.1.

For the two-sample *t*-test, the null hypothesis is $\mu_1 = \mu_2$, and the alternatives to this take the form $\mu_1 \neq \mu_2$, or equivalently $\mu_1 - \mu_2 = k$ for $k \neq 0$, β will depend on the value of k. In general, if k is large then there is a stronger chance of obtaining a significant t value, and so a smaller chance of incorrectly failing to reject the null hypothesis. β also depends on the values of n_1 and n_2 the sizes of the two samples. The larger these quantities are, the smaller the probability of incorrectly failing to reject the null hypothesis. Given any

Table 11.1 Relationship between α and β

Probability	*Reject null hypothesis*	
	Yes	*No*
Null hypothesis true	$1 - \alpha$	α
Null hypothesis false	$1 - \beta$	β

three of k, β, n_1 and n_2 one can compute the fourth (although the computation is not always simple).

Estimating parameters

The other inferential task is that of estimating θ or $f(\theta)$. As with hypothesis testing, we cannot be sure that our estimate of θ or $f(\theta)$ is exact – indeed given the fact that it is estimated from a sample we can be almost certain that it is not. Thus, in classical inference the key method provides upper and lower bounds – the so-called *confidence interval* for θ or $f(\theta)$. Note that this assumes that θ or $f(\theta)$ are scalar quantities. The situation when they are not will be discussed later. A confidence interval is a pair of numbers a and b computed from the sample data, such that the probability that the interval (a, b) contains θ is $1 - \alpha$. This probability is computed on the assumption that the model M is known in advance, up to the specification of θ. A very important distinguishing characteristic of this approach is that the probability quoted for a confidence interval is *not* the probability that a random θ lies within the deterministic interval (a, b) – rather it is the other way round – θ is not a random variable – under classical inference it is a fixed but unobservable quantity. It is the variables a and b that are the random variables, since they are computed from the random sample of observations – and so the probability statement is made about the random quantities a and b.

In situations where θ is not a scalar, one may specify confidence region*s* from the data. For example, if θ is two-dimensional, we could represent it as a point in the plane. A confidence region is some sub-region of the plane determined from the sample data that has a $1 - \alpha$ probability of containing the true θ.

11.2.2. Other issues for classical inference

In the section on 'Hypothesis testing' it was assumed that the quantity α could be easily calculated. In some situations this is not the case, because the probability of the test statistic, although known, cannot be manipulated analytically – making α impossible to compute directly. In such situations, a *Monte Carlo* (Metroplois and Ulam, 1949) approach may be more helpful. In this approach, a large number of random numbers are drawn from the probability distribution of the test statistic that would apply under the null hypothesis, and the observed value of the statistic is compared against this list (see Manly (1991) for some examples). It may be checked that the percentage rank of the observed test statistic when it is merged with the list of randomly generated test statistics is itself a significance level. Thus, provided we may generate random numbers from the distribution of the test statistic, this provides an alternative approach to the classical significance test – albeit one with a very different computational approach. This approach may also be used to generate confidence intervals.

Another important observation is that the derivation of the test statistics hinges on the model for the distribution of the observational data being known – at least up to the parameters being estimated. Sometimes this is not the case. Attempts to draw inference from data when this is the case are known as non-parametric statistics (see, for example, Siegel (1957)). One particular non-parametric approach is the so-called *permutation test*. This is a technique used to test relationships between pairs of variables, or more generally data sets in which the order of the observations is of some consequence. For example, if we have data taken from two samples, say S_1 and S_2 with respective sizes n_1 and n_2 – then we

could write the data as one long list, with all of the observations in S_1 followed by those in S_2:

$$\{x_1, x_2, \ldots, x_{n1}, x_{n1+1}, \ldots, x_{n1+n2}\}. \quad (11.3)$$

In this case the ordering of the observations is of some consequence in the sense that an observation with an index greater than n_1 must have come from S_2. Now suppose we wish to test the hypothesis that both sets of observations come from distributions with the same mean. Consider the quantity:

$$d = \frac{1}{n_1} \sum_{i=1}^{i=n_1} x_i - \frac{1}{n_2} \sum_{i=n_1+1}^{i=n_2} x_i. \quad (11.4)$$

Suppose a null hypothesis that S_1 and S_2 come from the same distribution. Then, there is no difference between the processes generating the observations in $\{x_1, \ldots, x_{n1}\}$ and $\{x_{n1+1}, \ldots, x_{n1+n2}\}$ – so that in fact any ordering of $\{x_1, \ldots, x_{n2}\}$ is equally likely.

Then, *regardless of the distributions of* S_1 *and* S_2 we would expect sample mean of d to be zero. We could use this quantity as a test statistic, although we do not know its distribution. However, if the null hypothesis were true, we may make use of Monte Carlo methods. We simply randomly permute the ordering of the data set a large number of times, and obtain a corresponding set of values of d. We then compare the observed value of d against this set, to obtain a value of α as before. This in essence is the randomization test. Here, it was shown in the context of a test of difference of means, although it may be used to test any kind of statistic dependent on the ordering of the observations. The advantage of this approach is that it allows tests to be made when one has no strong evidence of the distribution generating the data. A price paid for this is that the computational overhead is much higher – and typically nonparametric tests are not as powerful as the simpler parametric equivalents, provided the assumptions underlying the parametric tests hold. A final point is that there is a subtle difference between randomization tests and standard classical tests, in that they are conditional on the exact set of observed x values, i.e., the null hypothesis only considers the same values of x_i in different orders unlike, for example, a t-test which considers a sampling frame that could generate *any* real values of x_i.

11.2.3. Simple classical inference in action

To illustrate some of the above ideas a simple example is given. Here, the data consists of a number of sale prices of houses from two adjacent districts in the greater London area in 1991. The location of the districts in the context of greater London as a whole is shown in Figure 11.1, as are the locations of the houses in the sample. There are 220 houses in district 1 and 249 in district 2 (the district to the west).

If we assume that house prices in both districts have independent normal distributions with equal variances, we may test the hypothesis that the mean house price is the same in each district. This null hypothesis, together with the assumptions set out above, lead to the use of t-test as set out in equation (11.1). The values of the relevant quantities are set out in Table 11.2.

Since we are interested in detecting differences in the mean value of either sign, we use the absolute value *of* t which is 2.37. However, from tables, the critical value of t for (two-tailed) $\alpha = 0.05$ is 1.96 – suggesting we should reject the null hypothesis at the 5% level. Thus, with a

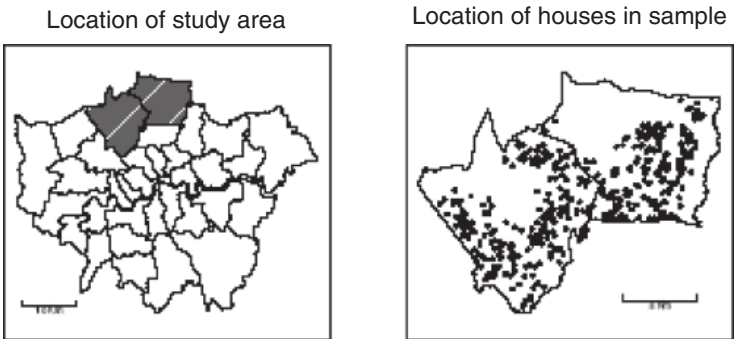


Figure 11.1 The location of study area (LHS) and the houses in the samples (RHS).

Table 11.2 Two sample *t*-test

District 1		*District 2*	
n_1	220	n_2	249
x_1	77.7	x_2	86.4
s_1	37.3	s_2	41.5
s_2	39.6		
ν	467		
t	−2.37		

5% chance of making an incorrect statement if the null hypothesis is true, we reject the null hypothesis – and state that there is a difference in average house price between two zones.

11.2.4. Bayesian Inference

The Bayesian approach views θ in a very different way. Whereas classical inference regarded θ as a deterministic but unknown quantity, Bayesian inference regards it as a random variable. The idea is that the probability distribution of θ represents the analyst's knowledge about θ – so that, for example, a distribution with very little variance suggests a great deal of confidence in knowing the value of θ. If we accept that θ is a random quantity, as is x, the observed data, we can consider the joint probability density of the two items given model M, say $f(x, \theta \mid M)$. Standard probability theory tells us that:

$$f(x,\ \theta \mid M) = f(x \mid \theta,\ M) f(\theta \mid M)$$
$$= f(\theta \mid x,\ M) f(x \mid M) \quad (11.5)$$

where $f(x \mid .)$ and $f(\theta \mid .)$ denote marginal distributions of x and θ, respectively. Dropping the M from the notation – as is conventional because everything is conditional on model M applying – we may write:

$$f(\theta \mid x) = f(x \mid \theta) f(\theta)/f(x). \quad (11.6)$$

Assuming we have a given observed data set x, we may regard $f(x)^{-1}$ as a normalizing constant and write:

$$f(x \mid \theta) \propto f(x \mid \theta) f(\theta). \quad (11.7)$$

This is essentially Bayes' theorem, and is the key to the inferential model here. If we regard $f(\theta)$ as the analysts knowledge about θ regardless of $\mathbf{x}$, then multiplying this by the

probability of observing x given theta (that is, $f(x \mid \theta)$, gives an expression proportional to $f(\theta \mid x)$. Note that in this framework, $f(x \mid \theta)$ is our process model, as set out in section 11.1. We can interpret this last expression as the knowledge the analyst has about θ *given* the observational data x. Thus, we have updated knowledge about θ in the light of the observations x – this is essentially the inferential step.

In standard Bayesian terminology $f(\theta)$ is referred to as the *prior* or *prior distribution* for θ and $f(x \mid \theta)$ is referred to as the *posterior* or *posterior distribution* for θ. Thus, starting out with a prior belief in the value of θ, the analyst obtains observational data x and modifies his or her belief in the light of these data to obtain the posterior distribution. The approach has a number of elegant properties – for example, if individual data items are uncorrelated and if data is collected sequentially, one can use the posterior obtained from an earlier subset of the data as a prior to be input to a later set of data. However, the approach does require a major change in world view. The requirement of a prior distribution for θ from an analyst could be regarded as removing objectivity from the study. Where does the knowledge to derive this prior come from?

One way of overcoming this is the use of *non-informative* priors which represent no knowledge of the value of θ prior to analysis. For example, if θ were a parameter between 0 and 1, then $f(\theta) = 1$ – a uniform distribution – would be a non-informative prior since no value of θ has a greater prior probability density than any other. Sometimes this leads to problems – for example if θ is variable taking any real value. In this case, $f(\theta) =$ const. is not a well-defined probability density function. However, this shortcoming is usually ignored provided the posterior probability thus created is valid (typically the posterior in this case could be regarded as a limiting value of an infinite sequence of posteriors derived from well-defined priors – for example if a sequence of priors with variances increasing without bound were supplied). A prior such as this is termed an improper *prior*.

Having arrived at a posterior distribution $f(x \mid \theta)$ we may begin to address the two key inferential questions:

(1) ***Estimate the value of θ or some function of θ, f(θ).*** Since we have a posterior distribution for θ we can obtain point estimates of θ using estimates of location for the distribution – such as the mean or median. Alternatively, we can obtain interval estimates such as the inter-quartile range derived from this distribution. Typically, one would compute an interval $[\theta_1, \theta_2]$ between which θ has a 0.95 probability of lying. Note that this is subtly different from the confidence interval of classical inference. The 95% in a confidence interval refers to the probability that the randomly sampled data provides a number pair that contains the unobserved, but non-random θ. Here we treat θ as a random variable distributed according to the posterior distribution obtained from equation (11.7). To emphasize that these Bayesian intervals differ from confidence intervals, they are referred to as *credibility intervals*.

(2) ***Infer whether some statement about θ is likely to be true.*** If our statement is of the form $a < \theta < b$ where either a or b are infinite, then this may be answered by computing areas underneath the posterior density function. For example, to answer the question 'is θ positive?' one computes:

$$\int_0^{\infty} f(\theta|x)\, \mathrm{d}\theta$$

and obtains the probability that the statement is true. However, questions of the form addressed by classical inference – such as 'is θ zero?' where typically one is concerned

with *exact* values of θ present more difficulties. Since the output is a probability density, the probability attached to any point value is zero. There are a number of workarounds to this. One quite sensible approach is to decide how far from zero θ could be for the difference to be unimportant, and term this ϵ. If this is done, we may then test the statement $-\epsilon < \theta < \epsilon$ using the above approach. Other approaches do attempt to tackle the exact value test directly – see Lee (1997) for further discussion.

Some final notes

Note that in the above sections θ is regarded as a univariate and continuous variable; however, the arguments may be extended to multivariate and discrete θ. In the discrete case, integrals are replaced by sums – and point hypothesis testing is no longer an issue. In the multivariate case, single integrals are replaced by multiple integrals – and instead of simple ranges for credibility intervals, regions in multidimensional parameter space may be considered.

11.2.5. Bayesian inference in action

In this section, we revisit the house price example, this time applying a Bayesian inferential framework to the problem. As before, we assume that house prices are independently normally distributed in each of the two districts. If we regard our list of house prices as x, then $\theta = (\mu_1, \mu_2)$ the respective means of the house price distribution for districts 1 and 2, and $f(x \mid \theta)$ is just the product of the house price probability densities for each observed price. Here we are interested in the quantity $\mu_1 - \mu_2$. In this case we have a non-informative prior in μ_1 and μ_2 and also in $\log \sigma$ where σ is the standard deviation of house prices in both districts. The choice of the prior for σ may seem strange, but essentially stems from the fact that this is a scale parameter, rather than one of location – see Lee (1997) for example. In this case, it can be shown that the posterior distribution for the quantity $\delta = \mu_1 - \mu_2$ is that of the expression:

$$(\bar{x}_1 - \bar{x}_2) - s\left(\frac{1}{n_1} - \frac{1}{n_2}\right)^{1/2} t \qquad (11.8)$$

where all variables are as defined in equation (11.1) except for t, which is a random variable with a t distribution with ν degrees of freedom (again ν is as defined earlier). The posterior distribution for δ is shown in Figure 11.2.

Here, the hypothesis under test differs from that of the classical test. Rather than a simple test of whether $\delta = 0$ – which makes little sense given the posterior curve above, we test whether $|\delta| < G$ where G is defined as some quantity below which a difference in means would be of little consequence. This is very different from the standard classical approach. In that framework, if a test were sufficiently powerful, differences in mean house prices of pennies could be detected. However, in terms of housing markets such a difference is of no practical importance. For this example we choose G to be £1,000 (UK). If this is the case, the probability that $|\delta| < G$ corresponds to the shaded area in Figure 11.2. This is equal to 0.014 – alternatively one could state that the probability that $|\delta|$ exceeds £1,000 is $1 - 0.014 = 0.986$. Thus, from a Bayesian perspective, it seems very likely that there is a non-trivial difference between the mean house prices for the two districts. Another possibility is to compute the probability that district 2 has a higher mean than district 1. This is just the posterior probability that $\delta > 0$, which, from the curve is equal to 0.99 – again suggesting this is highly likely.

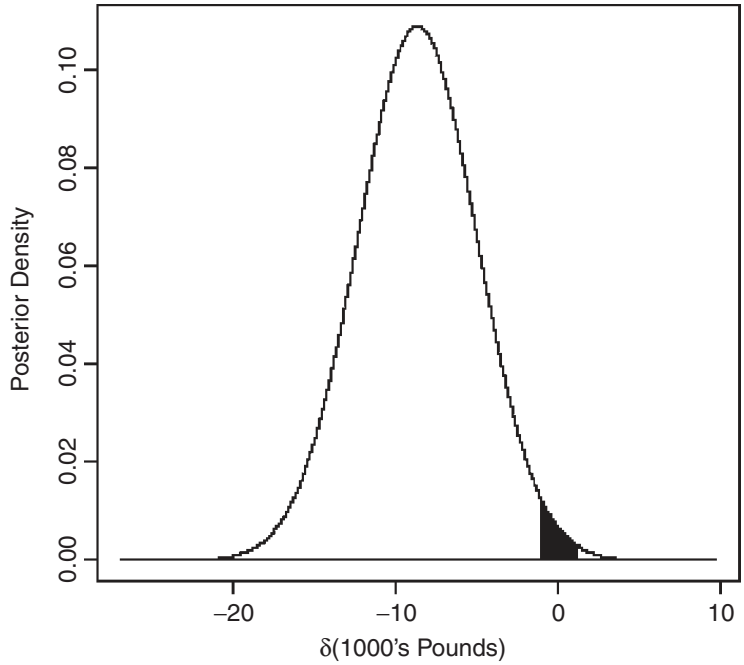


Figure 11.2 Posterior distribution for $\delta = \mu_1 - \mu_2$.

11.2.6. Bayesian approaches – some closing comments

The Bayesian approach is regarded by some as very elegant. Certainly the simplicity of the underpinning equation (11.7) and the natural way that hypotheses may be assessed, and parameters estimated from the posterior distribution do have a directness of appeal. However, there is a sting in the tail. Equation (7) gives the posterior distribution *up to a constant* – implying that the expression for probability distribution can only be obtained by integrating its un-normalized form. Herein lies the problem – in many cases the integral is not analytically tractable. At the time of writing, this presents fewer problems than in the past – as numerical quadrature techniques may be used to estimate the integrals. Alternatively, techniques based on Monte Carlo simulation and the Metropolis algorithm allow random values of θ to be generated according to the posterior distribution. In this case, hypotheses about θ are investigated by generating large numbers of random values, and investigating their properties.

Advances of the kind described above are not made without a great deal of research to answer important operational questions such as:

- How accurate are the quadrature results?
- How large should the samples of random numbers drawn using the Metropolis algorithm be?
- How can very large-scale simulations be computed efficiently?

Thus, the computational approach to Bayesian methods is an issue of great importance. However, in recent years this area has been the focus of much research, and this combined with increasing trends in the speed and

capacity of computers have led to an increase in the popularity of Bayesian methods.

11.3. WHY GEOGRAPHY MATTERS IN STATISTICAL INFERENCE

In the above section, two of the most common approaches to formal statistical inference were discussed. However, this was done in a general sense – nothing stated in the previous section applied exclusively to geographical data. As hinted in the introduction, working with spatial data introduces a few specific problems.

This raises a number of issues:

1 What happens if one ignores spatial effects?

2 Does one need to modify the above ideas of inference when working with spatial data?

3 If some spatial effects are present, can they be represented as spatial patterns or images?

All of these issues lead to important questions – but none have unique answers. First, consider issue 1 – if there are no serious problems encountered when ignoring spatial effects then there is little that spatial analysis can add to the 'usual suspects' list of standard statistical methods. However, it is argued here that there are indeed serious consequences arising from ignoring such effects. There are many examples of such consequences – see, for example, Fotheringham *et al*. (1998), who follow the work of Rees (1995) in modelling the relationship between limiting long-term illness (LLTI) as defined in the 1991 UK census of population, and a number of predictor variables: Unemployment rate, Crowding, Proportion in Social Class 1, Population Density, and Proportion of Single Parent Families. Full definitions of these variables may be found in Fotheringham *et al*. (1998). The study area consists of the four counties Tyne and Wear, Durham, Cleveland and North Yorkshire, in the north-east of England. Of particular interest here is the population density variable. An ordinary least squares regression model was fitted to the data, giving a coefficient of -5.6. A t-test based on principles of classical inference showed this to be significantly different from zero. In general, this suggests that an increase in population density leads to a decrease in LLTI. This is perhaps counter-intuitive. Normally one associates higher morbidity rates with urban areas, which have higher population densities. However, the study went on to consider geographically weighted regression (GWR) (Brunsdon *et al*., 1996) – a technique using an underlying model in which regression parameters vary over space. When this was carried out, it was found that the regression parameter for population density was at its most negative in areas in the region around the coalfields of east Durham. Here, it is likely that LLTI is linked to employment in the coalfields, and that most people in such employment lived in settlements near to the coalfields, where population density is low. However, those people living in urbanized areas in that part of the region are less likely to be employed in occupations associated with high LLTI. Thus, in that locality a negative relationship between population density and LLTI holds. However this is unusual in general, and in other parts of the study area (west Durham, North Yorkshire), there is a positive relationship. Here, low population density corresponds to a more typical rural environment, and in these places a more conventional urban/rural trend occurs. The key point here is that the global model told only one story, while the spatially-oriented GWR identified two different processes occurring in different parts of the study area. The 'moral' here is that ignoring geography

can lead to mis-interpretation. This example is a cautionary tale about the consequences of ignoring spatial effects in an inferential framework.

So ignoring geography can lead to inferential problems. How can this difficulty be overcome? In particular this raises another key question – 'Does one need to modify the above ideas of inference when working with spatial data?' To answer this, we return to the four aspects of statistical inference listed in section 11.1 once again: both Bayesian and classical *inferential frameworks* can handle the key *inferential tasks* of hypothesis evaluation and parameter estimation for spatial processes. However, for spatial data the *process model* must allow for geographical effects. Finally, it is also the case that the *computational approach* must also be altered on some occasions. These two key issues will be considered in turn.

11.3.1. Process models for spatial data

The process models for spatial data can differ from more commonly used ones in a number of ways. The two most common ones are that they exhibit *spatial non-stationarity* and *spatial autocorrelation*. Spatial non-stationarity is essentially the characteristic of the LLTI example above. The unknown parameter θ is not a constant, but in fact a function of spatial location. In this case, a technique like GWR may be used to estimate θ at a set of given localities. Using this approach, one can apply the classical inferential framework to obtain estimates of θ, and test hypotheses such as 'is θ a global fixed value'. A classical inferential framework for GWR is detailed in Fotheringham *et al*. (2002).

The phenomenon of spatial autocorrelation occurs when each of the observed x values are not drawn from statistically independent probability distributions, but are in fact correlated. In the geographical context, the correlation is generally related to proximity – nearby x values are more correlated than values located far apart. Typical examples are the SAR (spatial autoregression) and CAR (conditional autoregression) models. Unlike GWR, these regression models do not assume that the regression parameters vary over space – however they do assume that the dependent variables are correlated. Typically here, each record of variables is associated with a spatial unit, such as a census tract, and the spatial dependence occurs between adjacent spatial units. As well as the regression coefficients and the variance of the error term, CAR and SAR models have an extra parameter controlling the degree to which adjacent dependent variables are related. In the classical inference case, parameter estimation is typically based on maximum likelihood, with the parameter vector θ containing the extra parameter described above as well as the usual regression parameters. There is much work on the classical inferential treatment of such models: see, for example, Cressie (1991). LeSage (1997) offers a Bayesian perspective.

11.3.2. The computational approach

Computational issues for geographical data are generally complex. The whole field of geocomputation has grown to address this. As well as problems of data storage, data retrieval and data mining, there are many computational overheads attributable to inference in spatial data, for a number of reasons. In some cases, the issue is related to Monte Carlo or randomization methods – this is particularly true of the Monte Carlo Markov Chain approach to Bayesian analysis. In others, it is linked

to developing efficient algorithms to access large geographical data sets – this can be an issue in localized methods such as GWR. In each case, it is true than specific algorithms may need to be created to handle the geographical situation. A very good example is found in Diggle *et al*. (1998). Another example of this is shown graphically in figures 11.3 and 11.4. Figure 11.3 shows a map of crime rates. The intention here is to test whether the spatial autocorrelation (as measured by Moran's I) is zero. This is done by randomly permuting the rates to geographical zones. The distribution after 1000 simulations is shown in figure 11.4 – the observed value is far greater than any of these simulations, suggesting a highly significant ($p < 0.001$) result.

The final question in the earlier list also raises some interesting problems. The formal (Bayesian or classical) approach to hypothesis testing is essentially founded on the notion of testing a single hypothesis. However, many geographers would like answers to more complex hypotheses. In the spatial context, one of the key questions is 'Is there an unusually high or low value of some quantity in region R?' Typically this quantity might be the average price of a house, or an incidence rate of some disease. This phenomenon is often termed *clustering* (see the chapter by Jacquez in this volume for more discussion on this topic). In some situations R is known in advance – for example it may represent the catchment area of a particular school in the house price example. If it is known in advance the approach is relatively simple. One creates a proximity measure to reflect how close to R each observation is, or creates a 'membership function' of R for each observation, and then builds this into a model, using a parameter that may vary the influence of this new variable. Then one goes on to test the hypothesis that this parameter is zero (or whatever value of the parameter implies that proximity to R has no influence on the quantity of interest).

This approach fits in well with conventional theory – there is one single hypothesis to test, and it may be tested as set out above. However, on many occasions we have no prior knowledge of R, possibly even on whether R is a single region or a number of disjoint areas. On such occasions, a typical approach would be to carry out a test such as that described above on every possible region, and map the ones that have a significant result. This is essentially the approach of the Geographical Analysis Machine (GAM; Openshaw, 1987) – here the Rs are circular regions of several radii centred on grid points covering the study area. However, there is a difficulty with this approach. Suppose we carry out a significance test on each of the Rs. There could be a large number of tests, possibly hundreds. Even if no clustering were present, the chance of obtaining a false positive is α, the significance level of the test. If $\alpha = 0.05$ as is common practice, we would expect to find $N\alpha$ significant results even when no clustering occurs, where N is the number of regions to be tested. For example, if $N = 200$ and $\alpha = 0.05$, we would expect to find 10 significant regions *even when in reality no clustering occurs*. Thus, in an unadjusted form, this procedure is very prone to false positive findings. Essentially this is a problem of *multiple hypothesis testing*. Because the test has a positive probability of incorrectly rejecting the null hypothesis, carrying out enough tests will give some positive results even if in reality there are no effects to detect. A typical way of tackling the problem is to apply the *Bonferroni* adjustment to the significance levels of the test. For example, this is done by Ord and Getis (1995) for assessing local autocorrelation statistics.

The correction is derived by arguing that to test for clustering, we wish to test that *none* of the regions R have a significant

cluster centred on them. Thus, the probability of a false positive overall is the probability that any one of the regions has a false positive result. If it is assumed that each test is independent, then it can be shown that this probability (which will be called α') is given by:

$$\alpha' = 1 - (1 - \alpha')^N. \qquad (11.9)$$

Now, if we wish to develop an overall test for clustering, with say $\alpha' = 0.05$ then equation (11.9) may be solved for α – giving a significance level for the individual tests needed in order to achieve the overall level of significance. For example, if $N = 200$ and we require $\alpha' = 0.05$ then $\alpha = 0.000256$. This is a fairly typical result. To counter the risk of false positives, the individual tests must have very low values of α.

However, one thing of note about the above approach is that the assumption that the tests are independent is often incorrect for geographical studies. Typically, a large number of regions R are used, and many overlap, sharing part of the sample data used for the local tests – and for this reason the results of these tests cannot be independent. It is usually argued that the Bonferroni procedure provides conservative tests[3] – and in the situation where the tests are correlated the estimate of α in equation (11.9) is an underestimate. In an attempt to avoid false positives, we insist on very strong evidence of clustering around each of the test regions. Thus, we will be insisting on much stronger evidence than is actually necessary to detect clustering and there is some chance that genuine clustering is overlooked. In a nutshell this is a typical dilemma when looking for clusters – ignoring multiple hypothesis testing leads to false positives, but overcompensation for this could lead to false negatives. There are no inferential free lunches!

11.4. FURTHER ISSUES

In the previous sections, the two most common approaches to formal inference were discussed, and following this, some of the particular issues encountered when applying these principles to spatial data and spatial models were discussed. In this final section, other matters arising will be considered.

11.4.1. Population versus process

Throughout this chapter, the concept of inference has been applied to processes. However, another view is that one makes inferences about *Populations* given samples taken from these populations. In many respects, there are similarities between the two situations. In reality, every population is finite, and therefore the population that items in a sample are drawn from is discrete. Therefore, strictly speaking tests such as the *t*-test are inappropriate in this situation, as they assume observations are drawn from a normal distribution – which is continuous. However, when the population is very large, this distribution (or in some cases another continuous distribution) is a very good approximation of reality – for example in the UK a population of around 20 million would represent every household – but a continuous distribution for household income may well be quite close to the real situation of a discrete distribution with around 20 million values!

Thus, it is argued here that the random *process* of drawing from a normal (or other continuous) distribution is a very good approximation to drawing from a very large population – and that in many cases, hypothesis tests related to the population can be reasonably proxied by hypotheses relating to such an approximating process. Given this argument, all of the arguments based on the concept of process inference here may

be applied to making inferences about large populations.

For smaller populations, the discrete nature of the sampling frame may suggest that such continuous approximations are not valid. Here, we are faced with two choices. First, if we assume that this population really is the item of interest, and it is not particularly large, then one approach might be to collect observations for the *entire* population. In this case, the conventional framework for statistical hypothesis testing becomes meaningless – to test a hypothesis relating to the population simply look at the data and see if it is true or not!

A second alternative is to assume that the population itself is of less interest than the process generating it. In this case, we return to the process hypothesis framework. Note that on some occasions, we may have the entire population represented in our data, but even so it may be of interest to understand the process(es) that brought about that data. For example, we look at daily records of rainfall from a one month period of the previous century. In this case, the list of rainfall measurements is our population, but the process of generating these can be modelled as a random process – and we may wish to test hypotheses about whether average levels are similar to those in the present day. In this case, we wish to test the (process-based) hypothesis that the mean daily rainfall is equal to some given level. It is the author's opinion that in most cases when an entire population may be measured, it is the underlying process and not the values of the population itself that is of most interest.

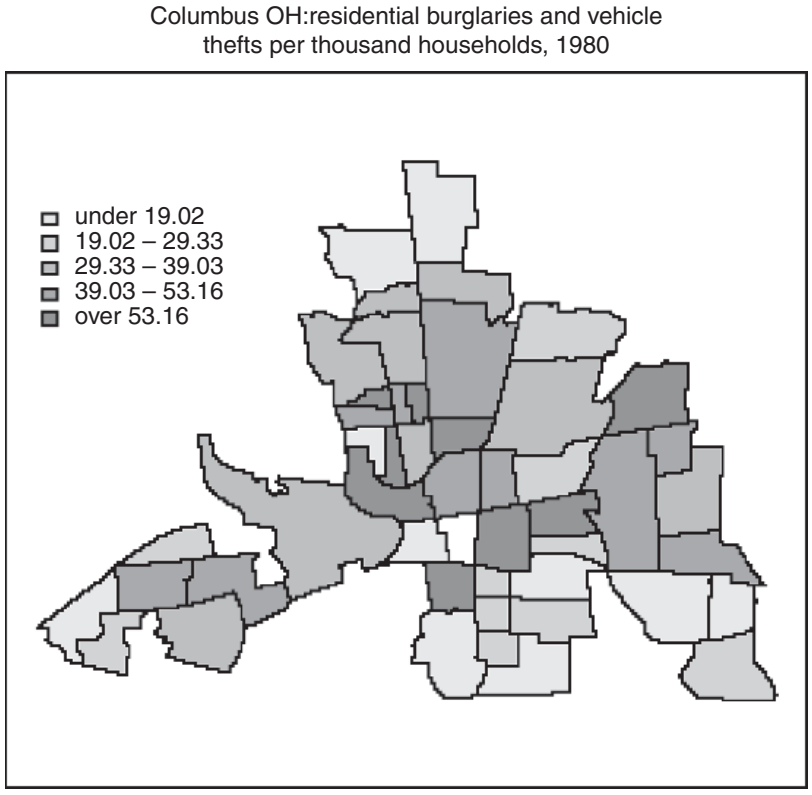


Figure 11.3 Crime rate distribution: vehicle thefts and residential burglaries per 1000 households (1980).

11.4.2. Other types of inference

Although classical and Bayesian methods are both covered in this chapter, these are not the only possible approaches. For example Burnhan and Anderson (1998) outline ways in which Akaike's An Information Criterion (AIC; Akaike 1973) may be used to compare models. This approach is quite different in terms of its inferential task – rather than testing whether a statement about a particular model is true – or assuming a specific model holds and then attempting to estimate a parameter of that model, this approach takes several models and attempts to identify which one is 'best' in the sense that it best approximates reality. The AIC is an attempt to measure the 'nearness' of the model to reality – obviously the true model is not known, but the observations have arisen from that model, and this is where the 'clues' about the true model come from. This is very different from the other approaches because it regards all potential models as compromises – none is assumed to be perfect – and attempts to identify the best compromise. This area may prove fruitful in the future – for example Fotheringham *et al*. (2002) use a method based on this idea to calibrate GWR models. The idea of finding a 'best approximation' also sits comfortably with the idea of approximating a large finite sample with a continuous distribution put forward in the previous section.

Of course, exploratory data analysis can be thought of as yet another inferential framework, albeit a less formal one. Although this can provide a very powerful framework for discovering patterns in data, it could be argued that this is an entire subject in its own right, and that there will be many examples elsewhere in this book, where the production of maps and associated graphics by various software packages provide excellent examples exhibiting the power and utility of graphical data exploration.

11.4.3. Software

No chapter about inference would be complete without some discussion of software. Having argued that making inferences about data is central to knowledge discovery in spatial analysis, one has every right to expect that software for inferential procedures will be readily available. However, as mentioned in the introduction, most readily available GIS packages do not currently contain code for many of the procedures outlined here. Unfortunately, although several commercial statistics packages do contain code for carrying out general inferential procedures, such as the *t*-test example discussed earlier in the chapter, they offer less support for more specific inferential tasks developed for spatial data. Until recently, for a number of spatial inferential tasks one was forced to write one's own code. However this situation is now improving. A number of packages that are either dedicated to the analysis of spatial data or sufficiently flexible that they may be extended to provide spatial data analysis now exist. Although by no means the only option, the statistical programming language *R* provides good spatial analysis options – all of the examples (most notably the spatial one) in this chapter were based on calculations done in *R*. There are a number of spatial data analysis libraries written in *R*, enabling this kind of geostatistical computation. For example:

- ***sp*** provides basic spatial data handling facilities;
- ***maptools*** provides map drawing functionality as well as the ability to import geographical data in a number of common formats, such as ArcGis shapefiles;
- ***spdep*** provides a number of hypothesis tests and model calibration facilities relating to models allowing for spatial dependencies; and

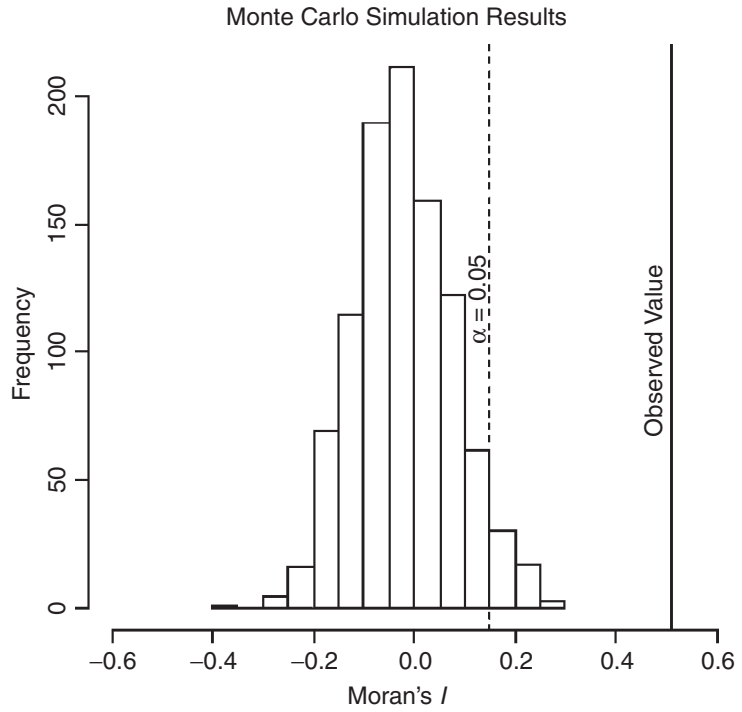


Figure 11.4 Results of Monte Carlo tests on *I*.

- ***spgwr*** provides a number of tools for *Geographically Weighted Regression* (GWR) analysis.

The package is also 'Open Source' so it provides an easy entry option for anyone wishing to experiment more with inferential approaches for geographical data.

ACKNOWLEDGEMENT

I am grateful to the Nationwide Building Society for providing the house price data first introduced in section 11.2.3.

NOTES

1 See http://stat.fsu.edu/brouchure/stat/whystat.htm for additional details.

2 These principles are the *significance test*, and the *confidence interval* respectively.

3 *Conservative* here means that the test has a significance level of 5% *or lower*.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B. and Csaki, F. (eds), *Proceedings of the Second International Symposium on Information Theory*, pp. 267–278. Budapest: Akademiai Kiado.

Bayes, T. (1763/1958). Studies in the history of probability and statistics: IX, Thomas Bayes' essay towards solving a problem in the doctrine of chances. *Biometrika* **45**: 296–315 (Bayes' essay in modernized notation).

Brunsdon, C. Fotheringham, A.S. and Charlton, M. (1996). Geographically weighted regression: A method for exploring spatial non-stationarity. *Geographical Analysis* **28**: 281–289.

Burnhan, K.P. and Anderson, D.R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer.

Cressie, N.A.C. (1991). *Statistics for Spatial Data*. New York: John Wiley and Sons.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998). Model-based geostatistics. *Applied Statistics* **47**: 299–350.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (1998). Scale issues and geographically weighted regression. In Tate, N. (ed.), *Scale Issues and GIS.* Chichester: John Wiley and Sons.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: John Wiley and Sons.

Fotheringham, A.S. and Brunsdon, C. (2004). Some thought on inference in the analysis of spatial data. *International Journal of Geographical Information Science* **18**: 447–57.

Lee, P.M. (1997). *Bayesian Statistics: An Introduction*. London: Arnold.

LeSage, J. (1997). Bayesian estimation of spatial autoregressive models. *International Regional Science Review* **20**: 113–129.

Manly, B. (1991). *Randomization and Monte Carlo Methods in Biology*. London: Chapman and Hall.

Metroplois, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association* **44**: 335–341.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Norwich: Quantitative Methods Research Group, Royal Geographical Society and Institute of British Geographers, Concepts and Techniques in Modern Geography Publication No. 38.

Openshaw, S. (1987). A mark *i* geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* **1**: 335–358.

Ord, J.K. and Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* **27**: 286–306.

Rees, P. (1995). Putting the census on the researcher's desk. In Openshaw, S. (ed.), *Census Users' Handbook*, pp. 27–82. Cambridge: GeoInformation International.

Siegel, S. (1957). *Nonparametric Methods for the Behavioral Sciences.* New York: McGraw-Hill.

12

Fuzzy Sets in Spatial Analysis

Vincent B. Robinson

12.1. INTRODUCTION

Shortly after the theory of fuzzy sets was introduced by Zadeh (1965) researchers began to argue that fuzzy sets theory could serve as an appropriate foundation for spatial analysis (Gale, 1972). It was argued early on that fuzziness is a major factor contributing to the uncertainty of spatial behavior. Thus, achieving exactitude in representing, analyzing, and predicting spatial behaviors, over space and through time is difficult, or impossible, to accomplish in a fuzzy environment characterized by ambiguous or incomplete information and inexact cognitive and decision-making processes. Although many of the earliest works focused on spatial behavior, policy, and planning (Leung, 1983; Lundberg, 1982; Pipkin, 1978), it was not long before its relevance was recognized in other areas of spatial analysis such as soil science (McBratney and Odeh, 1997) and the then developing field of geographic information science (Robinove, 1989; Robinson and Strahler 1984; Robinson, 1988).

For many spatial phenomena there are no crisp boundaries that can be identified to differentiate regions or zones. For examples, the boundary between beach and foreshore, between woodland and grassland, and between urban and rural areas may be gradual rather than defined by a crisp boundary. It is well known that when we use remotely sensed imagery to extract spatial objects of interest, there are pixels that may contain sub-pixel objects, trans-pixel objects, boundary pixels, and/or natural intergrades (Foody, 1999). The mixture of spectral information at the sub-pixel scale can lead to uncertain classification and indeterminate boundaries. This is not unrelated to the general region classification problem highlighted in Leung's (1983) evaluation of fuzzy sets in spatial

analysis and planning. In addition, it is often the case that concepts, or parameters, in spatially explicit models are inherently inexact. These, and other problems of uncertainty, have led many to use techniques based on fuzzy set theory. Ironically, fuzzy sets can be used to help make analyses less fuzzy because the inexactness is managed explicitly rather than implicitly. However, like other efforts at formalization, it can help lay bare assumptions and force us to be explicit about their meaning.

The basic idea underlying fuzzy set theory is that an element can be classified as being a member of more than one set and to varying degrees hold membership in each class. In the usual Boolean, or crisp, set theory, membership of an element x in a set $\boldsymbol{A}$, is defined by a characteristic function that indexes the degree to which the object in question is in the set. It should be noted that it is customary, but not strictly necessary, for the index to range from 0, for full non-membership, to 1.0 for full membership. Hence, the membership function is the fundamental element necessary to use fuzzy sets. A membership function measures the fractional truth value a statement such as 'Object Y is a member of set S'.

A variety of other works contain in-depth explanations of the relevant fundamental concepts of fuzzy set theory. Studies such as that by Klir *et al*. (1997), Buckley and Eslami (2002), and Zimmerman (2001) cover many aspects of fuzzy sets in considerable depth with applications as examples. More relevant to those interested in spatial analysis is the geographic information systems (GIS) textbook by Burrough and McDonnell (1998). Other informative perspective pieces have appeared in the social sciences (Verkuilen, 2005), soil science (McBratney and Odeh, 1997) and GIS (Robinson, 2003). Hence there are many sources that one can turn to for background on the fundamentals of fuzzy sets and their relevant use in spatial analysis.

This chapter will first briefly review some of the more noteworthy accomplishments using fuzzy sets in spatial analysis. Then it will discuss the issue of assigning fuzzy membership and how it has been approached for use in spatial analysis. Finally, it will briefly discuss some issues and challenges of using fuzzy sets in spatial analysis.

12.2. FUZZY SETS AND SPATIAL ANALYSIS: SOME ACCOMPLISHMENTS

Spatial analysis is a broad field not relegated to just social science, ecology, soil science, geography, or engineering. Fuzzy set theory has been specifically noted as being a more natural way of representing and analyzing phenomena in such diverse areas as social science (Ragin and Pennings, 2005), soil science (McBratney and Odeh, 1997), ecology (Schaefer and Willson, 2002) as well as geographical analysis and engineering. Thus, there are many areas in which it has been shown to be of value in spatial analysis. Some of the more recent and important accomplishments are noted in this section.

Fundamental to some types of spatial analysis is the generation of a surface from data that generally contains some level of uncertainty. These data are generally represented as a set of points. Not only may there be uncertainty regarding the data measurement, but there may be duplicate data records in the spatial database that may confound an analysis. Torres *et al*. (2004) present an asymptotically optimal algorithm for eliminating duplicates that incorporates the handling of fuzzy uncertainty. The generation of surfaces from point data entails some form of interpolation. In this regard there have

been several promising approaches presented for the interpolation of spatial surfaces from point data (Anile *et al*., 2003; Gedeon *et al*., 2003; Lodwick and Santos, 2003).

Sampling of spatial data is fundamental to spatial analysis. Fuzzy set theory has been used relatively rarely in this regard. It has been shown how a combination of fuzzy clustering and regionalized variables can be used to estimate the optimal spacing of sample collection sites for soils mapping (Odeh *et al*., 1990). Developments in mapping systems that integrate mobile computing, GIS, and one or more sensors to take physical measurements (Arvanitis *et al*., 2000) suggest that it is becoming realistic to think in terms of spatial data collection agents that use fuzzy logic in an adaptive spatial sampling strategy. Simulation results of a prototypical system for adaptive sampling along a transect suggest that the fuzzy adaptive sampler usually produced better results and on average required fewer sampling locations (Graniero and Robinson, 2003).

When using spatial analysis in support of spatial decision making, it is sometimes noted that the results of using a crisp, or nonfuzzy, approach to provide an information space for making decisions virtually guarantees that an analysis will ignore potentially useful information (Morris and Jankowski, 2005; Oberthur *et al*., 2000; Yanar and Akyurek, 2006). Thus, a nonfuzzy, or crisp, approach may have the effect of hiding important spatially explicit information from decision makers, hence increasing the risk of incurring additional costs by forgoing an opportunity because it was not known to the decision maker. This feature of fuzzy versus crisp approaches has been noted in studies as varied as landfill site selection (Charnpratheep *et al*., 1997), real estate evaluation (Zeng and Zhou, 2001), and soil erosion potential (Ahamed *et al*., 2000).

Often spatial decision making is represented using decision tables (DT). However, the problem of strict, crisp boundaries is viewed as a significant problem in the use of DTs for locational decision making and spatial analysis. Witlox and Derudder (2005) have demonstrated how fuzzy decision tables can be formulated and used effectively. They show that it is possible to explicate the imprecision involved in the decision-making process using FDTs. However, like DTs, when the number of conditions becomes large then knowledge-based techniques may be more effective and manageable.

The use of fuzzy sets in spatial analysis has been shown to improve the accuracy of representing spatial phenomena in a variety of domains. Often times this improvement is also coincident with a reduction in cost. Using a fuzzy similarity approach, Hwang and Thill (2005) found that the rate of success of a typically used georeferencing procedure went from 86% up to 94% of all fatal accidents. This may not sound like a great many instances, but in a mission-critical application such as locating fatal accidents, this represents a significant gain in accuracy. In a different domain, the soil–land inference model (SoLIM) based on fuzzy set theory (Zhu *et al*., 2001; Zhu, 1997) has been estimated to have increased the accuracy of spatially explicit soils data by as much as 20% at a third of the cost of tradition techniques (Zhu, 2004). A similar result of lower cost and higher accuracy when using a fuzzy logic-based methodology has been suggested by work on ecological landscape mapping (MacMillan *et al*., 2003; MacMillan *et al*., 2000).

For spatial analysis, map comparison is useful for purposes of studying dynamic processes such as land cover change, comparing simulation model results with empirical data, map creation/revision and translating between maps using different semantics. Translating between map products from

differing sources with differing semantics is a problem when assembling spatial data for analysis. To address this problem, Ahlqvist (2005) used rough fuzzy sets to analyze the semantic similarity of map products having differing classification semantics. Fritz and Lee (2005) used a fuzzy logic based methodology to compare two land cover datasets and found the fuzzy agreement approach superior to the nonfuzzy approach in identifying areas of severe disagreement. Using a hierarchical fuzzy pattern matching technique, Power *et al*. (2001) were able to convincingly demonstrate the superiority of the use of fuzzy logic to address the problems of map comparison. They noted the deficiencies of using a map comparison statistic such as the Kappa measure that relies on crisp, nonfuzzy categories. This has subsequently been addressed by Hagen and others in their development of the K-fuzzy (or fuzzy kappa) (Hagen-Zanker *et al*., 2005) that takes into consideration the fuzziness of both location and attribute quality (Hagen, 2003). This is one of the more promising approaches to comparing spatial fields (Wealands *et al*., 2005).

A variety of multi-criteria decision making efforts have used fuzzy techniques to address spatially explicit problems. A methodology for assessing land for allocation to restoration projects demonstrated that the additional information afforded by fuzzy classification can be of significance in avoiding misallocations that would result in unnecessary cost (Guneralp *et al*., 2003). Similar conclusions could be drawn from an earlier study on allocation of land for industrial use. Jiang and Eastman (2000) showed that results can vary significantly as a function of the method of aggregation. In their review of fuzzy-based approaches Kahraman *et al*. (2003) suggest that nonfuzzy, conventional approaches to the facility location problem tend to be less effective in dealing with the imprecise nature of linguistic assessments that are often part of the qualitative criteria.

It is not uncommon for fuzzy set theory applications to be incorporated in a component of a larger decision support system. For example, in the DISCUSS system of spatially disaggregating cost–benefit analyses fuzzy logic is used in only one component. The fuzzy spatial disaggregation method uses standard membership curves operating on spatial variables. If the initial method of spatial disaggregation is not accepted, then a fuzzy disaggregation method is used that is based on membership functions on distance variables and fuzzy addition. Using fuzzy sets this work has shown how cost benefit analyses can be spatially disaggregated, something that has rarely been accomplished in the past (Paez *et al*., 2006).

In some cases, when compared with more traditional methods that are statistics-based, fuzzy techniques have provided superior results. For example, when they replaced their principal components model with a fuzzy set analysis. Taylor and Derudder (2004) noted that the fuzzy-based analysis provided an exceptionally clear picture of regional and hierarchical tendencies among world cities. In a similar vein, Katz *et al*. (2005) concluded that regression analysis did not provide meaningful results while fuzzy set analysis did provide meaningful results. In quite a different domain, Kuo *et al*. (2003) incorporated a fuzzy analytical hierarchical process (AHP) to support the locational decision for convenience stores. Since the mean standard error (MSE) for the fuzzy AHP was 0.0173 as opposed to 0.091 for the regression model, Kuo *et al*. (2003) concluded that fuzzy AHP plus artificial neural network (ANN) decision support system provided more accurate results than did a regression model. In geographical soil science, Oberthur *et al*. (2000) showed that nonfuzzy approaches severely misclassified land while fuzzy

approaches were much more successful. In particular, Boolean classification allocated nearly 2,000 hectares of land to the group with low potential for plant recovery than did the fuzzy approach. Hence, leading to a result where less land was erroneously shown to have a high potential for plant recovery. In another physical science domain, Fisher *et al*.'s (2005) multi-scale fuzzy-based analysis provided significant insights over more conventional, nonfuzzy, analysis of the accumulation and erosion of material as reflected in elevation changes.

In addition to Taylor and Derudder's (2004) application of fuzzy clustering to world cities, Heikkila *et al*. (2003) used a two-pass Bayes classification method to assign membership values to urban objects that are ultimately used in the context of Kosko's (1992) fuzzy hypercube. They claim that the main contribution of their fuzzy urban set formulation is the introduction of a unifying conceptual framework for measures of urbanization. This is made possible by the representation of an entire study area as a single point within a fuzzy hypercube. In a fuzzy hypercube each axis corresponds to the membership of a particular fuzzy set. Hence, each axis is defined on the interval [0, 1]. A fuzzy system with n sets would generate a fuzzy hypercube of dimension $[0, 1]^n$.Using three dichotomies with fuzzy set interpretations, they exploited the geometric interpretation of fuzzy sets afforded by the fuzzy hypercube to show how study area could be 'located' within a three-dimensional hypercube. Aggregate measures were used to calculate the degree of membership a study area has in each of the three aggregate measures. These are used to locate the study area as fuzzy set in the hypercube.

In some spatially explicit applications hand-drawn sketch maps are used as a means of collecting spatial data. However, the inherent uncertainty of such maps is self-evident. Therefore, an important step towards automating the analysis of such spatial data is represented by Skubic *et al*. (2004) who use force histograms and fuzzy rules to generate a linguistic description from hand-drawn sketch maps. The use of force histograms combined with fuzzy set theory has been shown to be able to extract directional as well as topological information about spatial objects with relative ease (Matsakis and Nikitenko, 2005).

In addition to the forgoing applications of fuzzy sets to spatial analysis, there have been reformulations of nonfuzzy techniques for spatial analysis. One of the early reformulations was that of fuzzy kriging which can be used for analysis or interpolation of spatial data (Bardossy *et al*., 1989). The formulation of fuzzy kappa (Hagen-Zanker *et al*., 2005) has already been mentioned and is an important step towards using fuzzy techniques for analyzing fuzzy spatial data. One of the most widely used reformulations, or extensions, is the fuzzy c-means (also known as k-means) algorithm that is a fuzzification of the nonfuzzy c-means clustering algorithm (Bezdek *et al*., 1984;Wilson and Burrough, 1999).

As simulation models have become more commonly used to address spatially explicit problems, fuzzy sets has been used in a number of ways to address the uncertainties inherent at various levels of such models. Not only is there uncertainty in the spatial data that may affect the model, but uncertainty surrounding the precise value of a parameter can affect the outcome of the modeling exercise. Wu (1998) and Bone *et al*. (2006) are examples of using fuzzy sets in cellular automata for urban and ecological modeling respectively. Bossomaier *et al*. (2005) and Robinson and Graniero (2005) use fuzzy sets in individual-based modeling of housing transactions and animal dispersal movements respectively. Not only have fuzzy sets improved the ability of simulation models

to formally accommodate uncertainty, it has also been shown to enable model self-evaluation thus avoiding semantic errors in complex process models that unknowingly compromise the integrity of an analysis (Mackay and Robinson, 2000).

With the advent of GIS the representation and query of uncertain spatial to support spatial analysis has developed substantially. Cross and Firat (2000) discuss the issues involved in construction of fuzzy spatial objects with specific reference to GIS. Morris (2003) describes a fuzzy object-oriented framework to model spatial objects with uncertain boundaries. Another object-based effort is that of Bordogna *et al*. (2006) who developed a fuzzy object-based data model as a tool for supporting spatial analysis. It is based on the management of a linguistic granule.

Verstraete *et al*. (2005) presented detailed techniques for modeling fuzzy spatial information represented as triangular irregular networks (TINs) and raster (grid) layers. They show how processing, as well representation, can be carried out using fuzzy set theory to represent the uncertainty in spatial data. One of the significant aspects of this work is its presentation of the use of type-2 fuzzy sets. In other words, it detailed how to formally represent and process uncertainty not just about the spatial data, but also uncertainty about the fuzzy membership functions themselves.

12.3. ASSIGNING FUZZY MEMBERSHIPS

Crucial to any spatial analysis using fuzzy sets is the assignment of membership. With regard to fuzzy membership, it is important to realize that fuzzy memberships have special characteristics. First, although fuzzy membership values are typically normalized to fall between 0.0 and 1.0, they are not probabilities. Probabilities and fuzzy membership values measure very different things. For example, one of the most important properties of probability is additivity. There is no such inherent restriction on fuzzy memberships. In fact, the sum of membership values is interpreted as fuzzy cardinality (i.e., the size of a fuzzy set). Second, membership values are not a simple quantitative variable of the interval level. It is because the end points (i.e., 0 and 1) have more meaning than just being artifacts of the membership function. Verkuilen (2005) suggests it is really a generalization of the case of dichotomous dummy variables that are often used to represent ordinary crisp sets.

In the spatial analytic and GIS literature it is common to refer to either the Semantic Import (SI) or Similarity Relation (SR) model (Burrough and McDonnell, 1998; Robinson, 1988). However, it may be more useful to consider that fuzzy memberships are usually a function of a direct assignment (DA), indirect assignment (ID), or an assignment by transformation (AT) methodology (Verkuilen, 2005).

12.3.1. Direct assignment

Studies where membership functions are provided directly by an 'expert' is characteristic of the direct assignment (DA) method. It is also common in the DA method of assignment to make use of standard membership functions such as the triangular, trapezoidal, bell, and others (Robinson, 2003). For example, in consultation with experts, DeGenst *et al*. (2001) made use of a standard curve to describe a basic spatial relation in their study of squirrel dispersal. Often, as in Braimoh *et al*. (2004), and Zeng and Zhou (2001), the choice of membership function is based on 'the literature', 'common-sense', and/or expert opinion. Sometimes these

membership functions are available as part of a geographic information system so that experts can specify them directly in an automated geospatial environment (Yanar and Akyurek, 2006). Neverthless, they are still directly assigned by an expert. It should be noted that this approach is sometimes criticized because of these deficiencies:

1 Interpretation is difficult because rarely is there anything tangible underlying the number.

2 It may be too difficult for the expert(s) to do reliably, especially if they are not well-versed in fuzzy set theory.

3 Can be biased. In particular, subjects may systematically be biased towards the end points (Thole *et al.*, 1979).

4 Difficulty in combining assignments from multiple experts. This is especially difficult when the assignments are at extreme variance from one another (Verkuilen, 2005).

Despite these deficiencies, direct assignment remains a commonly used strategy for assigning membership values.

12.3.2. Indirect assignment

Indirect assignment elicits responses of some kind from experts and applies a model to the judgments to generate membership values. There have been a number of approaches used to formalize the process of generating fuzzy set memberships from expert knowledge. One of the simpler approaches showed how an intelligent, interactive question/answer system could be used to generate fuzzy representations of a spatial relation such as 'near'. In this approach the expert need only provide a yes/no answer to a question posed by the software. From those crisp answers the system generates a fuzzy representation of a spatial concept (Robinson, 2000). This approach may be useful to generalize for obtaining fuzzy representations of individual concepts; it is not suitable for use in studies where more complex expert knowledge representations are required.

One of the reasons indirect assignment is less often used is the difficulty of the knowledge elicitation process. Zhu (1999) used personal construct theory to formulate a rigorous methodology for eliciting expert knowledge about soils. Part of the process included the expert interacting with a graphical user interface (GUI) to assist in formalizing the relations. The result of this intensive knowledge elicitation process was used to populate a fuzzy soil similarity model. This is one of the rare studies in the geographical literature where knowledge consistency and validation were explicitly incorporated into the knowledge elicitation process. Although the process is rigorous and thorough, the interviews with the expert that are essential to the process can be very tense and often frustrating for the expert as well as the knowledge engineer. Hence, it can be difficult to secure an expert's cooperation. This is perhaps why there are so few studies in the spatial analytic literature where a rigorous indirect assignment process is followed.

Paired comparisons have been used in conjunction with fuzzy sets and spatial analysis, but generally not in the construction of membership functions/values themselves. For example, Charnpratheep *et al*. (1997) used paired-comparison analytic hierarchy process (AHP) methodology to arrive at weights that were subsequently used in a convex combination model of fuzzy aggregation. However, their membership functions were by direct assignment. In another instance, Kuo *et al*. (2003) used a fuzzy AHP methodology that made use of a questionnaire to acquire data on store location

decisions from 16 business 'experts.' The results of this questionnaire exercise provided enough information to estimate the weight assigned to each factor (e.g., the competition dimension received the highest single weight of 0.1922). They show that weights provided by fuzzy AHP can be applied as criteria for selecting important factors to subsequently be used in an artificial neural network location analysis. These works are suggestive of a linkage to discrete choice modeling (e.g., Fotheringham, 1988; Train, 2003). Some work in the transportation field has explored the use of fuzzy sets in modeling route choice (Vythoulkas and Koutsopoulos, 2003). Since preferences play an important role in discrete choice modeling, Ridwan (2004) introduced a model of route choice based on fuzzy preference relations. The elements the fuzzy relations were specified as fuzzy pairwise comparisons between alternative routes. Since the use of logit models are commonly used to estimate the probability of alternatives being chosen, it is interesting to note that Henn (2000) presents a fuzzy formulation that suggests the logit model is a special case of his fuzzy based model when the similarity measure has a given shape.

Questionnaires have been reportedly used in some studies as an instrument for constructing fuzzy memberships. Although details are not given, Lin *et al*. (2006) describe a process using results of a questionnaire survey to construct a fuzzy rule base. They were able then to make some tentative statements about changes in activity centers in relation to a subway line. Similarly, Fritz *et al*. (2000) used a web-based questionnaire where distances specified by respondents were used to construct fuzzy sets for defining the concepts near, medium and far for visible features and close and far away for nonvisible features. They then detailed a methodology that combined the resulting fuzzy rules to aid in mapping of wild lands. Although, respondents detailed what distances represented concepts like 'near,' the use of a default triangular membership function meant that the actual membership function was not obtained directly from respondent data. Nevertheless, it does represent a more formalized approach for proceeding from questionnaire responses to construction of a fuzzy set or rule base.

12.3.3. Assignment by transformation

In this approach a numerical variable is taken and mapped into membership values by some transformation. There are many different approaches that assign fuzzy membership using some version of assignment by transformation. In this section many of the approaches used to address problems in spatial analysis are briefly discussed.

Among the more typical approaches to assignment is the use of a fuzzy clustering algorithm. Perhaps the most commonly used method across the spatial sciences for assigning membership is based on the fuzzy *c*-means algorithm originally developed by Dunn (1973) later generalized by Bezdek (1974, 1981). It is also known as the fuzzy *k*-means (FKM) or fuzzy ISODATA algorithm. It is derived to minimize an objective function with respect to the membership functions and centroids of *c* clusters. Hence it is useful for clustering multivariate data into a finite number of fuzzy sets (Brown, 1998; Cheng *et al*., 2002; Irvin *et al*., 1997; McBratney and Odeh, 1997; Stefanakis *et al*., 1999). In spatial analytic studies each spatial object would be classified as a member of all classes but to varying degrees. Although used in numerous studies since the algorithm was published (Bezdek *et al*., 1984), it continues to figure prominently in

applications in physical geography (Bragato, 2004; Burrough *et al*., 2001; Scull *et al*., 2003). In addition, it has been used to address spatially explicit problems in fields as diverse as wildlife ecology (Schaefer *et al*., 2001), marketing (Wanek, 2003) and urban geography (Taylor and Derudder, 2004).

Since the objective function does not take into consideration spatial dependence between observations, 'noisy' spatial data can adversely affect the performance of the algorithm. Few attempts to incorporate spatial information in an FCM algorithm have been published outside the image analysis community. Liew *et al*. (2000) presented a modification of the FCM whereby the normed distance computed at each pixel within an image was replace with the weighted sum of distances from within a neighborhood of the pixel. Pham (2001) followed with a more general solution that uses penalty functions to constrain the membership value of a class to be negatively correlated with the membership values of the other classes at neighboring pixels. Both approaches produced promising results. It remains to be seen if, or when, these adaptations of FCM will develop and be applied outside the image analysis community.

Another problem is that the number of classes needs to be specified *a priori*. In their extension of the FCM to the spatio-temporal domain, Liu and George (2005) address the number of clusters problem using the Xie–Beni validity index to develop a stopping condition. Given a starting number of clusters, their technique will successively merge clusters until a stopping condition based on the Xie–Beni validity index is met. They illustrate its use on spatio-temporal meteorological data where it is able to detect interesting climatic phenomena.

Another approach that has been used to map from data to a fuzzy membership is characterized by the application of artificial neural network (ANN) methods adjusted so that the output is a fuzzy membership value. Note that this differs from the use of ANN by Kuo *et al*. (2003) in that they used fuzzy AHP to develop the weights that were used by ANN to produce nonfuzzy results. Here ANN is considered as a method that is used to directly produce a fuzzy classification (Foody and Boyd, 1999) or to extract a fuzzy rule base from data (Zheng and Kainz, 1999). In either case, ANNs are composed of a set of simple processing units, or nodes, that are interconnected by some predefined architecture which can be trained. The processing nodes are generally arranged in a layered architecture where the first layer is the input, or fuzzification, layer where there is one node per input channel (i.e., input variable). The second, or implication, layer(s) is comprised of a number of processing units. These processing nodes do most of the thinking of the ANN. The third layer is the output, or defuzzification, layer. In general, there is one output node associated with each class to be output. Each node in a layer is connected to every node in an adjacent layer by a weighted link. The weights are typically set randomly and iteratively adjusted during a training phase during which the ANN attempts to generate a model capable of correctly assigning class membership.

Related to ANN is the adaptive neuro-fuzzy inference system (ANFIS) (Jang, 1993). Using a given input/output data set the objective is to construct a fuzzy inference system whose membership functions best suit the data set. Using a back-propagation algorithm or a least-squares method, the membership parameters are tuned in a training exercise similar to ANN (The Math Works Inc., 2002). ANFIS has been used for map revision (Teng and Fairbairn, 2002) and land cover classification (Peschel, 2002).

These neural network approaches have advantages and disadvantages. The advantages include an ability to learn from training data and they can handle noisy, incomplete data. Once trained, an ANN can respond to a new set of data instantly. However, they can take a long time to train, especially since training is still largely by trial and error complicated by the fact that incomplete training data can cause the network to provide incorrect results. Perhaps the most important disadvantage is that it is difficult to explain the specific reasoning leading to the output product. Hence it can be a kind of black-box approach.

The presence, or absence, of an association, interaction or interconnectedness between elements of two or more sets is represented by a crisp relation. Rather than presence/absence of association, degrees of association can be represented by membership grades in a fuzzy relation in much the same way as degrees of set membership are represented in a fuzzy set. Thus, the classical notion of relation can be generalized into matter of degree as a fuzzy relation. Fuzzy relations have been used to formally represent fuzzy regions and their relationships (Zhan and Lin, 2003). In addition, Kahraman *et al*. (2003) present an example of using fuzzy relations in a model of group decision making for the facility location selection problem.

Statistical data analysis has been suggested as another way to choose fuzzy membership functions and form fuzzy rules (Hanna *et al*., 2002). However, it has not been used widely in spatial analysis. An example of its application to a spatially explicit problem is illustrated by the problem of estimating parameters to use in a regional ecohydrological simulation model. Mackay *et al*. (2003) use a two stage methodology where in the first stage many simulations are run in which parameters affecting stomatal conductance are assigned values using Monte Carlo sampling. Then each simulation result is evaluated by regressing simulated evaporative fraction from RHESSys and surface temperature from thermal remote sensing data. For each regression, the coefficient of determination (R^2) is calculated and used as a fuzzy measure of the goodness-of-fit for its respective simulation result. Hence the fuzzy set is composed of the set of R^2 measures for all simulations, to which an information-theoretic tool based on ordered possibility distributions is applied to form a restricted set in which only 'good' simulations retained. A restricted set is used as an ensemble solution in the second stage of parameter estimation. Note that a separate ensemble solution is produced for each hillslope (Mackay *et al*., 2003).

12.4. COMBINING MEMBERSHIPS

A common requirement of fuzzy spatial analysis is the combination of several fuzzy sets in a desirable manner to produce a single fuzzy set (Klir et al., 1997). This combination is often accomplished using aggregation operators. In fuzzy set theory there are many aggregation operators from which to choose with the most common being the min (intersection) and max (union) operators (Robinson, 2003). The choice of operator depends on the nature of the underlying decision model. For example, in their fuzzy-base cellular automata model of insect infestation, Bone *et al*. (2006) used a compensatory operator rather than the noncompensatory operators (max or min) because the compensatory aggregation operator allows for the influence of each set to contribute to the final result.

The basic aggregation operators have been further developed using various

weighting schemes. Both the convex combination and a modified ordered weighted averaging operator (OWA) have been used in various studies (Charnpratheep *et al*., 1997; Oberthur *et al*., 2000; Zeng and Zhou, 2001). However, use of the weighted aggregation models has highlighted the subjectivity inherent in the weighting scheme, hence care should be taken when formulating the weighting scheme as small differences in subjective weights can lead to large variations in the results (Jiang and Eastman, 2000).

Another common method is to use fuzzy rules that have the general structure of the form:

IF(*antecedent*)**THEN**(*consequent*).

To evaluate the rule base and arrive at an answer requires the application of an inference, or implication, method. One of the most common inference methods is known as Mamdani-type inference. Whether named, or not, it is often the one used for spatial analytic studies because it supports outputs as fuzzy sets. The use of fuzzy rule bases in spatial analysis include applications for the conflation of vector maps (Cobb *et al*., 1998), real estate evaluation (Zeng and Zhou, 2001), land fill location (Charnpratheep *et al*., 1997), and prediction of weed infestation (Chiou and Yu, 2001). The alternative Takagi–Sugeno type inference model tends to produce a single, crisp value as output rather than a fuzzy set. This is why many applications have avoided its use (e.g., Power *et al*., 2001). However, such a characteristic may be useful for spatial interpolation purposes. For example, a Takagi–Sugeno rule base has been used in the spatial interpolation of solar radiation (Botia *et al*., 2001).

12.5. CHALLENGES AND RESEARCH ISSUES

Although this chapter has detailed a varied set of accomplishments using fuzzy sets in fields allied with spatial analysis, there remain many significant challenges and research issues. All the areas of accomplishment noted above remain open to further studies to refine research issues in their work.

Even though there has been substantial research on the representation and processing of spatial data (Bordogna *et al*., 2006; Verstraete *et al*., 2005), especially in a GIS context, the specification of fuzzy membership remains a challenge. Although a variety of methods for assigning fuzzy memberships have been presented, there remains much to be done in formalizing the process. In the field of spatial analysis, there is still a need for the development of methodologies for the acquisition of fuzzy memberships from experts. Perhaps an even more pressing challenge is that posed by formalizing the acquisition of fuzzy memberships directly from spatial data in such a way as to have meaning in the context of the problem domain.

Another chapter of this book deals with the topic of spatial autocorrelation. In spatial analysis spatial autocorrelation is usually presented as a special topic in the statistical analysis of spatial data. With a few exceptions, it is a topic that has for the most part been neglected in many applications of fuzzy sets to spatial analysis. For example, in image analysis it is known to affect the fuzzy clustering results, hence its incorporation in some fuzzy clustering algorithms. However, there are few, if any, investigations of the degree to which spatial autocorrelation affects fuzzy-based results in areas such as soil science, geodemographics, landscape ecology, etc.

An issue that has not been fully developed in spatial analysis is the linkage between

fuzzy sets and mainstream spatial analysis. There are examples of the fuzzification of mainstream methods such as in the case of kriging and the kappa statistic. However, other aspects of fuzzy statistics have yet to be explored in depth for the analysis of spatial data. For example, although regression analysis is often used in spatial analyses the use of fuzzy regression techniques is virtually unheard of in spatial analysis even though fuzzy regression techniques address the case where the relations of the variables are subject to fuzziness or where the variables themselves are fuzzy (Taheri, 2003).

Many efforts in spatial analysis are concerned with the testing of hypotheses. Mainstream methods rely upon classical statistics to determine whether a hypothesis should be rejected. Little investigation of fuzzy hypothesis testing has been done in the context of spatial analysis. However, as Smithson (2005) points out, fuzzy sets and statistics work better together. There are a few cases where this is demonstrated mostly in relation to the process of assigning membership values (Ahn *et al*., 1999; Brown, 1998; Mackay *et al*., 2003), not in the explicit testing of hypotheses.

There are several broad issues that will face researchers attempting to use fuzzy sets in spatial analysis. Perhaps, the most fundamental issue is when, or when not, to use fuzzy-based analysis. This is not easily answered and demands considerable knowledge of both the problem at hand as well as both mainstream methods as well as fuzzy-based methods. However, fuzzy-based approaches are showing great promise, yet are still not as widely known, or understood, as many of the mainstream approaches detailed in other chapters of this book. Another issue is whether or not a fuzzy-based spatial analysis should be evaluated against nonfuzzy-based techniques or are they now developed enough to stand on their own. This implies that they are competing paradigms when they may be more properly viewed as complementary paradigms of analysis.

REFERENCES

Ahamed, T.R.N., Rao, G.K. and Murthy, J.S.R. (2000). Fuzzy class membership approach to soil erosion modelling. *Agricultural Systems*, **63**: 97–110.

Ahlqvist, O. (2005). Using uncertain conceptual spaces to translate between land cover categories. *International Journal of Geographical Information Science*, **19**: 831–857.

Ahn, C.-W., Baumgardner M.F. and Biehl L.L. (1999). Delineation of soil variability using geostatistics and fuzzy clustering analyses of hyperspectral data. *Soil Sci. Soc. Am. J.*, **63**: 142–150.

Anile, M.A., Furno, P., Gallo, G. and Massolo, A. (2003). A fuzzy approach to visibility maps creation over digital terrains. *Fuzzy Sets and Systems*, **135**: 63–80.

Arvanitis, L.G., Ramachandran, B., Brackett, D.P., Abd-El Rasol, H. and Xu, X.S. (2000). Multire-source inventories incorporating GIS, GPS and database management systems: a conceptual model. *Computers and Electronics in Agriculture*, **28**: 89–100.

Bardossy, A., Bogardi I. and Kelly W.E. (1989). Geostatistics utilizing imprecise (fuzzy) information. *Fuzzy Sets and Systems*, **31**: 311–328.

Bezdek, J.C. (1974). Cluster validity with fuzzy sets. *Journal of Cybernetics*, **3**: 58–73.

Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press.

Bezdek, J.C., Ehrlich, R. and Full, W. (1984). FCM: the fuzzy *c*-means clustering algorithm. *Computers and Geosciences*, **10**: 191–203.

Bone, C., Dragicevic, S. and Roberts, A. (2006). A fuzzy-constrained cellular automata model of forest insect infestations. *Ecological Modelling*, **192**: 107–125.

Bordogna, G., Chiesa, S. and Geneletti, D. (2006). Linguistic modelling of imperfect spatial information as a basis for simplifying spatial analysis. *Information Sciences*, **176**: 366–389.

Bossomaier, T., Amri, S. and Thompson, J. (2005). Agent-based modelling of house price evolution. In: *Proceedings of CABM-HEMA-SMAGET 2005 Joint Conference on Multi-Agent Modelling for Environmental Management*, Bourg StMaurice-Les Arc, France.

Botia, J.A., Gomez-Skarmeta, A.F., Valdes, M., and Padilla, A. (2001). Fuzzy and hybrid methods applied to GIS interpolation. In: *The 10th IEEE International Conference on Fuzzy Systems*, 453–456, Melbourne, Australia.

Bragato, G. (2004). Fuzzy continuous classification and spatial interpolation in conventional soil survey for soil mapping of the lower Piave plain. *Geoderma*, **118**: 1–16.

Braimoh, A.K., Vlek, P.L., Stein, A. (2004). Land evaluation for maize based on fuzzy set and interpolation. *Environmental Management*, **33**: 226–238.

Brown, D.G. (1998). Classification and boundary vagueness in mapping presettlement forest types. *International Journal of Geographical Information Science*, **12**: 105–129.

Brown, D.G. (1998). Mapping historical forest types in Baraga County Michigan, USA as fuzzy sets. *Plant Ecology*, **134**: 97–118.

Buckley, J.J. and Eslami, E. (2002). *An Introduction to Fuzzy Logic and Fuzzy Sets*. New York: Physica-Verlag.

Burrough, P.A., McDonnell, R.A. (1998). *Principles of Geographical Information Systems*. New York: Oxford University Press.

Burrough, P.A., Wilson, J.P., van Gaans Pauline, F.M. and Hansen, A.J. (2001). Fuzzy *k*-means classification of topo-climatic data as an aid to forest mapping in the Greater Yellowstone area, USA. *Landscape Ecology*, **16**: 523–546.

Charnpratheep, K., Zhou, Q. and Garner, B. (1997). Preliminary landfill site screening using fuzzy geographical information systems. *Waste Management & Research*, **15**: 197–215.

Cheng, T., Molenaar, M. and Lin, H. (2002). Formalizing fuzzy objects from uncertain classification results. *International Journal of Geographical Information Science*, **15**: 27–42.

Chiou, A. and Yu, X. (2001). Prediction of Parthenium weed infestation using fuzzy logic applied to geographic information system (GIS) spatial image. In: *The 10th IEEE International Conference on Fuzzy Systems*, pp. 1363–1366, Melbourne, Australia.

Cobb, M.A., Chung, M.J., Foley III, H., Petry, F.E. and Shaw, K.B. (1998). A rule-based approach for the conflation of attributed vector data. *Geoinformatica*, **2**: 7–35.

Cross, V., Firat, A. (2000). Fuzzy objects for geographical information systems. *Fuzzy Sets and Systems*, **113**: 19–36.

DeGenst, A., Canters, F. and Gulink, H. (2001). Uncertainty modeling in buffer operations applied to connectivity analysis. *Transactions in GIS*, **5**: 305–326.

Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, **3**: 32–57.

Fisher, P., Wood, J. and Cheng, T. (2005). Fuzziness and ambiguity in multi-scale analysis of landscape morphometry. In: Petry, F.E., Robinson, V.B. and Cobb, M.A. (eds.), *Fuzzy Modeling with Spatial Information for Geographic Problems*. pp. 207–232. Heidelberg: Springer.

Foody, G.M. (1999). The continuum of classification fuzziness in thematic mapping. *Photogrammetric Engineering and Remote Sensing*, **65**: 443–451.

Foody, G.M. and Boyd, D.S. (1999). Fuzzy mapping of tropical land cover along an environmental gradient from remotely sensed data with an artificial neural network. *Journal of Geographical Systems*, **1**: 23–35.

Fotheringham, A.S. (1988). Consumer store choice and choice set definition. *Marketing Science*, **7**: 299–310.

Fritz, S. and See, L. (2005). Comparison of land cover maps using fuzzy agreement. *International Journal of Geographical Information Science*, **19**: 787–807.

Fritz, S., Carver, S. and See, L. (2000). New GIS approaches to wild land mapping in Europe. In: *Wilderness Science in a Time of Change Conference – Volume 2: Wilderness Within the Context of Larger Systems*, Missoula, MT, pp. 120–127.

Gale, S. (1972). Inexactness, fuzzy sets, and the foundations of behavioral geography. *Geographical Analysis*, **4**: 337–349.

Gedeon, T.D., Wong, K.W., Wong, P. and Huang, Y. (2003). Spatial interpolation using fuzzy reasoning. *Transactions in GIS*, **7**: 55–66.

Graniero, P.A. and Robinson, V.B. (2003). A real-time adaptive sampling method for field mapping in

patchy, heterogeneous environments. *Transactions in GIS*, **7**: 31–54.

Guneralp, B., Mendoza, G., Gertner, G. and Anderson, A. (2003). Spatial simulation and fuzzy threshold analyses for allocating restoration areas. *Transactions in GIS*, **7**: 325–343.

Hagen, A. (2003). Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science*, **17**: 235–249.

Hagen-Zanker, A., Straatman, B., Uljee, I. (2005). Further developments of a fuzzy set map comparison approach. *International Journal of Geographical Information Science*, **19**: 769–785.

Hanna, A.S., Lotfallah, W.B. and Lee, M.J. (2002). Statistical-fuzzy approach to quantify cumulative impact of change orders. *Journal of Computing in Civil Engineering*, **16**: 252–258.

Heikkila, E.J., Shen, T.-Y., Yang, K.-Z. (2003). Fuzzy urban sets: theory and application to desakota regions in China. *Environment and Planning B: Planning and Design*, **30**: 239–254.

Henn, V. (2000). Fuzzy route choice model for traffic assignment. *Fuzzy Sets and Systems*, **116**: 77–101.

Hwang. S., Thill. J.-C. (2005). Modeling localities with fuzzy sets and GIS. In: Petry, F.E., Robinson, V.B. and Cobb, M.A. (eds.), *Fuzzy Modeling with Spatial Information for Geographic Problems*. pp. 71–104. Heidelberg: Springer.

Irvin, B.J., Ventura, S.J. and Slater, B.K. (1997). Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Winsconsin. *Geoderma*, **77**: 137–154.

Jang, J.-S.R. (1993). ANFIS: Adaptive-network-based fuzzy inference systems. *IEEE Trans. Systems, Man & Cybernetics*, **23**: 665–685.

Jiang, H. and Eastman, J.R. (2000). Application of fuzzy measures in multi-criteria evaluation in GIS. *International Journal of Geographical Information Science*, **14**: 173–184.

Kahraman, C., Ruan, D. and Dogan, I. (2003). Fuzzy group decision-making for facility location selection. *Information Sciences*, **157**: 135–153.

Katz, A., Vom, H.M. and Mahoney, J. (2005). Explaining the great reversal in Spanish America: fuzzy set analysis versus regression analysis. *Sociological Methods and Research*, **33**: 539–573.

Klir, G.J., Ute, S.C., Yuan, B. (1997). *Fuzzy Set Theory: Foundations and Applications*. Upper Saddle River, NJ: Prentice Hall.

Kosko, B. (1992). *Neural Networks and Fuzzy Systems*. Englewood Cliffs, NJ: Prentice-Hall.

Kuo, R.J., Chi, S.C. and Kao, S.S. (2003). A decision support system for selecting convenience store location through integration of fuzzy AHP and artificial neural network. *Computers in Industry*, **47**: 199–214.

Leung, Y. (1983). Fuzzy sets approach to spatial analysis and planning, a nontechnical evaluation. *Geografiska Annaler, Series B, Human Geography*, **65**: 65–75.

Liew, A.W.C., Leung, S.H. and Lau, W.H. (2000). Fuzzy image clustering incorporating spatial continuity. *IEE Proc Vision, Image, and Signal Processing*, **147**: 185–192.

Lin, J.-J., Feng, C.-M., Hu, Y.-Y. (2006). Shifts in activity centers along the corridor of the Blue subway line in Taipei. *Journal of Urban Planning and Development*, **132**: 22–28.

Liu, Z. and George, R. (2005). Mining weather data using fuzzy cluster analysis. In: Petry, F.E., Robinson, V.B. and Cobb, M.A. (eds.), *Fuzzy Modeling with Spatial Information for Geographic Problems*, pp. 105–119. Heidelberg: Springer.

Lodwick, W.A. and Santos, J. (2003). Constructing consistent fuzzy surfaces from fuzzy data. *Fuzzy Sets and Systems*, **135**: 259–277.

Lundberg, C.G. (1982). Modeling constraints and anticipation: linguistic variables, foresight-hindsight and relative alternative attractiveness. *Geographical Analysis*, **14**: 347–355.

Mackay, D.S. and Robinson, V.B. (2000). A multiple criteria decision support system for testing integrated environmental models. *Fuzzy Sets and Systems*, **113**: 53–67.

Mackay, D.S., Samanta S., Ahl, D.E., Ewers, B.E., Gower, S.T., Burrows, S.N. (2003). Automated parameterization of land surface process models using fuzzy logic. *Transactions in GIS*, **7**: 139–153.

MacMillan, R.A., Martin, T.C., Earle, T.J. and McNabb, D.H. (2003). Automated analysis and classification of landforms using high-resolution digital elevation data: applications and issues. *Canadian Journal of Remote Sensing*, **29**: 592–606.

MacMillan, R.A., Pettapiece, W.W., Nolan, S.C. and Goddard, T.W. (2000). A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules, and fuzzy logic. *Fuzzy Sets and Systems*, **113**: 81–109.

Matsakis, P. and Nikitenko, D. (2005). Combined extraction of directional and topological relationship information from 2D concave objects. In: Petry, F.E., Robinson, V.B. and Cobb, M.A. (eds.), *Fuzzy Modeling with Spatial Information for Geographic Problems*. pp. 143–158. Berlin: Springer.

McBratney, A.B. and Odeh, I.O.A. (1997). Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma*, **77**: 85–113.

Morris, A. (2003). A framework for modeling uncertainty in spatial databases. *Transactions in GIS*, **7**: 83–103.

Morris, A. and Jankowski, P. (2005). Spatial decision making using fuzzy GIS. In: Cobb, M.A., Petry, F. and Robinson, V.B. (eds.), *Fuzzy Modeling with Spatial Information for Geographic Problems*. pp. 275–298. Heidelberg: Springer.

Oberthur, T., Dobermann, A. and Aylward, M. (2000). Using auxiliary information to adjust fuzzy membership functions for improved mapping of soil qualities. *International Journal of Geographical Information Science*, **14**: 431–454.

Odeh, I.O.A., McBratney, A.B. and Chittleborough, D.J. (1990). Design of optimal sample spacings for mapping soil using fuzzy *k*-means and regionalized variable theory. *Geoderma*, **47**: 93–112.

Paez, D., Bishop, I.D. and Williamson, I.P. (2006). DISCUSS: a soft computing approach to spatial disaggregation in economic evaluation of public policies. *Transactions in GIS*, **10**: 265–278.

Peschel, J.M. (2002). Creating land cover input datasets from the SWAT 2000 model using remotely sensed data. Texas A&M University, *http://ceprofs.tamu.edu/folivera/TxAgGIS/Spring2002/Peschel/peschel.htm*, visited on April 14, 2006.

Pham, D.L. (2001). Spatial models for fuzzy clustering. *Computer Vision and Image Understanding*, **84**: 285–297.

Pipkin, J.S. (1978). Fuzzy sets and spatial choice. *Annals of Association of American Geographers*, **68**: 196–204.

Power, C., Simms, A. and White, R. (2001). Hierarchical fuzzy pattern matching for the regional comparison of land use maps. *International Journal of Geographical Information Science*, 15: 77–100.

Ragin, C.C. and Pennings, P. (2005). Fuzzy sets and social research. *Sociological Methods and Research*, **33**: 423–430.

Ridwan, M. (2004). Fuzzy preference based traffic assignment problem. *Transportation Research Part C*, **12**: 209–233.

Robinove C.J. (1989). Principles of logic and the use of digital geographic information systems. In: Ripple, W.J. (ed.), *Fundamentals of GIS: A Compendium*, Washington, D.C.: American Society for Photogrammetry and Remote Sensing. pp. 61–79.

Robinson, V.B. and Strahler, A.H. (1984). Issues in designing geographic information systems under conditions of inexactness. In: *Proceedings of 10th International Symposium on Machine Processing of Remotely Sensed Data*, pp. 179–188, Terre Haute, IN.

Robinson, V.B. (1988). Some implications of fuzzy set theory applied to geographic databases. *Computers, Environment, and Urban Systems*, **12**: 89–97.

Robinson, V.B. (2000). Individual and multipersonal fuzzy spatial relations acquired using human-machine interaction. *Fuzzy Sets and Systems*, **113**: 133–145.

Robinson, V.B. (2003). A perspective on the fundamentals of fuzzy sets and their use in geographic information systems. *Transactions in GIS*, **7**: 3–30.

Robinson, V.B. and Graniero, P.A. (2005). Spatially explicit individual-based ecological modeling with mobile fuzzy agents. In: Petry, F.E., Robinson, V.B. and Cobb, M.A. (eds.), *Fuzzy Modeling with Spatial Information for Geographic Problems*. pp. 299–334. Heidelberg: Springer.

Schaefer, J.A., Veitch, A.M., Harrington, F.H., Brown, W.K., Theberge, J.B. and Luttich, S.N. (2001). Fuzzy structure and spatial dynamics of a declining woodland caribou population. *Oecologia*, **126**: 507–514.

Schaefer, J.A. and Willson, C.C. (2002). A fuzzy structure of populations. *Canadian Journal of Zoology*, **80**: 2235–2241.

Scull, P., Franklin, J., Chadwick, O.A. and McArthur, D. (2003). Predictive soil mapping: a review. *Progress in Physical Geography*, **27**: 171–197.

Skubic, M., Blisard, S., Bailey, C., Adams, J.A., Matsakis, P. (2004). Qualitative analysis of sketched route maps: translating a sketch into linguistic descriptions. *IEEE Trans. on Systems, Man, and Cybernetics*, **34**: 1275–1282.

Smithson, M. (2005). Fuzzy set inclusiong: linking fuzzy set methods with mainstream techniques. *Sociological Methods and Research*, **33**: 431–461.

Stefanakis, E., Vazirgiannis, M. and Sellis, T. (1999). Incorporating fuzzy set methodologies in a DBMS repository for the application domain of GIS. *International Journal of Geographical Information Science*, **13**: 657–675.

Taheri, S.M. (2003). Trends in fuzzy statistics. *Austrian Journal of Statistics*, **32**: 239–257.

Taylor, P.J. and Derudder, B. (2004). Porous Europe: european cities in global urban arenas. *Tijdschrift voor Economishe en Sociale Geogra phie*, **95**: 527–538.

Teng, C.H. and Fairbairn, D. (2002). Comparing expert systems and neural fuzzy systems for object recognition in map dataset revision. *International Journal of Remote Sensing*, **23**: 555–567.

The Math Works Inc. (2002). *Fuzzy Logic Toolbox User's Guide*. Natick, MA. USA: The Math Works Inc.

Thole, U., Zimmermann, H.-J. and Zysno, P. (1979). On the suitability of minimum and product operators for the intersection of fuzzy sets. *Fuzzy Sets and Systems*, **2**: 167–180.

Torres, R., Keller, G.R., Kreinovich, V., Longpre, L. and Starks, S.A. (2004). Eliminating duplicates under interval and fuzzy uncertainty: an asymptotically optimal algorithm and its geospatial applications. *Reliable Computing*, **10**: 401–422.

Train, K.E. (2003). *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge University Press.

Verkuilen, J. (2005). Assigning membership in a fuzzy set analysis. *Sociological Methods and Research*, **33**: 462–496.

Verstraete, J., De Tre, G., De Caluwe, R. and Hallez, A. (2005). Field based methods for the modeling of fuzzy spatial data. In: Petry, F.E., Robinson, V.B. and Cobb, M.A. (eds.), *Fuzzy Modeling with Spatial Information for Geographic Problems*. pp. 41–70. Heidelberg: Springer.

Vythoulkas, P.C. and Koutsopoulos, H.N. (2003). Modeling discrete choice behavior using concepts from fuzzy set theory, approximate reasoning and neural networks. *Transportation Research Part C*, **11**: 51–73.

Wanek, D. (2003). Fuzzy spatial analysis techniques in a business GIS environment. In: *European Regional Science Association 2003 Congress*, Jyvaskla, Finland [CD-ROM (paper no. 177)].

Wealands, S.R., Grayson, R.B. and Walker, J.P. (2005). Quantitative comparison of spatial fields for hydrological model assessment – some promising approaches. *Advances in Water Resources*, **28**: 15–32.

Wilson, J.P., Burrough, P.A. (1999). Dynamic modeling, geostatistics, and fuzzy classification: new sneakers for a new geography? *Annals of the Association of American Geographers*, **89**: 736–746.

Witlox, F. and Derudder, B. (2005). Spatial decision-making using fuzzy decision tables: theory, application and limitations. In: Petry, F.E., Robinson, V.B. and Cobb, M.A. (eds.), *Fuzzy Modeling with Spatial Information for Geographic Problems*. pp. 253–275. Berlin: Springer.

Wu, F. (1998). Simulating urban encroachment on rural land with fuzzy-logic-controlled cellular automata in a geographical information system. *Journal of Environmental Management*, **53**: 293–308.

Yanar Tahsin, A. and Akyurek, Z. (2006). The enhancement of the cell-based GIS analyses with fuzzy processing capabilities. *Information Sciences*, **176**: 1067–1085.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, **8**: 338–353.

Zeng, T.Q. and Zhou, Q. (2001). Optimal spatial decision making using GIS: a prototype of a real estate geographical information system (REGIS). *International Journal of Geographical Information Science*, **15**: 307–321.

Zhan, F.B., Lin, H. (2003). Overlay of two simple polygons with indeterminate boundaries. *Transactions in GIS*, **7**: 67–81.

Zheng, D. and Kainz, W. (1999). Fuzzy rule extraction from GIS data with a neural fuzzy system for decision making. In: *Proceedings of the Seventh ACM International Symposium on Advances in Geographic Information Systems*, Kansas City, MO, USA, pp. 79–84.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K. and Simonson, D. (2001). Soil mapping using GIS, expert

knowledge, and fuzzy logic. *Soil Science Society of America Journal*, **65**: 1463–1472.

Zhu, A.-X. (1997). A similarity model for representing soil spatial information. *Geoderma*, **77**: 217–242.

Zhu, A.-X. (1999). A personal construct-based knowledge acquisition process for natural resource mapping. *International Journal of Geographical Information Science*, **13**: 119–141.

Zhu, A.-X. (2004). Personal Communication. Department of Geography, University of Wisconsin.

Zimmermann, H.-J. (2001). *Fuzzy Set Theory and Its Applications*. Boston, MA: Kluwer Academic.

13

Geographically Weighted Regression

A. Stewart Fotheringham

13.1. INTRODUCTION

Spatial data contain locational information as well as attribute information. It is increasingly recognized that most data sets are spatial in that the attribute being measured is typically recorded either at some specific location or as a representation of a general area. It is also increasingly recognized that spatial data exhibit special properties which distinguish them from aspatial data and which necessitate the development of specialized statistical techniques. For instance, spatial data almost invariably exhibit some form of spatial dependence whereby locations in close proximity tend to have more similar attributes than do locations further apart. This tends to invalidate the assumption of the independence of error terms, a fundamental property of classical aspatial statistical inference. Another property of many spatial data sets, perhaps slightly less recognized but becoming increasingly well-known, is that the processes generating the data might exhibit spatial *heterogeneity* or *nonstationarity*. That is, the processes generating observed attributes might vary over space rather than being constant as is assumed in the use of most traditional types of statistical analysis.

Nowhere is this more evident than in the use of what is undoubtedly the most frequently used statistical modelling approach in the analysis of spatial data – that of regression. In a typical linear regression model applied to spatial data we assume a stationary process (often without giving this any thought!). That is, we assume that the same relationships hold throughout the

entire study area we are investigating and that the same stimulus provokes the same response in all parts of the study region. In a linear framework, we can represent these relationships with the following general model:[1]

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_n x_{ni} + \varepsilon_i \tag{13.1}$$

where y_i is the value of the dependent variable observed at location i, $x_{1i}, x_{2i}, \ldots, x_{ni}$ are the values of the independent variables observed at i, $\beta_0, \beta_1, \ldots, \beta_n$ are parameters to be estimated, and ε_i is an error term which is assumed to be normally distributed.

The parameter estimates obtained in the calibration of such a model are constant over space and are obtained from the following estimator:

$$\beta' = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{13.2}$$

That is, for each relationship between y and an x variable, a single parameter is estimated which is assumed to be constant across the study region. Consequently, if there is spatial nonstationarity, the resulting single parameter estimate would then represent an average of the different processes operating over space and we would only get an inkling of this through the residuals of the model. We might map these to determine whether there are any spatial patterns. Or we might compute an autocorrelation statistic for the residuals or we might even try to 'model' the error dependency with various types of spatial regression models. However, spatial dependency in the residuals can result from other processes apart from spatial heterogeneity so examining the residuals is not an ideal solution. It seems much more obvious to allow the parameter estimates in the model to vary over space rather than to calibrate a stationary model and then trying to examine a possible error in the model through the spatial patterning of the residuals. The specification of a model that allows the parameter estimates to vary over space is the essence of geographically weighted regression (GWR).

13.2. GWR MECHANICS

The geographically weighted version of the regression model described in equation (13.1) is:

$$y_i = \beta_{0i} + \beta_{1i} x_{1i} + \beta_{2i} x_{2i} + \cdots \beta_{ni} x_{ni} + \varepsilon_i \tag{13.3}$$

where i refers to a location at which data on y and x are measured and at which local estimates of the parameters are obtained. In this model, the parameter estimates are now local to location i instead of being global constants. The estimator for the parameters is then:

$$\beta'(i) = (\mathbf{X}^T\mathbf{W}(i)\,\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(i)\,\mathbf{Y} \tag{13.4}$$

where $\mathbf{W}(i)$ is a matrix of weights specific to location i such that observations nearer to i are given greater weight than observations further away. The matrix $\mathbf{W}(i)$ has the form:

$$\mathbf{W}(i) = \begin{bmatrix} w_{i1} & 0 & \ldots & \ldots & 0 \\ 0 & w_{i2} & \ldots & \ldots & 0 \\ 0 & 0 & w_{i3} & \ldots & 0 \\ . & . & . & & . \\ 0 & 0 & 0 & \ldots & w_{in} \end{bmatrix} \tag{13.5}$$

where w_{in} is the weight given to data point n for the estimate of the local parameters at location i.

There are many possible weighting functions that could be specified which relate the weighting of an observed value at location j to the distance location j is from the regression point i but they tend to be Gaussian or Gaussian-like, reflecting the nature of many spatial processes. The operation of a typical weighting function is shown in Figure 13.1.

Data points that are located close to the regression point are weighted highly whereas data points that are far from the regression point get a very low weight. Hence, the weighting matrix will change every time the regression point changes. GWR thus produces a model that effectively answers the question 'what do the relationships in my model look like around this location?' The question can be answered for many different locations as we will see below.

Although the exact specification of the weighting function can take many forms, there are two broad categories of weighting functions: *fixed* or *adaptive*. An example of a fixed spatial weighting function is shown in Figure 13.2. In this case, the specified weighting function or kernel is constant across the study area and therefore has the undesirable property that in areas where data points are relatively sparse, the resulting local parameter estimates will have high standard errors attached to them reflecting the added uncertainty in the estimates caused by the relative lack of data.

There are many functions that could be used to represent a fixed spatial weighting function. One is a Gaussian expression:

$$w_{ij} = \exp\left[-\tfrac{1}{2}(d_{ij}/h)^2\right] \qquad (13.6)$$

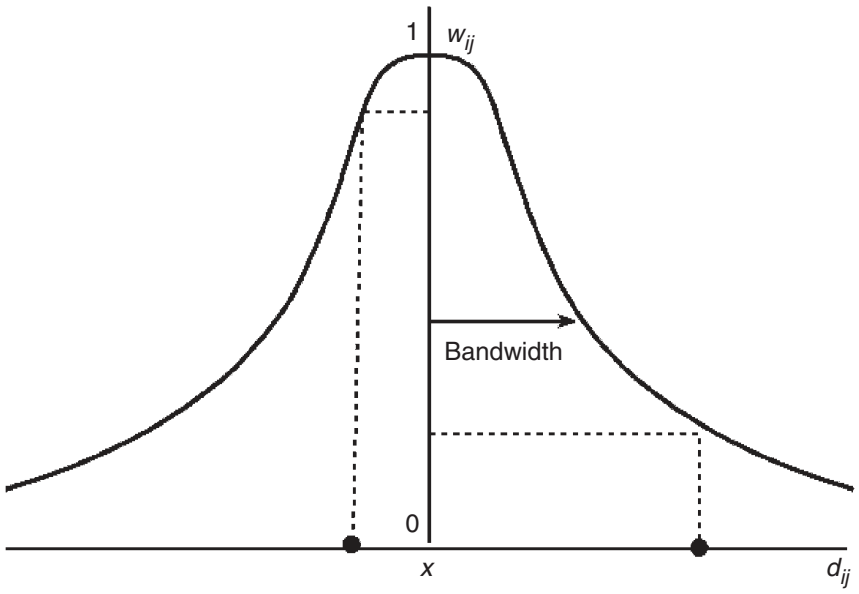


Figure 13.1 A typical spatial weighting function.

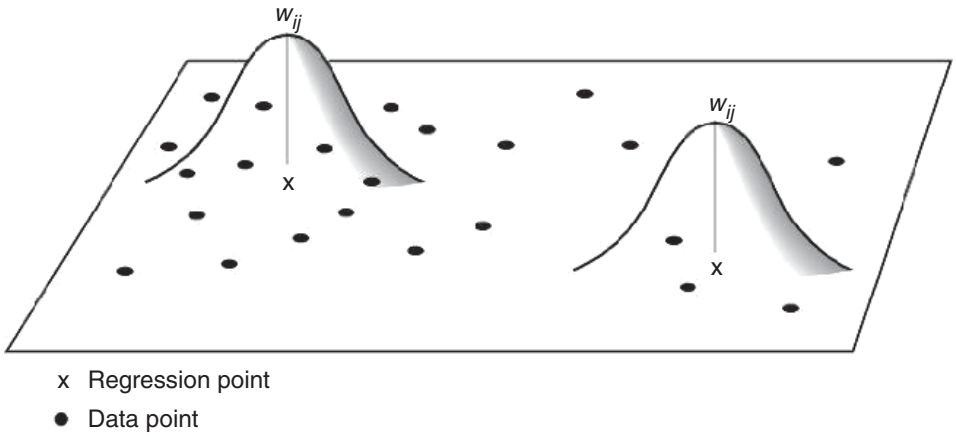


Figure 13.2 A fixed spatial weighting function.

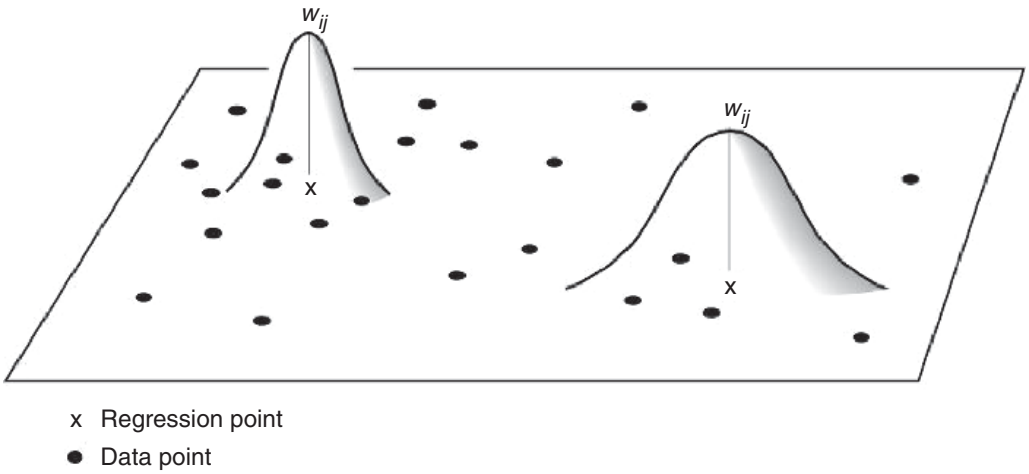


Figure 13.3 A spatially adaptive weighting function.

where d_{ij} is the distance between locations i and j, and h is a parameter often referred to as the bandwidth – as h increases, the gradient of the kernel becomes less steep and more data points are included in the local calibration.

An alternative, and generally preferred, alternative is an adaptive kernel where the spatial extent of the kernel is dictated by the underlying density of data points. In areas where data are plentiful, the kernel is relatively tightly defined around the regression point; in areas where the data are relatively sparse, the kernel has to extend outwards in order to capture more data. The operation of an adaptive kernel is shown in Figure 13.3.

Again, there are several functions that one could use to produce a spatially adaptive weighting function. One, for example, is the following:

$$w_{ij} = [1 - (d_{ij}^2/h^2)]^2 \quad \text{if } j \text{ is one of the } N\text{th nearest neighbours of } i$$
$$= 0 \quad \text{otherwise} \qquad (13.7)$$

where h is the bandwidth and N is a parameter to be estimated.

The results of GWR appear to be relatively insensitive to the choice of weighting function *as long as it is a continuous distance-based function* but whichever weighting function is used, the results will, however, be sensitive to the degree of distance-decay. Therefore an optimal value of either h or N has to be obtained. This can be found by minimizing a cross-validation score (CV) or the Akaike information criterion (AICc) where:

$$\mathrm{CV} = \sum_i [y_i - y \neq i^*(h)]^2 \qquad (13.8)$$

where $y \neq i^*(h)$ is the fitted value of y_i with data from point i omitted from the calibration and:

$$\mathrm{AICc} = \mathrm{Deviance} + 2k[n/(n-k-1)] \qquad (13.9)$$

where n is the number of data points and k is the number of parameters in the model. Lower values of both statistics indicate better model fits.

Optimal bandwidth selection is a trade-off between bias and variance:

- too small a bandwidth leads to a large variance in the local estimates because of the relatively small number of data points used in the local calibration;
- too large a bandwidth leads to large bias in the local estimates because data are drawn from locations further away from the regression point.

As the bandwidth $\to \infty$, the local model will tend to the global model with number of parameters $= k$.

As the bandwidth $\to 0$, the local model 'wraps itself around the data' so the number of parameters $= n$.

The number of parameters in local models therefore ranges between k and n and depends on the bandwidth. This number need not be an integer and is referred to as *the effective number of parameters in the model*.

13.3. GWR OUTPUT

The main output from GWR is a set of location-specific parameter estimates that can be mapped and analysed to provide information on spatial nonstationarity in relationships. However, any diagnostic from regression can be replicated in geographically weighted format so within the GWR framework we can also:

- estimate local standard errors;
- derive local t statistics;
- calculate local goodness-of-fit measures;
- calculate local leverage measures;
- perform tests to assess the significance of the spatial variation in the local parameter estimates; and
- perform tests to determine if the local model performs better than the global one, accounting for differences in degrees of freedom.

13.4. A SIMULATION EXPERIMENT

Consider the following model:

$$y_i = \alpha_i + \beta_{1i}\, x_{1i} + \beta_{2i}\, x_{2i} \qquad (13.10)$$

and data on x_1 and x_2 drawn randomly for 2500 locations on a 50×50 matrix subject to the correlation between x_1 and x_2, $r(x_1, x_2)$, being controlled. In fact, the results of this experiment can be shown to be independent of $r(x_1, x_2)$ so we will ignore this feature of the experiment here.

13.4.1. Experiment 1 (parameters spatially invariant)

In this experiment, we set the three parameters in the model to known, constant values:

$$\alpha_i = 10 \quad \text{for all } i$$

$$\beta_{1i} = 3 \quad \text{for all } i$$

$$\beta_{2i} = -5 \quad \text{for all } i.$$

With everything on the right-hand side of equation (13.10) now known, we can derive a value of y_i at each location and then use these data to calibrate the model both by ordinary least squares regression and by GWR. The results are as follows:

Global model calibrated by OLS

$$\text{Adj. } R^2 = 1.0$$

$$\text{AIC} = -59{,}390$$

$$K = 3$$

$$\alpha(\text{est.}) = 10$$

$$\beta_1(\text{est.}) = 3$$

$$\beta_2(\text{est.}) = -5$$

In this case, where there is no spatial nonstationarity (the parameters are the same everywhere), the global model is clearly appropriate and replicates the y variable perfectly and the estimated parameters are equal to their known values. K represents the number of parameters estimated in the model. The results are not surprising – the processes being modelled are stationary so the global model works well. The question is, how well does the GWR model perform in this situation? The results of the GWR calibration are given below.

Local model calibrated by GWR

$$\text{Adj. } R^2 = 1.0$$

$$\text{AIC} = -59{,}386$$

$$K = 6.5$$

$$N = 2{,}434$$

$$\alpha_i(\text{est.}) = 10 \quad \text{for all } i$$

$$\beta_{1i}(\text{est.}) = 3 \quad \text{for all } i$$

$$\beta_2 i(\text{est.}) = -5 \quad \text{for all } i.$$

Reassuringly, the GWR model attempts to make itself as similar to the global model as possible. N, the number of nearest neighbours used to calibrate each local model, is optimized at 2434 data points out of the 2500. That is, the kernel is trying to become as broad as possible to use all the data on each local calibration. Consequently, the local parameter estimates are the same everywhere and the model replicates the y variable almost perfectly. The AIC values, another goodness of fit measure, are almost identical. Notice that k, here the *effective* number of parameters, is not an integer and

is 6.5. This is the equivalent number of independent parameters used in the model.

These results are useful because they demonstrate that the GWR model is not picking up spurious nonstationarity when the processes are stationary and the global model is appropriate. However, what happens if the processes being examined *are* nonstationary?

13.4.2. Experiment 2 (parameters spatially varying)

Given that the locations of the data points lie on a 50×50 grid, we can use this to assign spatially varying values to each of the three parameters in our model. If the coordinates of a representative grid point are defined as (i, j), we know:

$$0 \leq i \leq 50, \quad 0 \leq j \leq 50$$

so that we can make the parameters functions of i and j. In this case, we chose the following relationships:

$$\alpha_i = 0 + 0.2i + 0.2j \tag{13.11}$$

so that α_i ranges between 0 and 20:

$$\beta_{1i} = -5 + 0.1i + 0.1\,j \tag{13.12}$$

so that β_{1i} ranges between -5 and 5; and:

$$\beta_{2i} = -5 + 0.2i + 0.2\,j \tag{13.13}$$

so that β_{2i} ranges between -5 and 15.

Values of y_i are then obtained as before and the data used to calibrate the model by global regression and by GWR. The results are as follows.

Global model calibrated by OLS

Adj. $R^2 = 0.04$

AIC $= 17{,}046$

$K = 3$

α(est.) $= 10.26$

β_1(est.) $= -0.1$

β_2(est.) $= 5.28$.

These are close to the averages of the local estimates (10; 0; 5).

In this case, the OLS calibration performs very poorly because it is trying to fit a global model to a situation in which the processes are nonstationary. The resulting parameter estimates are very close to the averages of the spatially varying local values but this 'average' model is not representative of any situation across the study region and hence the model cannot replicate the y data at all well. In addition, of course, the model provides no indication on how the processes being examined vary spatially.

Local model calibrated by GWR

Adj. $R^2 = 0.997$

AIC $= 2{,}218$

$K = 167$

$N = 129$

α_i(est.) range $= 2$ to 18.6

β_{1i}(est.) range $= -4.3$ to 4.7

β_{2i}(est.) range $= -3.9$ to 13.6.

The local model clearly captures the spatial nonstationarity in relationships extremely well. The *y* variable is replicated accurately with the adjusted *r*-squared statistic being close to 1.0 and the AIC value being much lower than the comparable value from the global model (2,218 versus 17,046). In this case, the local model is trying to make itself as local as possible and the number of nearest neighbours in each local regression is only 129. Recall also that the data on these 129 observations are not weighted as 1 but will have a weight somewhere between 0 and 1 depending on their distance from the regression point. The effective number of parameter estimates is 167 reflecting the spatially varying nature of the processes underlying the model and the ranges of local parameter estimates are close to their known values. The local parameter estimates are geocoded and can easily be mapped to display the nature of their spatial variation.

The conclusion from these two experiments is that calibration of local models by GWR allows the identification of spatial nonstationarity where it exists. Further, the GWR calibration procedure does not appear to introduce any spurious nonstationarity in situations where a global model is appropriate.

13.5. SOFTWARE FOR GWR

Software for running GWR (GWR 3.1) is available from the author and runs on any Windows platform. It has a very simple point-and-click interface which makes it very easy to calibrate models by GWR. The user can select from a Gaussian, Poisson, or binary logit GWR models. The current restrictions on data size are a maximum of 80,000 observations and 50 variables. The software also calibrates a global model for comparison and the output consists of various model diagnostics plus geocoded local parameter estimates, their local standard errors, local *t*-values and local goodness-of-fit measures.

An example of the interface is shown in Figure 13.4. The user is asked to input a data file from which the variable names are stripped off and loaded into the GWR model editor for placement in the appropriate model form. The user defines the dependent variable and a set of independent variables for the model from the variable list. The *x* and *y* coordinates of the data locations must also be designated. A kernel type (either fixed or adaptive), a calibration criterion (either CV or AICc), and an output format for the geocoded information must then be selected before the model is saved and run. The output is presented in both a listing file on the screen and an output file which is saved for subsequent processing – generally mapping of the output to see the spatial variation in local parameter estimates and goodness-of-fit statistics.

The model editor also allows extra computations. The user can select a Monte Carlo simulation exercise to examine the significance of any spatial variability in local parameter estimates and various other diagnostics can be chosen. The user also has the facility to by-pass the optimization routine for the bandwidth and input his/her own bandwidth. This can be useful to examine the effects of scale on the output: large bandwidths essentially perform regional calibrations on the data; small bandwidths perform very local calibrations.

The software is distributed on a self-loading CD which also contains sample data.

13.6. RESEARCH TOPICS

Although the initial development of GWR took place over a decade ago and it is

now becoming a relatively well-established technique, being used in many disciplines, much research remains to be done. For instance, the investigation of how the GWR format can be linked with that of spatial regression models would seem quite fruitful. One of the advantages of using GWR is that it generally accounts for much of the spatial autocorrelation in the residuals often found in global modelling. In the past, such autocorrelation has necessitated the use of various spatial regression techniques, some of which are quite complex and which has possibly hindered their adoption. This raises the question 'to what extent is the spatial autocorrelation of residuals often seen in the application of global models, a result of assuming a stationary process when the relationships being examined vary over space?' There are other reasons why

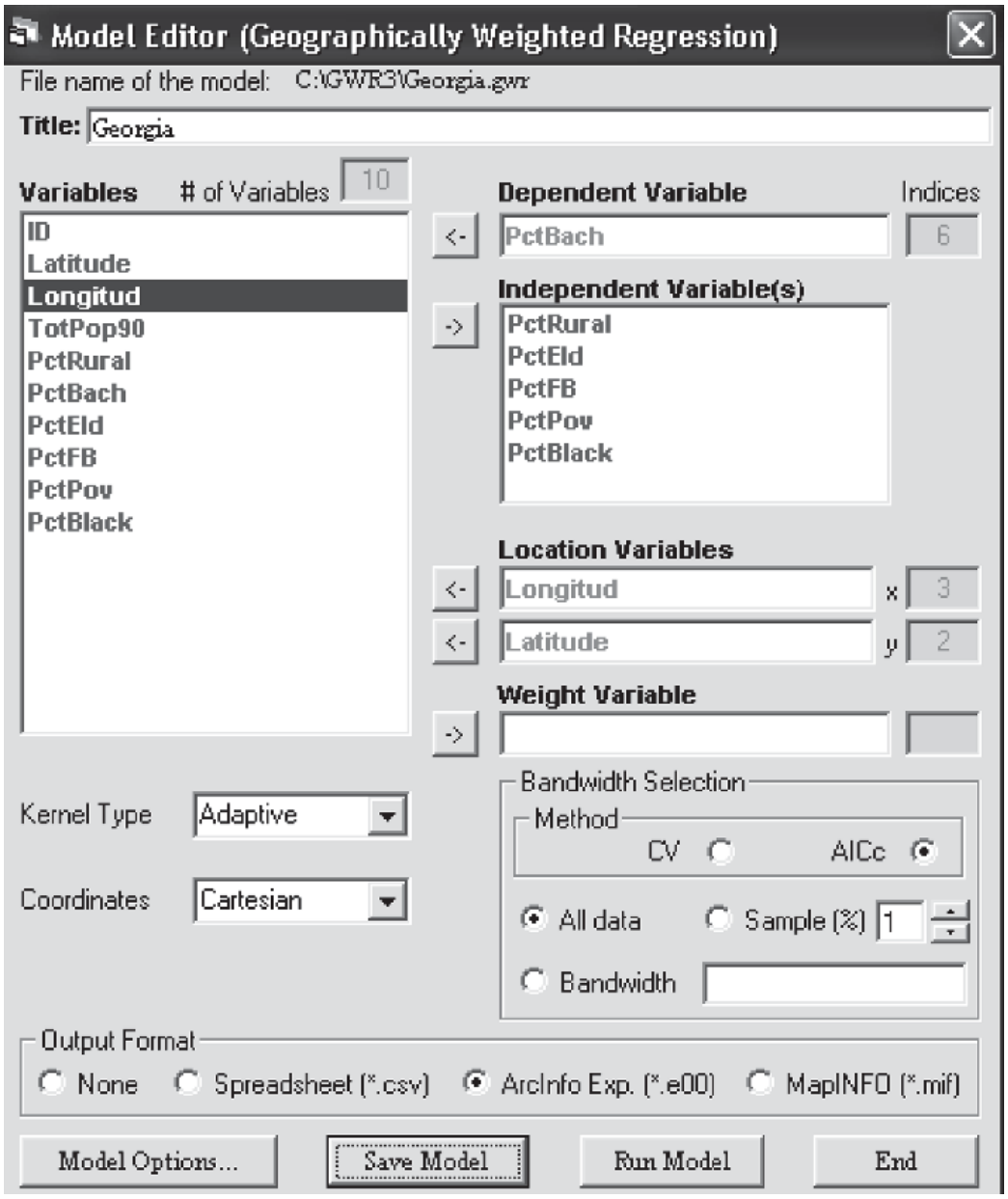


Figure 13.4 The model editor in GWR 3.1.

the residuals from global models applied to spatial data might be autocorrelated but we now have the means of examining the relative contributions of different processes to such autocorrelation. That said, there is still a potentially useful merger of spatial regression models and GWR – one could have, for example, a GWR version of a spatial regression model. If the spatial regression model were an autoregressive model, for example, this would provide an easy way of calibrating local spatial autocorrelation statistics which are free from covariate effects.

A second research area is that of the development of what are termed 'mixed' or 'semi-parametric' GWR models where some of the parameters are allowed to vary spatially whilst others are fixed globally. In some instances, for example, there is no reason to suspect that a particular relationship would be spatially varying and it makes sense to set such a parameter in the model as 'fixed'. The calibration of such models, however, is somewhat more complex than the full GWR model.

This topic leads into a related one which concerns variable selection in GWR. It should be realized that simply because a variable is insignificant at the global level, does not mean it might not be important locally. Consequently, variable selection should ideally be at the level of the GWR model and not at that of the global model. Following from the above, however, variables could either be: unimportant at the local level, important but with a stationary effect, or important with a spatially varying effect. Consequently, variable selection, along the lines of stepwise regression, is considerably more complex in GWR.

Another topic that needs further research is that of statistical inference in GWR. It is necessary to distinguish the degree of spatial variation in local parameter estimates that could reasonably be attributed to sampling variation from that which is likely to be attributable to something more interesting. Currently, this is done via Monte Carlo simulation but more formal methods might be developed. One aspect of inference that is well known in these situations is that of the multiple hypothesis testing problem which suggests that the traditional cut-off points on a statistical distribution for rejecting a null hypothesis is too liberal. Bonferroni-type adjustments should be made although recognizing that the hypothesis tests in GWR are not independent. Probably the ratio of the effective number of parameters in the GWR model to the number of parameters in the global model should be used as the adjustment factor rather than the number of tests.

Although the primary rationale for calibrating a GWR model is to uncover facets of possible nonstationarity in the processes being examined, a common question is to what extent can the methodology be used for prediction? To answer this, research is currently being undertaken to compare GWR as a prediction method with various forms of kriging. The results so far suggest GWR provides much better estimates of unknown values than do many types of kriging and about the same level of predicative ability as universal kriging with external covariates. Of course, the advantage of GWR is that much more information is yielded on the processes at work.

Finally, the most powerful aspect of GWR is the concept of geographically weighting models. Anything that can be weighted can be geographically weighted. The models need not be linear nor even in a regression format. One can generate, for example, GW versions of any descriptive spatial statistic or GW versions of any multivariate statistical method such as GWR PCA or GW discriminant analysis. The task in these latter cases is probably to handle the large volumes of output that will be generated.

13.7. SUMMARY

GWR appears to be a useful method to investigate spatial nonstationarity – simply assuming relationships are stationary over space is no longer tenable and is easily testable. GWR can be likened to a 'spatial microscope' in that it allows us to see variations in relationships that were previously unobservable. It provides a whole new 'geography of relationships' that needs explanation.

GWR can be viewed as both a model diagnostic tool or as a method to identify interesting locations for further investigation. In doing so, it conforms to two previously disparate philosophical views. From a post-modernist view relationships can be intrinsically different across space caused by differences in attitudes, preferences or different administrative, political or other contextual effects and GWR helps identify such differences. From a positivist view, global statements about relationships can be made but our models might not be properly specified to allow us to make them. GWR is then a good indicator of when and in what way a global model is mis-specified and how it can be improved. If the assumption that global statements can be made is correct and a global model fails to make them, then clearly the model is mis-specified. GWR can thus be a useful model-building tool.

Finally, GWR is a good example of a spatial statistical method. It uses locational information as well as attribute information as input, it employs a spatial weighting function with the assumption that near places are more similar than distant ones, and it produces outputs that are location-specific and geocoded so they can easily be mapped and subject to further spatial analysis. The concept of GW can be extended to many statistical techniques and there is still a great deal of work to be done.

ACKNOWLEDGEMENT

Research presented in this paper was supported by a grant to the National Centre for Geocomputation by Science Foundation Ireland (03/RP1/1382) and by a Strategic Research Cluster grant (07/SRC1/1168) from Science Foundation Ireland under the National Development Plan. The author gratefully acknowledges this support.

NOTE

1 Note that the model need not be a linear one but this is used here for convenience and because it is probably the most frequently encountered type of regression. The software described subsequently allows geographically weighted Poisson regression models and geographically weighted binary logit models to be calibrated and in theory there is no limit to what model forms can be geographically weighted.

14

Spatial Regression

Luc Anselin

14.1. INTRODUCTION

Spatial regression deals with the specification, estimation, and diagnostic checking of regression models that incorporate spatial effects. Two broad classes of spatial effects may be distinguished, referred to as spatial dependence and spatial heterogeneity (Anselin, 1988b). In this chapter, attention will be limited to the former, since spatial heterogeneity is addressed in Chapter 13, on Geographically Weighted Regression. The focus will be on ways to incorporate spatial correlation structures into a linear regression model, and the implications of this for estimation and specification testing.

Early interest in the statistical implications of estimating spatial regression models dates back to the pioneering results of the statistician Whittle (1954), followed by other by now classic papers in statistics, such as Besag (1974) and Ord (1975), and the book by Ripley (1981). It was introduced in quantitative geography through the works of Cliff and Ord (1973, 1981) and Upton and Fingleton (1985). Paralleling this was the development of the field of spatial econometrics, started by regional scientists who were concerned with spatial correlation in multiregional econometric models (Paelinck and Klaassen, 1979; Anselin, 1980). By the late 1980s and early 1990s, several compilations had appeared that included technical reviews of a range of models, estimation methods and diagnostic tests, including Anselin (1988b), Griffith (1988) and Haining (1990). In addition, the publication of the text by Cressie (1993) provided a near-comprehensive technical treatment of the statistical foundations for the analysis of spatial data.

In recent years, the interest in spatial analysis in general and spatial data analysis in particular has seen an almost exponential

growth, especially in the social sciences (Goodchild *et al*., 2000). Spatial regression analysis is a core aspect of the 'spatial' methodological toolbox and several recent texts covering the state of the art have appeared, such as Haining (2003), Waller and Gotway (2004), Banerjee *et al*. (2004), Fortin and Dale (2005), Schabenberger and Gotway (2005), and Arbia (2006). There have also been a number of edited volumes, dealing with more advanced topics, such as Bartels and Ketellapper (1979), Anselin and Florax (1995a), Anselin *et al*. (2004), Getis *et al*. (2004), and LeSage and Pace (2004). In addition, several journal special issues have recently been devoted to the topic, and they provide an excellent overview of important research directions. Such special issues include Anselin (1992, 2003), Anselin and Rey (1997), Pace *et al*. (1998), Nelson (2002), Florax and van der Vlist (2003), Pace and LeSage (2004b), and LeSage *et al*. (2004).

This chapter provides a concise overview of some of the central methodological issues related to spatial regression analysis. It consists of four sections, starting with a treatment of the specification of spatial dependence in a regression model. Next, specification tests are considered to detect the presence of spatial autocorrelation. This is followed by a review of the estimation methods, including maximum likelihood, instrumental variables/method of moments and semi-parametric methods. The chapter closes with some concluding remarks.

The treatment in this brief chapter is not intended to be comprehensive, but instead aims to provide a guide to both the current state of the art as well as to ongoing research and remaining gaps. A number of topics are not included, since they are (partially) addressed in other chapters in this volume, such as Bayesian techniques (Chapter 17). The focus is therefore entirely on regression models in a simple cross-sectional setting, leaving out other promising applications, such as the spatial econometrics of panel data (Elhorst, 2001, 2003; Anselin *et al*., 2008), the spatial econometrics of origin-destination flow models (LeSage and Pace, 2005; Fischer *et al*., 2006), the analysis of spatial latent variables (Pinkse and Slade, 1998; LeSage, 2000; Beron *et al*., 2003; Fleming, 2004), and spatial generalized linear mixed models (Gotway and Stroup, 1997; Zhang, 2002; Gotway and Wolfinger, 2003). Finally, it should be pointed out that this chapter derives from several earlier and more technical reviews dealing with various methodological aspects of spatial regression analysis, specifically, Anselin and Bera (1998), and Anselin (2001a, b, 2002, 2006). A more in-depth technical discussion can be found in those reviews.

14.2. SPECIFYING THE SPATIAL REGRESSION MODEL

The point of departure is the familiar specification of a linear regression model, where for each observation (location) i, with $i = 1, \ldots, N$, the following relationship holds:

$$y_i = \sum_k x_{ik}\beta_k + \varepsilon_i, \qquad (14.1)$$

where y_i is an observation on the dependent variable, x_{ik} an observation on an explanatory variable, with $k = 1, \ldots, K$ (including a constant term, or 1), β_k as the matching regression coefficient, and ε_i is a random error term. For ease of notation, the K explanatory variables and matching coefficients are expressed as a $K \times 1$ vector,

respectively $\mathbf{x}_i$ and β, such that the regression becomes:

$$y_i = \mathbf{x}_i\beta + \varepsilon_i. \tag{14.2}$$

In the classic regression specification, the error terms have mean zero ($E[\varepsilon_i] = 0, \forall i$), and they are identically and independently distributed (i.i.d.). Consequently, their variance is constant, $\text{Var}[\varepsilon_i] = \sigma^2$, and they are uncorrelated, $E[\varepsilon_i\varepsilon_j] = 0$, for all i, j.

In matrix notation, the N observations on the dependent variable are stacked in an $N \times 1$ vector $\mathbf{y}$, the observations on the explanatory variables in an $N \times K$ matrix $\mathbf{X}$, and the random error terms in an $N \times 1$ vector ε, such that:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{14.3}$$

with $E[\varepsilon] = \mathbf{0}$ (an $N \times 1$ vector of zeros), and $E[\varepsilon\varepsilon] = \sigma^2\mathbf{I}$ (with $\mathbf{I}$ as the identity matrix).

Spatial dependence is introduced into this specification in two major ways, one referred to as spatial lag dependence, the other as spatial error dependence (Anselin, 1988b). While the former pertains to spatial correlation in the dependent variable, the latter refers to the error term. Spatial autocorrelation can also be introduced in the explanatory variables, in so-called spatial cross-regressive models (Florax and Folmer, 1992). However, in contrast to the lag and error models, cross-regressive models do not require the application of special estimation methods. They will therefore not be further considered here.

14.2.1. Spatial lag models

A spatial lag model is a formal representation of the equilibrium outcome of processes of social and spatial interaction. Since the observations are for a single point in time, the actual dynamics of the interaction among agents (peer effects, neighborhood effects, spatial externalities) cannot be observed, but the correlation structure that results once the process has reached equilibrium is what can be modeled (Brock and Durlauf, 2001, 2004). This is also referred to as a spatial reaction function (Brueckner, 2003). In the spatial regression equation, this is accomplished by including a function of the dependent variable observed at other locations on the right-hand side:

$$y_i = g(y_{J_i}, \theta) + \mathbf{x}_i\beta + \varepsilon_i \tag{14.4}$$

where J_i includes all the neighboring locations j of i, with $j \neq i$. The function g can be very general (and non-linear), but typically is simplified by using a spatial weights matrix (see also Chapter 8 in this volume). The $N \times N$ spatial weights matrix $\mathbf{W}$ has non-zero elements w_{ij} in each row i for those columns j that are 'neighbors' of location i. The notion of neighbors is very general, and not limited to geographical concepts, but can readily be extended to neighbors in social network space (Leenders, 2002).

A so-called mixed regressive, spatial autoregressive model (Anselin, 1998b) then takes on the form:

$$y_i = \rho \sum_j w_{ij}y_j + \mathbf{x}_i\beta + \varepsilon_i \tag{14.5}$$

where ρ is the spatial autoregressive coefficient, and the error term ε_i is i.i.d. Alternatively, in matrix notation:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\beta + \varepsilon. \tag{14.6}$$

With a row-standardized spatial weights matrix (i.e., the weights standardized

such that $\sum_j w_{ij} = 1, \forall i$), this amounts to including the average of the neighbors as an additional variable into the regression specification. This added variable is referred to as a spatially lagged dependent variable, or a spatial lag. For example, in a model for tax rates of local communities, this would add the average of the tax rates in the neighboring locations as an explanatory variable.

The inclusion of the spatial lag is similar to an autoregressive term in a time series context, hence it is called a spatial autoregressive model, although there is a fundamental difference. Unlike time dependence, dependence in space is multidirectional, implying feedback effects and simultaneity. More precisely, if i and j are neighboring locations, then y_j enters on the right-hand side in the equation for y_i, but y_i also enters on the right-hand side in the equation for y_j (the neighbor relation is symmetric). This endogeneity must be accounted for in the estimation process.

The proper solution to the equations for all observations is the so-called reduced form, which no longer contains any spatially lagged dependent variables on the right-hand side. After some matrix algebra, this follows as:

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\beta + (\mathbf{I} - \rho\mathbf{W})^{-1}\varepsilon \tag{14.7}$$

a model that is nonlinear in ρ and β and has a spatially correlated error structure (more precisely, a spatial autoregressive structure, see below). More importantly, this reveals the *spatial multiplier*, i.e., the notion that the value of y at any location i is not only determined by the values of x at i, but also of x at all other locations in the system. This can be seen after a simple expansion of the inverse matrix term (for $| \rho | < 1$ and with a row-standardized $\mathbf{W}$), and using the expected value (since the errors all have mean zero):

$$E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta + \rho\mathbf{W}\mathbf{X}\beta + \rho^2\mathbf{W}^2\mathbf{X}\beta + \cdots \tag{14.8}$$

The powers of ρ matching the powers of the weights matrix (higher orders of neighbors) ensures that a distance decay effect is present.

Even when the spatial lag specification is not necessarily the result of a process of interaction among agents, it remains a useful model to deal with spatial autocorrelation, and can be interpreted as a filtering model. More precisely, moving the spatial lag term to the left-hand side reveals:

$$y_i^* = y_i - \rho \sum_j w_{ij} y_j = \mathbf{x}_i\beta + \varepsilon_i \tag{14.9}$$

i.e., a standard regression model in a dependent variable y_i^* from which the spatial correlation has been removed (filtered). Unlike detrending time series data, however, the ρ parameter cannot take on the value of 1 and must be estimated jointly with the other parameters of the model. The spatial filtering interpretation is often useful when there is a mismatch between the spatial scale of observations and the spatial scale at which the phenomenon of interest manifests itself. For example, this would be the case when a regional phenomenon (e.g., a labor market or housing market) is measured at a subregional scale, resulting in a high degree of positive spatial autocorrelation (very little change across the sub-regional scale). In that situation, the estimation of the spatial lag model will yield estimates for the β parameters that properly control for the spatial autocorrelation.

14.2.2. Spatial error models

In spatial error models, the spatial autocorrelation does not enter as an additional variable in the model, but instead affects the covariance structure of the random disturbance terms. The typical motivation for this is that unmodeled effects spill over across units of observation and hence result in spatially correlated errors. For example, in hedonic house price models, it is often assumed that neighborhood effects that are hard (or impossible) to quantify are shared by houses in similar locations and thus appear as spatially correlated errors (Dubin, 1988). More recently, a theoretical framework based on common shocks has been suggested as a mechanism to motivate spatially correlated errors (Andrews, 2005).

Spatial error autocorrelation is a special case of a non-spherical error covariance matrix, in which the off-diagonal elements are non-zero, i.e., $E[\varepsilon_i \varepsilon_j] \neq 0$, for $i \neq j$, or, in matrix notation, $E[\varepsilon\varepsilon] = \mathbf{\Sigma}$. The value and pattern of the non-zero covariances are the outcome of a spatial ordering. In a cross-section, it is impossible to extract this ordering from the data directly, since there are potentially $N \times (N-1)/2$ covariance parameters and only N observations to estimate them from. Hence, it is necessary to impose structure and to obtain estimates from a more parsimonious specification.

The spatial covariance structure can be obtained in a number of ways. One of the earliest suggestions was a so-called *direct representation*. In this, each covariance between a pair of observations $i-j$ is specified as a parameterized function f (with parameter vector ϕ) of the distance d_{ij} between them, or, $E[\varepsilon_i \varepsilon_j] = \sigma^2 f(d_{ij}, \phi)$. Early applications of this approach were Cook and Pocock (1983) and Mardia and Marshall (1984). More recently, it is commonly used in analyses of real estate markets, as reviewed in Dubin *et al.* (1999).

The choice of the function and of the distance metric needs to be made very carefully, in order to ensure that the resulting variance–covariance matrix is positive definite. A common choice is a negative exponential distance decay function. This results in an error variance–covariance matrix of the form:

$$E[\varepsilon\varepsilon] = \sigma^2[\mathbf{I} + \gamma\mathbf{\Psi}] \tag{14.10}$$

where the variance is accounted for in the first term, γ is a non-negative scaling parameter, and the off-diagonal elements of $\mathbf{\Psi}$ are $\Psi_{ij} = e^{-\phi d_{ij}}$.

A second approach obtains structure for the error covariance matrix by specifying a *spatial process* for the random disturbance. A number of processes may be considered, each yielding a different covariance structure, expressed as a function of one or two parameters. The most common choice is a spatial autoregressive process, or SAR:

$$\varepsilon_i = \lambda \sum_j w_{ij}\varepsilon_j + u_i \tag{14.11}$$

with λ as the autoregressive parameter and u_i as a random error term, typically assumed to be i.i.d. In matrix notation, this is equivalent to:

$$\varepsilon = \lambda\mathbf{W}\varepsilon + \mathbf{u}. \tag{14.12}$$

Solving this for the full vector or errors ε yields:

$$\varepsilon = (\mathbf{I} - \lambda\mathbf{W})^{-1}\mathbf{u} \tag{14.13}$$

with $E[\mathbf{uu}] = \sigma^2\mathbf{I}$, so that the complete error variance–covariance matrix follows as:

$$E[\varepsilon\varepsilon] = \sigma^2(\mathbf{I} - \lambda\mathbf{W})^{-1}(\mathbf{I} - \lambda\mathbf{W}')^{-1}. \quad (14.14)$$

Even though the spatial weights matrix $\mathbf{W}$ may contain only a few neighbors for each observation, the variance–covariance structure that results from the SAR process is a non-sparse matrix, representing a *global* pattern of spatial autocorrelation. Moreover, unless the number of neighbors is constant for each observation, the diagonal elements in the variance–covariance matrix will not be constant, resulting in heteroskedasticity. This induced heteroskedasticity is a distinguishing characteristic for spatial processes, and it complicates specification testing and estimation. More precisely, since many of the theoretical asymptotic results in time series analysis are based on assumptions of constant variance, they do not translate directly to spatial processes; for technical details, see, e.g., Anselin (2006).

Other spatial processes used to provide structure to the error variance–covariance matrix include a conditional autoregressive process (CAR) and a spatial moving average process (SMA). The CAR model is often used as a prior in hierarchical Bayesian specifications, whereas the SMA specification is appropriate for *local* patterns of spatial autocorrelation (for details, see Anselin, 2006).

Error component models have been suggested as well, and some recent theoretical results provide the basis for a wide range of structures for error spatial autocorrelation. In Kelejian and Robinson (1992), an error decomposition was proposed that combined a local or location-specific component with a spillover component, yielding an error variance–covariance structure similar to that of an SMA (see also Anselin and Moreno, 2003). The common shocks framework outlined in Andrews (2005) can encompass general factor structures yielding different specifications for the range and strength of spatial autocorrelation. This approach has seen increased application in recent work on spatial autocorrelation in panel data models (Pesaran, 2005).

A final approach to provide structure to spatial error variance–covariance matrices is based on a non-parametric rationale, which is particularly appropriate for local patterns of spatial autocorrelation. Using the formal properties for a kernel estimator of spatial autocovariance established by Hall and Patil (1994), a general non-parametric covariance matrix estimator has been suggested by Conley (1999), and, more recently, by Kelejian and Prucha (2007).

14.3. HIGHER ORDER MODELS

In addition to the basic spatial lag and spatial error models just reviewed, higher order models can be specified as well, by including multiple weights matrices, by combining lag and error structures, and by including specification for spatial heterogeneity jointly with spatial dependence. An extensive review of these specifications can be found in Anselin (2006).

14.4. SPECIFICATION TESTS

In empirical practice, there are often no strong *a priori* reasons to consider a spatial lag or spatial error model in a cross-sectional situation. Instead, the need for such a specification follows from the result of model diagnostics. Specifically, diagnostic tests derived from the residuals of

a regression carried out by means of ordinary least squares (OLS) may point to violations of the underlying assumptions, including the uncorrelatedness of errors.

Ignoring spatial autocorrelation when it is in fact present has different consequences, depending on whether the correct model is a spatial lag or a spatial error specification. Ignoring a spatially lagged dependent variable is equivalent to an omitted variable error, and will yield OLS estimates for the model coefficients that are biased and inconsistent. On the other hand, ignoring spatially correlated errors is mostly a problem of efficiency, in the sense that the OLS coefficient standard error estimates are biased, but the coefficient estimates themselves remain unbiased. However, to the extent that the spatially correlated errors mask an omitted variable, the consequences of ignoring this may be more serious.

The problem at hand is the extent to which any systematic spatial patterning in the residuals provides evidence to reject the null hypothesis of uncorrelated errors. There are two complications in this respect. One is that the null hypothesis pertains to the error terms, which are not observable. Instead, one has to deal with residuals. OLS regression residuals are already correlated by construction, since they are derived from a common set of data. Hence, simply concluding from correlated residuals that the errors are also correlated may be spurious. Another complication is that the rejection of the null hypothesis does not necessarily suggest a given spatial model as the proper alternative. Both spatial lag and spatial error alternatives will, when ignored, lead to OLS residuals that are spatially correlated. In addition, since several spatial processes also result in heteroskedastic errors, distinguishing true heteroskedasticity from this type of induced heteroskedasticity will constitute an added complication.

Specification tests against spatial autocorrelation are either based on a specific alternative model, referred to as focused tests, or are diffuse, in that the alternative is an unspecified form of spatial correlation. In the remainder of the section, diffuse or spatial autocorrelation tests are considered first, followed by focused tests based on the maximum likelihood principle. The section concludes with a discussion of the practice of a specification search.

14.4.1. Spatial autocorrelation tests

Arguably the best known test statistic against spatial autocorrelation is the application of Moran's I statistic for spatial autocorrelation (Moran, 1948) to regression residuals (Moran, 1950), popularized in the work of Cliff and Ord (1972, 1973, 1981). This statistic corrects the well known Moran's I for the fact that the random variable under consideration is a regression residual. As a result, inference is based on analytical and asymptotic results, but should not rely on the familiar permutation approach (Anselin and Rey, 1991; Schmoyer, 1994).

Moran's I for regression residuals is then:

$$I = \frac{\mathbf{e}'\mathbf{W}\mathbf{e}/S_0}{\mathbf{e}'\mathbf{e}/N} \qquad (14.15)$$

where $\mathbf{e}$ is a $N \times 1$ vector of OLS residuals $\mathbf{y} - \mathbf{X}\hat{\beta}$, $\mathbf{W}$ is a spatial weights matrix, and $S_0 = \sum_i \sum_j w_{ij}$, a normalizing factor.

In practice, inference in Moran's I test can be based on a normal approximation, using a standardized value, or z-value. This is obtained by subtracting the mean under the null and dividing by the square root of the variance. The first two moments were derived in Cliff and Ord (1972) as:

$$E[I] = \text{tr}(\mathbf{MW})/(N - K) \qquad (14.16)$$

and:

$$\text{Var}[I] = \frac{\text{tr}(\mathbf{MWMW}') + \text{tr}(\mathbf{MWMW}) + [\text{tr}(\mathbf{MW})]^2}{(N-K)(N-K+2)} - (E[I])^2 \quad (14.17)$$

where tr is a matrix trace operator and $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The normality of the z-value is an approximation, which works well in large samples. Alternatives are to use exact inference (under the assumption of Gaussian error terms, as in Tiefelsdorf and Boots, 1995), or a saddlepoint approximation (Tiefelsdorf, 2002).

Moran's I has been shown to have certain optimal properties, similar to the Durbin–Watson test against serial correlation in the time domain (King, 1981). Also, it turns out to be asymptotically equivalent to a likelihood ratio (LR) test and to a Lagrange multiplier (LM) test (Cliff and Ord, 1972; Burridge, 1980), and therefore shares the asymptotic properties of these statistics.

Moran's I has power against any alternative of spatial correlation, including spatial lag dependence, as demonstrated in a large number of Monte Carlo simulation experiments (see, e.g., Anselin and Rey, 1991; Anselin and Florax 1995b; Florax and de Graaff, 2004). In addition, not unlike the Durbin–Watson statistic, the test has power against heteroskedasticity as well (Anselin and Griffith, 1988). In practice, this complicates specification testing in that without further evidence, it will be difficult to conclude whether a spatial model, a heteroskedastic model, or a combination of the two is the proper alternative.

Moran's I test statistic is very general and can be applied in many contexts other than the classic regression model. For example, in Anselin and Kelejian (1997) it is extended to residuals from a two stage least squares (2SLS) regression estimation. Kelejian and Prucha (2001) formulate a general framework to obtain the asymptotic properties of the statistic in a wide range of contexts. Ellner and Seifu (2002) use Moran's I as a model diagnostic to select the proper bandwidth for kernel estimators in semi-parametric models. In this application, the weights matrix does not pertain to geographic locations, but to locations in 'variable space'.

An alternative to Moran's I as a test statistic against an unspecified form of spatial autocorrelation was suggested by Kelejian and Robinson (1992). Theirs is a large sample test, which does not require an assumption of normality and can be applied in nonlinear models as well. In Kelejian and Robinson (1998, 2004), this principle is extended to include both heteroskedasticity and error autocorrelation as the alternative.

14.4.2. Maximum likelihood based tests

In contrast to diffuse spatial autocorrelation tests, focused tests are constructed with a specific alternative in mind, such as a spatial lag or a spatial error specification. In general, they boil down to a test of restrictions on the parameters of a spatial regression model. For example, for a spatial lag model, the null hypothesis would be $H_0 : \rho = 0$, such that the *restricted* model would then be $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. The alternative hypothesis is then that $H_1 : \rho \neq 0$, such that the *unrestricted* model is $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\beta + \varepsilon$. The three classic test statistics obtained under maximum likelihood (ML) estimation are the Wald, likelihood ratio, and Lagrange multiplier (or, Rao score) tests.

The Wald, or asymptotic t-test is simply a significance test on the spatial autoregressive parameter in a spatial lag or spatial error model, based on the results of estimation by means of maximum likelihood of the

unrestricted (spatial) model. This requires both the point estimate of the parameter as well as an estimate of the asymptotic variance matrix (for technical details, see Anselin, 1988b, Ch. 6).

The likelihood ratio test statistic is obtained in the standard manner as well, as twice the difference between the log-likelihood of the unrestricted (i.e., the spatial) model, and that of the restricted model (i.e., the standard regression without spatial autocorrelation). This thus requires the estimation of two models, and an assumption of normality for the OLS regression. The statistic is asymptotically distributed as $\chi^2(1)$ (see Anselin, 1988b, Ch. 6).

The Lagrange multiplier (LM) test only requires estimation of the model under the null hypothesis of no spatial dependence. It therefore lends itself well to specification searches in practice, since the extra step of estimating a spatial lag or spatial error model can often be avoided. In the spatial case, the LM statistic does not follow the standard result from econometrics, where in many instances it can be obtained as a measure of fit in an auxiliary regression. Instead, it needs to be derived explicitly, as in Burridge (1980) and Anselin (1988a) (for extensive technical details, see also Anselin and Bera, 1998; Anselin, 2001a).

Even though the LM statistic is constructed from the OLS residuals, a complete alternative model must be specified. In some instances, two different alternatives yield the same LM statistic. These are called locally equivalent alternatives (Godfrey, 1981). SAR and SMA error processes fall into this category. As a result, a LM test statistic against spatial error autocorrelation cannot distinguish between these two different processes. In practice, this affects the interpretation of the results, since SAR is a global spatial process, while SMA is local.

The LM error statistic is very similar to Moran's I. As shown in Burridge and (1980) and Anselin (1988a), the statistic is:

$$LM_\lambda = \frac{[\mathbf{e}'\mathbf{W}\mathbf{e}/(\mathbf{e}'\mathbf{e}/N)]^2}{\text{tr}[\mathbf{W}'\mathbf{W} + \mathbf{W}\mathbf{W}]} \qquad (14.18)$$

where $\mathbf{e}$ is a $N \times 1$ vector of OLS residuals, and tr stands for the trace operator (the sum of the diagonal elements of a matrix). Except for the scaling factor in the denominator, this statistic is essentially the square of Moran's I. It is asymptotically distributed as $\chi^2(1)$.

Using similar principles, the LM lag statistic follows as:

$$LM_\rho = [\mathbf{e}'\mathbf{W}\mathbf{y}/(\mathbf{e}'\mathbf{e}/N)]^2/D \qquad (14.19)$$

with $\mathbf{e}$ as the OLS residuals, and the denominator term:

$$D = [(\mathbf{W}\mathbf{X}\hat{\beta})'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'](\mathbf{W}\mathbf{X}\hat{\beta})/\hat{\sigma}^2] + \text{tr}(\mathbf{W}'\mathbf{W} + \mathbf{W}\mathbf{W}) \qquad (14.20)$$

where the estimates for $\hat{\beta}$ and $\hat{\sigma}^2$ are from OLS. The test statistic is asymptotically distributed as $\chi^2(1)$.

A related test statistic, also based on the maximum likelihood principle, applies the idea of double length artificial regressions (DLR, Davidson and MacKinnon, 1984, 1988) to tests for spatial error and spatial lag dependence (Baltagi and Li, 2001a). The DLR approach consists of expressing the regression model as a function of standard normal error terms. In the spatial models, this follows as a simple standardization (for technical details, see Baltagi and Li, 2001a).

The LM principle can be applied to alternatives other than the SAR/SMA error processes or the spatial lag model. Test statistics can be derived for higher-order

processes (multiple orders of contiguity), and for different error models, such as spatial error components or direct representation (Anselin, 2001a; Anselin and Moreno, 2003).

So far, only a single alternative has been taken into account. However, in practice, it is often more reasonable to consider an alternative hypothesis that contains both a spatial lag and spatial error autocorrelation:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\beta + \varepsilon \qquad (14.21)$$

with:

$$\varepsilon = \lambda\mathbf{W}\varepsilon + \mathbf{u} \qquad (14.22)$$

a SARSAR model, or, with:

$$\varepsilon = \lambda\mathbf{W}\mathbf{u} + \mathbf{u} \qquad (14.23)$$

a SARMA model.

In this more general case, there are three ways to proceed. One is as before, considering a one-directional alternative only and ignoring the other form of spatial autocorrelation. For example, the LM error test above has the null hypothesis $H_0 : \lambda = 0$, irrespective of the value of ρ, which is considered to be a nuisance parameter. This is referred to as a *marginal* test.

A problem with the marginal approach is that the LM_λ and LM_ρ test statistics are no longer $\chi^2(1)$ in the presence of local mis-specification in the form of the other type of spatial dependence, but they become non-central χ^2. In other words, in the presence of spatial lag dependence, the LM_λ test against error correlation becomes biased, and, in the presence of spatial error dependence, the LM_ρ test against lag dependence becomes biased. Using a result of Bera and Yoon (1993), robust versions of these test statistics have been developed in Anselin *et al.* (1996) (see also Anselin and Bera, 1998, pp. 273–278).

A second strategy is that of a *joint* test, where the null hypothesis is to set all spatial parameters equal to zero. For example, for the spatial lag model with a SAR or SMA error term, $H_0 : \rho = \lambda = 0$. In contrast to standard results in the econometric literature, the joint test statistic is not simply the sum of the marginal test statistics, i.e., $LM_{\lambda\rho} \neq LM_\lambda + LM_\rho$, but it takes on a far more complex form (Anselin, 1988a).

A third strategy is a so-called *conditional* approach, where a test on the null hypothesis $\rho = 0$ is carried out in a model with $\lambda \neq 0$, and vice versa. This can no longer be based on OLS estimates, but requires estimation of the proper spatial model by means of ML. Using the same principles as before, but now with the residuals of the ML estimation, a test statistic for $H_0 : \lambda = 0$ in the spatial lag model (i.e., with $\rho \neq 0$) can be derived. Similarly, a test statistic can be constructed for $H_0 : \rho = 0$ in the spatial error model (i.e., with $\lambda \neq 0$). While straightforward, the derivations are quite tedious and the resulting test statistics complex (for technical details, see Anselin, 1988a; Anselin *et al.*, 1996; Anselin and Bera, 1998).

The LM principle can also be extended to multiple sources of mis-specification, such as spatial dependence and heteroskedasticity (Anselin, 1988b), or spatial dependence and functional mis-specification (Baltagi and Li, 2001b).

14.4.3. Specification search

In practice, the sheer number of available test statistics can seem overwhelming and a strategy needs to be developed to move from the null model to a superior alternative (when appropriate). Given that tests may be based

on marginal, joint, or conditional approaches, the results of a specification search may be subject to the order in which tests are carried out, and whether or not adjustments are made for pre-testing (see, e.g., Florax and Folmer, 1992; Anselin and Florax, 1995b; Florax and de Graaff, 2004).

Based on a large number of simulation results, an *ad hoc* decision rule was suggested in Anselin and Rey (1991) for the simple case of choosing between a spatial lag or spatial (SAR) error alternative. There is considerable evidence that the proper alternative is most likely the one with the largest significant LM test statistic value. This was later refined in light of the robust forms of the statistics in Anselin *et al*. (1996). In a recent paper by Florax *et al*. (2003), this classic forward stepwise specification search is compared to a 'general-to-simple' model selection rule (for further discussion, see also Florax *et al*., 2006; Hendry, 2006).

14.5. ESTIMATION

The estimation problems associated with spatial regression models are distinct for the spatial lag and spatial error case. Spatial error models are special instances of specifications with a non-spherical error. As a result, OLS may still be applied, as long as the estimated standard errors are adjusted to take into account the error correlation. In contrast, the inclusion of a spatially lagged dependent variable in a regression specification yields a form of endogeneity. As a result, for most spatial weights used in practice, OLS in the spatial lag model is not an appropriate method, and the simultaneity must be accounted for explicitly. An exception to this general rule is when the weights represent subgroups in the data (i.e., all the observations in the same group are neighbors of each other), in which case OLS turns out to yield consistent estimates (Lee, 2002; Kelejian and Prucha, 2002).

Two general sets of methods have been developed to address the estimation of spatial regression models, one based on the maximum likelihood (ML) principle, the other on the (general) method of moments (GMM). Each will be considered in turn, followed by a brief overview of semi-parametric methods.

14.5.1. Maximum likelihood estimation

The point of departure for maximum likelihood estimation in spatial regression models is an assumption of normality for the error term. In general, allowing for heteroskedasticity and/or error correlation, the $N \times 1$ error vector has a multivariate normal distribution, $\varepsilon \sim N(0, \Sigma_\theta)$, with the subscript θ denoting that Σ may be a function of a $p \times 1$ vector θ of parameters. In the commonly considered i.i.d. case, this simplifies to $\varepsilon \sim N(0, \sigma^2\mathbf{I})$, with $\theta = \sigma^2$.

To move from the likelihood for the error vector to a likelihood for the observed dependent variable, a *Jacobian* of the transformation needs to be inserted, which corresponds to the determinant $|\mathbf{I} - \rho\mathbf{W}|$ in the spatial lag model, and $|\mathbf{I} - \lambda\mathbf{W}|$ in the spatial error model. The presence of the Jacobian term constitutes a major computational complication.

Using the standard result for a multivariate normal distribution, and taking into account the Jacobian term, the log-likelihood for the spatial lag model follows as:

$$\begin{aligned} L = &-(N/2)(\ln 2\pi) - (1/2)\ln|\Sigma_\theta| \\ &+ \ln|\mathbf{I} - \rho\mathbf{W}| - (1/2)(\mathbf{y} - \rho\mathbf{W}\mathbf{y} - \mathbf{X}\beta)' \\ &\times \Sigma_\theta^{-1}(\mathbf{y} - \rho\mathbf{W}\mathbf{y} - \mathbf{X}\beta). \end{aligned} \tag{14.24}$$

Maximizing the log-likelihood is *not* equivalent to minimizing weighted least squares (the last term in L), as in the standard linear regression model. The main difference is in the presence of the log-Jacobian term $\ln|\mathbf{I}-\rho\mathbf{W}|$. This illustrates informally how weighted least squares will not yield a consistent estimator in the spatial lag model, due to the endogeneity in the $\mathbf{Wy}$ term. The log-Jacobian also implies constraints on the parameter space for ρ, which must be such that $|\mathbf{I}-\rho\mathbf{W}|>0$.

Maximum likelihood estimates for β, ρ, and θ are obtained as solutions to the usual first-order conditions, requiring numerical optimization (for technical details, see Ord (1975), Cliff and Ord (1981), Anselin (1980, 1988b, 2006), Anselin and Bera (1998), among others). Inference is based on an asymptotic variance matrix, the inverse of the information matrix (see Anselin, 1980, 1988b).

Even though the principles of ML estimation in a spatial lag model were laid out more than 30 years ago by Ord (1975), it was only very recently that the formal proofs were developed that established the conditions under which consistency and asymptotic normality of this estimator are obtained (Lee, 2004).

Maximum likelihood estimation of the parameters in models with spatially dependent error terms follows as a special case of the results in Magnus (1978). For a general non-spherical error term Σ_θ, with θ as the parameters, the ML estimator for β is the familiar generalized least squares expression:

$$\hat{\beta}_{\mathrm{ML}}=(\mathbf{X}'\Sigma_\theta^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_\theta^{-1}\mathbf{y}. \qquad (14.25)$$

This follows as the solution of the first-order conditions, applied to the log-likelihood:

$$L=-(N/2)\ln(2\pi)-(1/2)\ln|\Sigma_\theta| -(\mathbf{y}-\mathbf{X}\beta)'\Sigma_\theta^{-1}(\mathbf{y}-\mathbf{X}\beta). \qquad (14.26)$$

With a consistent estimate for the parameters θ_i, consistent estimates for β are obtained through feasible generalized least squares (FGLS).

Each spatial error process will result in a specialized form for Σ_θ. For example, for a SAR error process without heteroskedasticity, the corresponding parameter vector is $\theta=[\sigma^2,\lambda]$. The FGLS estimator in this model simplifies to:

$$\hat{\beta}_{\mathrm{ML}}=[\mathbf{X}'(\mathbf{I}-\hat{\lambda}\mathbf{W})'(\mathbf{I}-\hat{\lambda}\mathbf{W})\mathbf{X}]^{-1} \times\mathbf{X}'(\mathbf{I}-\hat{\lambda}\mathbf{W})'(\mathbf{I}-\hat{\lambda}\mathbf{W})\mathbf{y} \qquad (14.27)$$

or, a regression of spatially filtered $\mathbf{y}_L=\mathbf{y}-\hat{\lambda}\mathbf{Wy}$ on spatially filtered $\mathbf{X}_L=\mathbf{X}-\hat{\lambda}\mathbf{WX}$. This is referred to as spatially weighted least squares. Unlike the time series counterpart, a consistent estimate for λ cannot be obtained from a simple auxiliary regression, but the first-order condition must be solved explicitly by numerical means. As for the spatial lag model, asymptotic inference is based on the inverse of the information matrix (for technical details, see Anselin, 1988b, Chapter 6).

Maximum likelihood estimation in spatial regression models involves the application of nonlinear optimization techniques to the log-likelihood function. A main computational obstacle follows from the presence of the log-Jacobian term $\ln|\mathbf{I}-\rho\mathbf{W}|$ in the log-likelihood. In addition, the first-order conditions and information matrix involve the traces of matrix products such as $\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1}$. For even medium-sized data

sets, the computation of these terms by 'brute force' is impractical.

An early solution was suggested by Ord (1975), who exploited the decomposition of the Jacobian in terms of the eigenvalues of the spatial weights matrix. This facilitates computation greatly, since the eigenvalues only need to be calculated once. The trace terms used in the information matrix can be expressed in terms of the eigenvalues as well (Anselin, 1980).

The computation of eigenvalues becomes impractical and computationally unstable for medium and large-sized data sets ($n > 1000$). This precludes the application of the Ord approach. Several alternatives have been suggested that either approximate or bound the Jacobian or log-Jacobian term (e.g., Martin, 1993; Griffith and Sone, 1995; Barry and Pace, 1999; Pace and LeSage, 2002, 2004a), or exploit the sparse nature of spatial weights (Pace and Barry, 1997a, b; Smirnov and Anselin, 2001).

A second important computational problem pertains to the presence of terms like $\text{tr}[\mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}]^2$ in the information matrix. The calculation of these inverse matrices is impractical in large data settings. As a result, most large data ML methods developed so far have not based inference on the asymptotic variance matrix, but instead use a sequence of likelihood ratio tests. Recently, Smirnov (2005) developed a solution to this problem, based on the use of a conjugate gradient approach.

14.6. INSTRUMENTAL VARIABLES/METHOD OF MOMENTS ESTIMATION

An alternative to maximum likelihood estimation is the use of the method of moments (including instrumental variables, generalized method of moments, and generalized moments). This approach does not require an assumption of normality and it avoids some of the computational problems associated with ML for very large data sets.

The spatial lag model can be formulated as a linear model that contains an endogenous variable ($\mathbf{Wy}$) and exogenous variables ($\mathbf{X}$):

$$\mathbf{y} = \mathbf{Z}\gamma + \varepsilon \tag{14.28}$$

with $\mathbf{Z} = [\mathbf{Wy}, \mathbf{X}]$ and $\gamma = [\rho, \beta]$. A classic solution to the endogeneity problem is to use instrumental variables. A matrix of additional variables $\mathbf{Q}\,(N \times q)$ is used to obtain an instrument for the spatially lagged dependent variable:

$$\widehat{\mathbf{Wy}} = \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\,\mathbf{Wy} \tag{14.29}$$

such that $\widehat{\mathbf{Z}} = [\widehat{\mathbf{Wy}}, \mathbf{X}]$, resulting in the spatial two-stage least squares estimator (S2SLS):

$$\hat{\gamma}_{\text{S2SLS}} = [\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}]^{-1}\widehat{\mathbf{Z}}\mathbf{y}. \tag{14.30}$$

Inference on the γ_{S2SLS} is based on the asymptotic variance matrix:

$$\text{AsyVar}[\hat{\gamma}_{\text{S2SLS}}] = \hat{\sigma}^2[\mathbf{Z}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{Z}]^{-1} \tag{14.31}$$

with $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{Z}\hat{\gamma}_{\text{S2SLS}})'(\mathbf{y} - \mathbf{Z}\hat{\gamma}_{\text{S2SLS}})/N$.

The application of instrumental variables to the spatial lag model was initially outlined in Anselin (1980, 1988b, pp. 82–86), where some *ad hoc* suggestions were made for the selection of the instruments (see also Land and Deane (1992) for an early discussion).

Specifically, the choice of a spatial lag of the predicted values of the **y** (using only the exogenous variables) or of spatially lagged exogenous variables was considered. In Kelejian and Robinson (1993), proof is provided of the consistency of $\hat{\gamma}_{S2SLS}$ and the selection of instruments is couched in terms of the reduced form. This suggests the use of a subset of columns from $\{\mathbf{X}, \mathbf{WX}, \mathbf{W}^2\mathbf{X}, \mathbf{W}^3\mathbf{X}, \ldots\}$ as the instruments (see also Kelejian and Prucha, 1998).

Recent work has focused on the selection of optimal instruments (Lee, 2003; Das *et al.*, 2003; Kelejian *et al.*, 2004), and on establishing formal proofs of consistency and asymptotic normality. In Lee (2007), the S2SLS estimator is compared to a GMM method with superior asymptotic properties. Extensions of the instrumental variables approach to systems of simultaneous equations are considered in Rey and Boarnet (2004) and Kelejian and Prucha (2004).

Moment methods have been developed to address spatial error autocorrelation as well, both in isolation as well as in combination with a spatial lag model (the SARSAR model). The basic results were obtained by Kelejian and Prucha (1998, 1999), who initially treated the spatial autoregressive coefficient in the error SAR process as a nuisance parameter. Specifically, attention focused on obtaining a consistent estimate for the nuisance parameter as the solution of a set of moment conditions. This consistent estimate could then be used in a second step of a FGLS estimation. One drawback of the nuisance parameter approach is that no inference can be carried out on the spatial autoregressive parameter, since no result existed on its asymptotic variance. In recent work by Lin and Lee (2005) and Kelejian and Prucha (2006), this problem has been alleviated, in the context of an extended set of moment conditions that account for both spatial autoregressive errors as well as heteroskedasticity of unspecified form. Their results also yield an asymptotic variance matrix, so that tests of significance can be carried out on the spatial parameters as well.

14.6.1. Semi-parametric methods

Semi-parametric methods provide a compromise between a full parametric specification and a non-parametric approach where the parameters are completely determined by the data, with very little prior structure. The combination of a full specification of the parts where theory or previous results provide a strong support for the model and relaxing the functional and distributional assumptions for the rest has become very attractive, especially when large data sets provide ample information (for a recent review, see Horowitz and Lee, 2002).

While by far the predominant paradigm in spatial regression analysis is the parametric approach, the use of semi-parametric techniques has seen a recent increase and is an area of very active research, both theoretical as well as applied. A semi-parametric approach has seen application in four main areas in spatial regression analysis.

One is as an alternative to specifying a specific spatial process for the error term. Instead, the error covariance may be estimated in a non-parametric fashion. This follows along the lines of the work in econometrics by White (1980) on a heteroskedastic-consistent approach, and its extension to both heteroskedasticity and serial correlation by Newey and West (1987), and others. The incorporation of spatial dependence in this framework was first considered by Conley (1999) in the context of GMM estimation, and recently elaborated upon in Kelejian and Prucha (2007) (see also Chen and Conley (2001), for a related approach). The basic idea is to avoid specifying a particular spatial process or spatial weights matrix and to extract

the spatial covariance terms from weighted averages of cross-products of residuals, using a kernel function. This yields a so-called heteroskedastic and spatial autocorrelation consistent (HAC) estimator. The HAC approach is asymptotic and in finite samples a major practical problem is to ensure that the estimated variance–covariance matrix is positive semidefinite. A number of suggestions have been formulated, but considerable research remains to be done to obtain insight into finite sample properties (see Kelejian and Prucha (2007), for some technical details).

In a second approach, the focus is on relaxing the requirements to specify a spatial weights matrix $\mathbf{W}$ in the construction of the spatially lagged dependent variable in a spatial lag model. In Pinkse *et al*. (2002), a model is considered of the form:

$$y_i = \sum_{j \neq i} g(d_{ij}) y_j + \mathbf{x}_i \beta + \varepsilon_i \qquad (14.32)$$

in which the unspecified function g relates the values of y at other locations j to that at i through a distance measure d_{ij}. The function g is approximated by a polynomial series expansion in distance measures, the coefficient of which are estimated jointly with the other parameters in the model.

In a third approach, suggested in the work of Gress (2004a) (see also Gress (2004b), and Basile and Gress (2005), for applications), the spatial weights specification is kept in the spatial lag part, but the other variables enter into the model in a non-parametric way. For example, a semi-parametric spatial lag model takes the form:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + g(\mathbf{X}) + \varepsilon \qquad (14.33)$$

where g is an unspecified function, to be estimated in a non-parametric way. A semi-parametric spatial error model is considered as well, using residuals from a non-parametric regression of $\mathbf{y}$ on $g(\mathbf{X})$, as a special application of local linear weighted least squares (Henderson and Ullah, 2005).

A fourth approach is akin to spatial filtering, and purports to model unspecified spatial spillover effects non-parametrically, in a so-called smooth spatial effects (SSE) estimator. In Gibbons and Machin (2003), the model considered is:

$$y_i = \mathbf{x}_i \beta + g(c_i) + \varepsilon_i \qquad (14.34)$$

where g is an unknown function, intended to capture all spatial correlation, and c_i represents the location of i. The model is estimated by means of the classic two-step procedure suggested by Robinson (1988). In the SSE estimator, both the dependent variable and the explanatory variables are replaced by deviations from the conditional expectation, which is obtained as a spatial kernel smoother. OLS can be applied to the transformed regression to obtain consistent estimates for β, (for a recent application, see Day *et al*., 2004).

14.7. CONCLUSION

The methodological toolbox for spatial regression has reached a certain maturity when it comes to the classical linear regression model. However, much less has been accomplished beyond this context and the development of new models, estimation techniques and specification tests is a very active area of research, both in statistics as well as in econometrics. Given space constraints, it was impossible to review all these efforts in a comprehensive way, but it is hoped that through the references provided an entry into this field has been facilitated.

Considerable theoretical research is ongoing to develop the formal conditions and proofs needed to obtain the asymptotic properties of estimators and tests in various settings. New techniques are being developed to deal with spatial effects in panel data, count models, probit and tobit, and other specifications that are the mainstay of applied empirical regression analysis. The growth in applications is encouraging as well, providing a greater empirical basis to document the importance of location and distance in explaining socioeconomic phenomena. Lastly, while in the past the lack of software may have been an impediment to the dissemination of spatial regression methods, this is no longer the case, as attested by several active open source developments (for a recent review, see Anselin, 2005, pp. 101–106).

REFERENCES

Andrews, D.W. (2005). Cross-section regression with common shocks. *Econometrica*, **73**: 1551–1585.

Anselin, L. (1980). *Estimation Methods for Spatial Autoregressive Structures*. Regional Science Dissertation and Monograph Series, Cornell University, Ithaca, New York.

Anselin, L. (1988a). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, **20**: 1–17.

Anselin, L. (1988b). *Spatial Econometrics: Methods and Models*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Anselin, L. (1992). Space and applied econometrics. Introduction. *Regional Science and Urban Economics*, **22**: 307–316.

Anselin, L. (2001a). Rao's score test in spatial econometrics. *Journal of Statistical Planning and Inference*, **97**: 113–139.

Anselin, L. (2001b). Spatial econometrics. In: Baltagi, B. (ed.), *A Companion to Theoretical Econometrics*, pp. 310–330. Oxford: Blackwell.

Anselin, L. (2002). Under the hood. Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, **27**(3): 247–267.

Anselin, L. (2003). Spatial externalities. *International Regional Science Review*, **26**(2): 147–152.

Anselin, L. (2005). Spatial statistical modeling in a GIS environment. In: Maguire, D.J., Batty, M. and Goodchild, M.F. (eds), *GIS, Spatial Analysis and Modeling*, pp. 93–111. Redlands, CA: ESRI Press.

Anselin, L. (2006). Spatial econometrics. In Mills, T. and Patterson, K. (eds), *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory*, pp. 901–969. Basingstoke: Palgrave Macmillan.

Anselin, L. and Bera, A. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In: Ullah, A. and Giles, D.E. (eds), *Handbook of Applied Economic Statistics*, pp. 237–289. New York: Marcel Dekker.

Anselin, L., Bera, A., Florax, R.J. and Yoon, M. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, **26**: 77–104.

Anselin, L. and Florax, R.J. (1995a). *New Directions in Spatial Econometrics*. Berlin: Springer-Verlag.

Anselin, L. and Florax, R.J. (1995b). Small sample properties of tests for spatial dependence in regression models: Some further results. In Anselin, L. and Florax, R.J. (eds), *New Directions in Spatial Econometrics*, pp. 21–74. Berlin: Springer-Verlag.

Anselin, L., Florax, R.J. and Rey, S.J. (2004). *Advances in Spatial Econometrics. Methodology, Tools and Applications*. Berlin: Springer-Verlag.

Anselin, L. and Griffith, D.A. (1988). Do spatial effects really matter in regression analysis? *Papers, Regional Science Association*, **65**: 11–34.

Anselin, L. and Kelejian, H.H. (1997). Testing for spatial error autocorrelation in the presence of endogenous regressors. *International Regional Science Review*, **20**: 153–182.

Anselin, L., Le Gallo, J. and Jayet, H. (2008). Spatial panel econometrics. In: Matyas, L. and Sevestre, P. (eds), *The Econometrics of Panel Data, Fundamentals and Recent Developments in Theory and Practice* (3rd Edition), pp. 627–662. Berlin: Springer-Verlag.

Anselin, L. and Moreno, R. (2003). Properties of tests for spatial error components. *Regional Science and Urban Economics*, **33**(5): 595–618.

Anselin, L. and Rey, S.J. (1991). Properties of tests for spatial dependence in linear regression models. *Geographical Analysis*, **23**: 112–131.

Anselin, L. and Rey, S.J. (1997). Introduction to the special issue on spatial econometrics. *International Regional Science Review*, **20**: 1–7.

Arbia, G. (2006). *Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence*. Berlin: Springer-Verlag.

Baltagi, B.H. and Li, D. (2001a). Double length artificial regressions for testing spatial dependence. *Econometric Reviews*, **20**(1): 31–40.

Baltagi, B.H. and Li, D. (2001b). LM tests for functional form and spatial error correlation. *International Regional Science Review*, **24**(2): 194–225.

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman & Hall/CRC.

Barry, R.P. and Pace, R.K. (1999). Monte Carlo estimates of the log determinant of large sparse matrices. *Linear Algebra and its Applications*, **289**: 41–54.

Bartels, C. and Ketellapper, R. (1979). *Exploratory and Explanatory Analysis of Spatial Data*. Boston: Martinus Nijhoff.

Basile, R. and Gress, B. (2005). Semi-parametric spatial auto-covariance models of regional growth in Europe. *Région et Développement*, **21**: 93–118.

Bera, A. and Yoon, M.J. (1993). Specification testing with misspecified alternatives. *Econometric Theory*, **9**: 649–658.

Beron, K.J., Murdoch, J.C., and Vijverberg, W.P. (2003). Why cooperate? Public goods, economic power, and the Montreal Protocol. *The Review of Economics and Statistics*, **85**(2): 286–297.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, **36**: 192–225.

Brock, W. and Durlauf, S. (2001). Discrete choice with social interactions. *Review of Economic Studies*, **59**: 235–260.

Brueckner, J.K. (2003). Strategic interaction among governments: An overview of empirical studies. *International Regional Science Review*, **26**(2): 175–188.

Burridge, P. (1980). On the Cliff–Ord test for spatial autocorrelation. *Journal of the Royal Statistical Society B*, **42**: 107–108.

Chen, X. and Conley, T.G. (2001). A new semiparametric spatial model for panel time series. *Journal of Econometrics*, **105**: 59–83.

Cliff, A. and Ord, J.K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical Analysis*, **4**: 267–284.

Cliff, A. and Ord, J.K. (1973). *Spatial Autocorrelation*. London: Pion.

Cliff, A. and Ord, J.K. (1981). *Spatial Processes: Models and Applications*. London: Pion.

Conley, T.G. (1999). GMM estimation with cross-sectional dependence. *Journal of Econometrics*, **92**: 1–45.

Cook, D. and Pocock, S. (1983). Multiple regression in geographic mortality studies, with allowance for spatially correlated errors. *Biometrics*, **39**: 361–371.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: Wiley.

Das, D., Kelejian, H.H. and Prucha, I.R. (2003). Finite sample properties of estimators of spatial autoregressive models with autoregressive disturbances. *Papers in Regional Science*, **82**: 1–27.

Davidson, R. and MacKinnon, J.G. (1984). Model specification tests based on artificial regressions. *International Economic Review*, **25**: 485–502.

Davidson, R. and MacKinnon, J.G. (1988). Double-length artificial regression. *Oxford Bulletin of Economics and Statistics*, **50**: 203–217.

Day, B., Bateman, I. and Lake, I. (2004). Omitted locational variates in hedonic analysis: A semiparametric approach using spatial statistics. Working Paper 04–04, Center for Social and Economic Research on the Global Environment (CSERGE), University of East Anglia, UK.

Dubin, R. (1988). Estimation of regression coefficients in the presence of spatially autocorrelated errors. *Review of Economics and Statistics*, **70**: 466–474.

Dubin, R., Pace, R.K. and Thibodeau, T.G. (1999). Spatial autoregression techniques for real estate data. *Journal of Real Estate Literature*, **7**: 79–95.

Durlauf, S.N. (2004). Neighborhood effects. In: Henderson, J. and Thisse, J.-F. (eds), *Handbook of Regional and Urban Economics, Volume 4*, pp. 2173–2242. Amsterdam: North Holland.

Elhorst, J.P. (2001). Dynamic models in space and time. *Geographical Analysis*, **33**: 119–140.

Elhorst, J.P. (2003). Specification and estimation of spatial panel data models. *International Regional Science Review*, **26**(3): 244–268.

Ellner, S.P. and Seifu, Y. (2002). Using spatial statistics to select model complexity. *Journal of Computational and Graphical Statistics*, **11**: 348–369.

Fischer, M.M., Reismann, M. and Scherngell, T. (2006). From conventional to spatial econometric models of spatial interaction. Paper presented at the Fifth International Workshop on Spatial Econometrics and Statistics, Rome, Italy, May 2006.

Fleming, M. (2004). Techniques for estimating spatially dependent discrete choice models. In: Anselin, L., Florax, R.J. and Rey, S.J. (eds), *Advances in Spatial Econometrics*, pp. 145–168. Heidelberg: Springer-Verlag.

Florax, R. and Folmer, H. (1992). Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators. *Regional Science and Urban Economics*, **22**: 405–432.

Florax, R.J. and de Graaff, T. (2004). The performance of diagnostic tests for spatial dependence in linear regression models: A meta-analysis of simulation studies. In Anselin, L., Florax, R.J. and Rey, S.J. (eds), *Advances in Spatial Econometrics. Methodology, Tools and Applications*, pp. 29–65. Berlin: Springer-Verlag.

Florax, R.J., Folmer, H. and Rey, S.J. (2003). Specification searches in spatial econometrics: The relevance of Hendry's methodology. *Regional Science and Urban Economics*, **33**(5): 557–579.

Florax, R.J., Folmer, H. and Rey, S.J. (2006). A comment on specification searches in spatial econometrics: The relevance of Hendry's methodology: A reply. *Regional Science and Urban Economics*, **36**: 300–308.

Florax, R.J.G.M. and van der Vlist, A. (2003). Spatial econometric data analysis: moving beyond traditional models. *International Regional Science Review*, **26**(3): 223–243.

Fortin, M.-J. and Dale, M. (2005). *Spatial Analysis: A Guide for Ecologists*. Cambridge: Cambridge University Press.

Getis, A., Mur, J. and Zoller, H.G. (2004). *Spatial Econometrics and Spatial Statistics*. London: Palgrave Macmillan.

Gibbons, S. and Machin, S. (2003). Valuing English primary schools. *Journal of Urban Economics*, **53**: 197–219.

Godfrey, L. (1981). On the invariance of the Lagrange Multiplier test with respect to certain changes in the alternative hypothesis. *Econometrica*, **49**: 1443–1455.

Goodchild, M.F., Anselin, L., Appelbaum, R. and Harthorn, B. (2000). Toward spatially integrated social science. *International Regional Science Review*, **23**(2): 139–159.

Gotway, C.A. and Stroup, W.W. (1997). A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological and Environmental Statistics*, **2**(2): 157–178.

Gotway, C.A. and Wolfinger, R.D. (2003). Spatial prediction of counts and rates. *Statistics in Medicine*, **22**: 1415–1432.

Gress, B. (2004a). *Semi-Parametric Spatial Autocovariance Models*. PhD thesis, University of California, Riverside, CA.

Gress, B. (2004b). Using semi-parametric spatial autocorrelation models to improve hedonic housing price prediction. Working paper, Department of Economics, University of California, Riverside, CA.

Griffith, D.A. (1988). *Advanced Spatial Statistics*. Dordrecht: Kluwer Academic.

Griffith, D.A. and Sone, A. (1995). Trade-offs associated with normalizing constant computational simplications for estimating spatial statistical models. *Journal of Statistical Computation and Simulation*, **51**: 165–183.

Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge: Cambridge University Press.

Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.

Hall, P. and Patil, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields*, **99**: 399–424.

Henderson, D.J. and Ullah, A. (2005). A nonparametric random effects estimator. *Economics Letters*, **88**: 403–407.

Hendry, D.F. (2006). A comment on specification searches in spatial econometrics: The relevance of Hendry's methodology. *Regional Science and Urban Economics*, **36**: 309–312.

Horowitz, J.L. and Lee, S. (2002). Semiparametric methods in applied econometrics: Do the models fit the data? *Statistical Modelling*, **2**: 3–22.

Kelejian, H.H. and Prucha, I. (1998). A generalized spatial two stage least squares procedures for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics*, **17**: 99–121.

Kelejian, H.H. and Prucha, I. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, **40**: 509–533.

Kelejian, H.H. and Prucha, I. (2001). On the asymptotic distribution of the Moran *I* test statistic with applications. *Journal of Econometrics*, **104**(2): 219–257.

Kelejian, H.H. and Prucha, I.R. (2002). 2SLS and OLS in a spatial autoregressive model with equal spatial weights. *Regional Science and Urban Economics*, **32**(6): 691–707.

Kelejian, H.H. and Prucha, I.R. (2004). Estimation of simultaneous systems of spatially interrelated cross sectional equations. *Journal of Econometrics*, **118**: 27–50.

Kelejian, H.H. and Prucha, I.R. (2005). HAC estimation in a spatial framework. Working paper, Department of Economics, University of Maryland, College Park, MD.

Kelejian, H.H. and Prucha, I.R. (2007). HAC estimation in a spatial framework. *Journal of Econometrics*, **140**: 131–154.

Kelejian, H.H. and Prucha, I.R. (2006). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. Working paper, Department of Economics, University of Maryland, College Park, MD.

Kelejian, H.H., Prucha, I.R. and Yuzefovich, Y. (2004). Instrumental variable estimation of a spatial autoregressive model with autoregressive disturbances: Large and small sample results. In: LeSage, J.P. and Pace, R.K. (eds), *Advances in Econometrics: Spatial and Spatiotemporal Econometrics*, pp. 163–198. Oxford, UK: Elsevier Science Ltd.

Kelejian, H.H. and Robinson, D.P. (1992). Spatial autocorrelation: A new computationally simple test with an application to per capita county police expenditures. *Regional Science and Urban Economics*, **22**: 317–333.

Kelejian, H.H. and Robinson, D.P. (1993). A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. *Papers in Regional Science*, **72**: 297–312.

Kelejian, H.H. and Robinson, D.P. (1995). Spatial correlation: A suggested alternative to the autoregressive model. In: Anselin, L. and Florax, R.J. (eds), *New Directions in Spatial Econometrics*, pp. 75–95. Berlin: Springer-Verlag.

Kelejian, H.H. and Robinson, D.P. (1998). A suggested test for spatial autocorrelation and/or heteroskedasticity and corresponding Monte Carlo results. *Regional Science and Urban Economics*, **28**: 389–417.

Kelejian, H.H. and Robinson, D.P. (2004). The influence of spatially correlated heteroskedacity on tests for spatial correlation. In: Anselin, L. and Florax, R.J. (eds), *Advances in Spatial Econometrics*, pages 79–97. Heidelberg: Springer-Verlag.

King, M. (1981). A small sample property of the Cliff–Ord test for spatial correlation. *Journal of the Royal Statistical Association B*, **43**: 264.

Land, K. and Deane, G. (1992). On the large-sample estimation of regression models with spatial or network-effect terms: A two stage least squares approach. In: Marsden, P. (ed.), *Sociological Methodology*, pp. 221–248. San Francisco: Jossey-Bass.

Lee, L.-F. (2002). Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric Theory*, **18**(2): 252–277.

Lee, L.-F. (2003). Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews*, **22**: 307–335.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, **72**: 1899–1925.

Lee, L.-F. (2006). GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*. Forthcoming.

Lee, L.-F. (2007). GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, **137**: 489–514.

Leenders, R.T.A.J. (2002). Modeling social influence through network autocorrelation: Constructing the weights matrix. *Social Networks*, **24**: 21–47.

LeSage, J.P. (2000). Bayesian estimation of limited dependent variable spatial autoregressive models. *Geographical Analysis*, **32**: 19–35.

LeSage, J.P. and Pace, R.K. (2004). *Advances in Econometrics: Spatial and Spatiotemporal Econometrics*. Oxford, UK: Elsevier Science Ltd.

LeSage, J.P. and Pace, R.K. (2005). Spatial econometric modeling of origin-destination flows. Paper Presented at the *52nd North American Meeting for the Regional Science Association International*, Las Vegas, NV, Nov. 2005.

LeSage, J.P., Pace, R.K. and Tiefelsdorf, M. (2004). Methodological developments in spatial econometrics and statistics. *Geographical Analysis*, **36**: 87–89.

Lin, X. and Lee, L.-F. (2005). GMM estimation of spatial autoregressive models with unknown heteroskedasticity. Working paper, The Ohio State University, Columbus, OH.

Magnus, J. (1978). Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix. *Journal of Econometrics*, **7**: 281–312. Corrigenda, *Journal of Econometrics* **10**: 261.

Mardia, K. and Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**: 135–146.

Martin, R. (1993). Approximations to the determinant term in Gaussian maximum likelihood estimation of some spatial models. *Communications in Statistics: Theory and Methods*, **22**: 189–205.

Moran, P.A. (1948). The interpretation of statistical maps. *Biometrika*, **35**: 255–260.

Moran, P.A. (1950). A test for the serial dependence of residuals. *Biometrika*, **37**: 178–181.

Nelson, G.C. (2002). Introduction to the special issue on spatial analysis. *Agricultural Economics*, **27**(3): 197–200.

Newey, W.K. and West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55**: 703–708.

Ord, J.K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, **70**: 120–126.

Pace, R.K. and Barry, R. (1997a). Quick computation of spatial autoregressive estimators. *Geographical Analysis*, **29**: 232–246.

Pace, R.K. and Barry, R. (1997b). Sparse spatial autoregressions. *Statistics and Probability Letters*, **33**: 291–297.

Pace, R.K., Barry, R. and Sirmans, C. (1998). Spatial statistics and real estate. *Journal of Real Estate Finance and Economics*, **17**: 5–13.

Pace, R.K. and LeSage, J.P. (2002). Semiparametric maximum likelihood estimates of spatial dependence. *Geographical Analysis*, **34**: 76–90.

Pace, R.K. and LeSage, J.P. (2004a). Chebyshev approximation of log-determinants of spatial weights matrices. *Computational Statistics and Data Analysis*, **45**: 179–196.

Pace, R.K. and LeSage, J.P. (2004b). Spatial statistics and real estate. *Journal of Real Estate Finance and Economics*, **29**: 147–148.

Paelinck, J. and Klaassen, L. (1979). *Spatial Econometrics*. Farnborough: Saxon House.

Pesaran, M.H. (2005). Estimation and inference in large heterogenous panels with cross section dependence. Working paper, Faculty of Economics and Politics, University of Cambridge, Cambridge, United Kingdom.

Pinkse, J. and Slade, M.E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. *Journal of Econometrics*, **85**: 125–154.

Pinkse, J., Slade, M.E. and Brett, C. (2002). Spatial price competition: A semiparametric approach. *Econometrica*, **70**(3): 1111–1153.

Rey, S.J. and Boarnet, M.G. (2004). A taxonomy of spatial econometric models for simultaneous equations systems. In Anselin, L., Florax, R.J. and Rey, S.J. (eds), *Advances in Spatial Econometrics*. pp. 99–119, Heidelberg: Springer-Verlag.

Ripley, B.D. (1981). *Spatial Statistics*. New York: Wiley.

Robinson, P.M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, **56**: 931–954.

Schabenberger, O. and Gotway, C.A. (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Schmoyer, R. (1994). Permutation tests for correlation in regression errors. *Journal of the American Statistical Association*, **89**: 1507–1516.

Smirnov, O. (2005). Computation of the information matrix for models with spatial interaction on a lattice.

Journal of Computational and Graphical Statistics, **14**: 910–927.

Smirnov, O. and Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: A characteristic polynomial approach. *Computational Statistics and Data Analysis*, **35**: 301–319.

Tiefelsdorf, M. (2002). The saddlepoint approximation of Moran's I and local Moran's I_i's reference distribution and their numerical evaluation. *Geographical Analysis*, **34**: 187–206.

Tiefelsdorf, M. and Boots, B. (1995). The exact distribution of Moran's I. *Environment and Planning A*, **27**: 985–999.

Upton, G.J. and Fingleton, B. (1985). *Spatial Data Analysis by Example. Volume 1: Point Pattern and Quantitative Data*. New York: Wiley.

Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: John Wiley.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**: 817–838.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, **41**: 434–449.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, **56**: 129–136.

15

Spatial Microsimulation

D. Ballas and G.P. Clarke

15.1. INTRODUCTION

Much modelling in human geography and related disciplines takes an aggregate or meso-scale approach to the issue of spatial resolution. That is, characteristics of individuals or households are summed to provide zonal population or demand totals and, if appropriate, individual companies or firms are similarly aggregated on the supply side of the economy. In spatial econometrics or regional science those zones can be as large as entire cities or regions. The most obvious reason for doing this is that detailed disaggregate data on persons or firms are typically not regularly available below the level of the region (especially economic or household survey data). In most countries census data are available to help disaggregate population totals to smaller geographical regions but the level of detail available for researchers is then limited by what is published in two- or three-dimensional tables (although special requests for different combinations can be made in certain countries but at additional expense). Models built on these more aggregate data sets are widespread and have proved very fruitful in many areas of policy analysis (see, for example, Fotheringham *et al*., 2000; Longley and Batty 2002; Stillwell and Clarke 2004). However, such modelling techniques often need to be highly disaggregated for real world applications and they also provide very little information concerning the *interdependencies* between household structure or type and their lifestyles, including the events they routinely participate in and hence their ability to raise and spend various types of income and wealth. For social policy evaluation such micro models allow analysts to monitor the effects of changes in taxation, family credit, property or council tax, pensions, social security payments, etc. (the actions of local and national governments) at the household level (and hence at any

more aggregate spatial scale). For area-based policy evaluation such models allow differential impacts between and within areas to be analysed more effectively. The necessity of predicting the impacts of social and area-based policies at the local or micro-level has also been emphasized by Openshaw (1995, p.60). 'Governments need to predict the outcomes of their actions and produce forecasts at the local level'.

For these reasons Wilson (2000, p. 98) identified microsimulation as one the most important methods in regional science modelling: 'Simulation is a critical concept in the future development of modelling because it provides a way of handling complexity that cannot be handled analytically. Microsimulation is a valuable example of a technique that may have increasing prominence in future research'.

This chapter reviews the history of spatial microsimulation and spells out a research agenda for the further exploitation of the technique. First, the semantics of microsimulation are revisited and we describe the different types of microsimulation models and how they can be formulated (section 15.2). We then provide a brief overview of applications of microsimulation models which includes use in economics, social policy, geography and regional science (section 15.3). Then, we spell out a research agenda for spatial microsimulation (section 15.4) and offer some concluding comments in section 15.5.

15.2. WHAT IS SPATIAL MICROSIMULATION?

Microsimulation can generally be defined as a methodology that is concerned with the creation of large-scale population microdata sets to enable the analysis of policy impacts at the micro-level. The approach dates back to the work of Orcutt (1957) and Orcutt *et al*. (1961) who argued for a new type of socio-economic system and described a simple model of demographic transitions based on micro-analytical simulation. In particular, microsimulation methods aim to examine changes in the characteristics or lifestyles of individuals within households and to analyse the impact of government policy changes on these individuals or households. Microsimulation models can be distinguished between two main types. First, there are *static models* that are based on simple snapshots of the current circumstances of a sample of the population at any one time. Second, there are *dynamic models* that vary or age the attributes of each micro-unit in a sample to build up a synthetic longitudinal database forecasting the sample members' lifestyles into the future.

The first geographical application of microsimulation was developed by Hägerstrand (1967) who employed micro-analytical techniques for the study of spatial diffusion of innovation. Nevertheless, it can be argued that the basis for *spatial microsimulation of households and individuals* was founded in the 1970s. In particular, Wilson and Pownall (1976) were among the first to address the aggregation difficulties that were associated with traditional comprehensive spatial models of urban systems. They suggested a new spatial modelling framework for representing the urban system based on the micro-level interdependence of household and individual characteristics. Further, they concentrated on the spatial distribution of population and its activities and suggested that persons and their associated attributes should be defined separately in the form of lists, rather than represented in the form of matrices. In this manner, there is no loss of information and the storage is computationally efficient. In their representational framework, they were interested in estimating all the characteristics of the individuals that

comprise the urban population. Formally, they defined variables for each person in the system separately by adding a person label r to each person attribute $x_1, x_2, x_3, \ldots, x_n$. The person attributes would therefore become $x_1^r, x_2^r, x_3^r, \ldots, x_n^r$ for the *rth* person of the population. This means that if there are M people in the population, there will be $N * M$ variables in total. After suggesting the above framework for representing individuals, Wilson and Pownall (1976) proposed a modelling procedure to estimate each characteristic for each person in turn. They formally expressed this procedure as follows:

$$x_j^r = (x_j^r\ (P_j(x/\ldots)R_j^r, \tau)$$

where $P_j(x/\ldots)$ is the probability of x_j taking the value x conditional on variables yet to be specified, R_j^r is a random number selected for person r and characteristic j, and τ represents a relevant constraint set (Wilson and Pownall, 1976). One of the most significant properties of the above model is its *causal structure*, which is largely reflected in the order in which the characteristics are estimated for each person.

Almost a decade later, Birkin and Clarke (1988) built a synthetic spatial information system for urban and regional analysis. It can be argued that this model is the first comprehensive spatial microsimulation model in the UK. Birkin and Clarke (1988) discussed the difficulties of performing micro-level spatial analysis using the existing published data sources and they proposed a methodology for generating synthetic micro-data from a number of different aggregate sources. This microsimulation methodology was underpinned by a technique known as iterative proportional fitting (IPF) (see Birkin and Clarke (1988) and Ballas (2001) for a more detailed discussion of this technique). Birkin and Clarke (1988) briefly discuss the theoretical properties of IPF and they demonstrate how they applied the method to estimate joint probability distributions of household attributes. The IPF procedure adopts a synthetic reconstruction method which calculates conditional probabilities of having particular attributes and it then assigns these attributes on the basis of random sampling procedures (Monte Carlo simulation). Table 15.1 depicts the steps that need to be followed in the procedure for allocating economic activity status for example.

More recently researchers have argued that reweighting existing survey data can produce more robust results than these synthetic probabilistic reconstruction models, which involve the use of random sampling (Williamson *et al*., 1998; Huang and Williamson, 2001; Ballas *et al*., 2005). Two well-used reweighting procedures are:

- Reweighting probabilistic approaches, which typically reweight an existing national microdata set to fit a geographical area description on the basis of random sampling and optimization techniques
- Reweighting deterministic approaches, which reweight a non geographical population microdata set to fit small area descriptions, but *without* the use of random sampling procedures

These new methods involve the reweighting of an existing microdata sample (which is usually only available at coarse levels of geography), so that it would fit small area population statistics tables. For instance, an existing microdata set such as the British Household Panel Survey (BHPS) described in Table 15.2 can be reweighted to 'populate' small areas.

The BHPS provides a detailed record for a sample of households and all of their occupants (Taylor *et al*. 2001). Reweighting methods aim to sample from all the microdata

Table 15.1 Microsimulation procedure for the allocation of *economic activity* status (after the similar example of tenure allocation procedure given by Clarke, 1996: 3)

Head of household (*hh*)

Steps	*1st*	*2nd*	...	*Last*
Age, sex and marital status and location (ED level) (given)	Age: 16–29 Sex: Male Marital status: SWD GeoCode: DAFA01	Age: 75–84 Sex: Female Marital status: married GeoCode: DAFA02	...	Age: 30–44 Sex: Male Marital status: married GeoCode: DAGK45
Probability of *hh* of given age, sex, and location (ED level) being economically active	0.7	0.4	...	0.7
Random number	0.55	0.5	...	0.45
Economic activity assigned to *hh* on the basis of random sampling	Economically active	Economically inactive	...	Economically active
Probability of economically active *hh* being an employee	0.6		...	0.5
Probability of economically active *hh* given age, sex, marital staus, and location (ED level) being self-employed	0.2		...	0.3
Probability of economically active *hh* given age, sex, marital staus, and location (ED level) being on a government scheme	0.05		...	0.15
Probability of economically active *hh* given age, sex, marital staus, and location (ED level) being unemployed	0.15		...	0.05
Random number	0.4		...	0.6
Economic activity category assigned on the basis of random sampling	Employee		...	Self-employed

records to find the set of household records that best matches the population described in the Small Area Statistics or Census Area Statistics tables for the small area under study. First, a series of small area tables (e.g., from the Census or other sources) that describe the small area of interest must be selected. For example, a reweighting method would sample from the BHPS to find a suitable combination of households that would fit the data described in Table 15.3.

This first stage of population estimation at the household level is primarily a data-fitting exercise. However once built the model can be used for *static what-if* simulations, in which the impacts of alternative policy scenarios on the population are estimated: for instance if there had been no poll tax in 1991 which communities would have benefited most and which would have had to have paid more tax in other forms? Second it can be used for dynamic modelling, to update a basic micro-dataset and future-oriented *what-if* simulations: for instance if the current government had raised income taxes in

Table 15.2 A population microdata example

Person	*AHID*	*PID*	*AAGE12*	*Sex*	*AJBSTAT*	...	*AHLLT*	*AQFVOC*	*ATENURE*	*AJLSEG*	...
1	1000209	10002251	91	2	4	...	1	1	6	9	...
2	1000381	10004491	28	1	3	...	2	0	7	−8	...
3	1000381	10004521	26	1	3	...	2	0	7	−8	...
4	1000667	10007857	58	2	2	...	2	1	7	−8	...
5	1001221	10014578	54	2	1	...	2	0	2	−8	...
6	1001221	10014608	57	1	2	...	2	1	2	−8	...
7	1001418	10016813	36	1	1	...	2	1	3	−8	...
8	1001418	10016848	32	2	−7	...	2	−7	3	−7	...
9	1001418	10016872	10	1	−8	...	−8	−8	3	−8	...
10	1001507	10017933	49	2	1	...	2	0	2	−8	...
11	1001507	10017968	46	1	2	...	2	0	2	−8	...
12	1001507	10017992	12	2	−8	...	−8	−8	2	−8	...

Note: The British Household Panel Survey data were made available through the UK Data Archive. The data were originally collected by the ESRC Research Centre on Micro-social Change at the University of Essex, now incorporated within the Institute for Social and Economic Research.

Person Person number.
AHID Household identifier (number of household to which the listed individual belongs).
PID Person identifier (a unique number to identify the individual).
AAGE12 Age at 1/12/1991.
Sex Sex
AJBSTAT Current labour force status (e.g., self-employed, in paid employment, unemployed, family care, etc.) in 1991.
AHLLT Health status in 1991.
AQFVOC Vocational qualifications in 1991.
AJBSEG Socio-economic group (e.g., employers, managers, professionals, skilled manual, unskilled, etc.) in 1991.
ATENURE Tenure status in 1991.
AJLSEG Socio-economic group: last job (in 1991).

Table 15.3 An example of small area data

Small area table 1 (household type)	*Small area table 2 (economic activity of household head)*	*Small area table 3 (tenure status)*
Area 1	*Area 1*	*Area 1*
60 Married couple households	70 Employed/ self-employed	60 Owner occupier
20 Single-person households	10 Unemployed	20 Local Authority or Housing Association
20 Other	20 Other	20 Rented privately
Area 2	*Area 2*	*Area 2*
40 Married couple households	50 Employed/ self-employed	60 Owner occupier
20 Single-person households	20 Unemployed	20 Local Authority or Housing Association
40 Other	30 Other	20 Rented privately

1997 what would the redistributive effects have been between different socio-economic groups and between central cities and their suburbs by 2007?

We shall explore applications based on these principles in the next section.

15.3. APPLICATIONS

15.3.1. Introduction

As mentioned above, microsimulation has been mainly developed and used by a

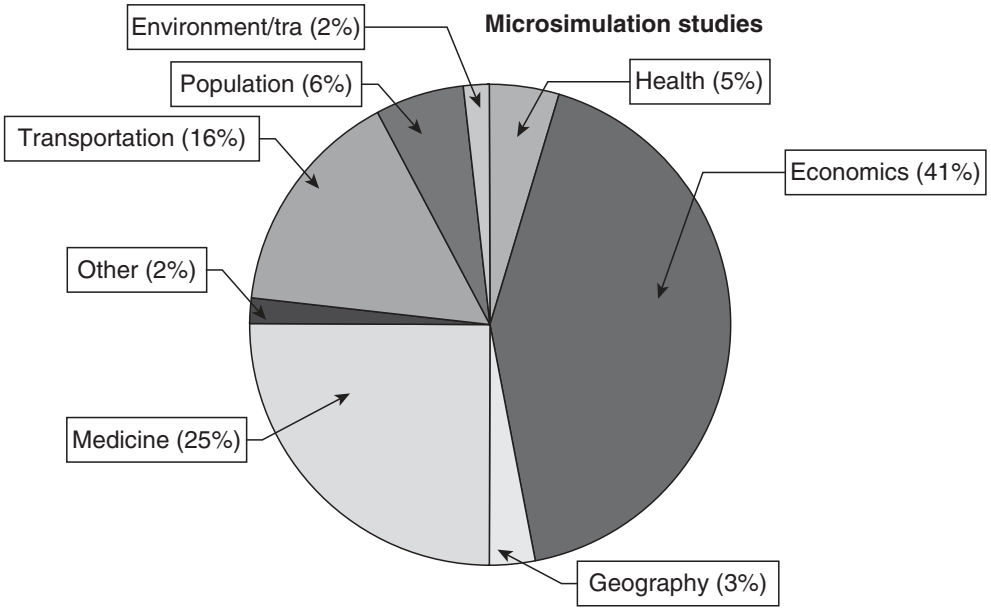


Figure 15.1 Distribution of microsimulation academic studies in the period 1967–2003. (Source: http://www.sciencedirect.com/; Accessed 15 October 2003; after Ballas *et al.*, 2005: p. 11).

variety of social sciences. Figure 15.1 shows the results of a basic keyword search in the *Sciencedirect* academic journal database, searching the word 'microsimulation' in the titles or abstracts of papers in the last 30 years. As can be seen, the majority of the papers were in economics (41%) with very few papers in geography (3%), although spatial applications may also lie in fields such as population, transport and health. There is also a relatively high number of microsimulation applications in medicine. However these are applications of a different nature, as their main focus is the effectiveness of medicines (e.g., simulating the impact of medicines on human well-being, etc.)

The rest of this section explores some well-known examples of microsimulation for certain types of policy work. This includes static models (simply run for one period of time) and dynamic models (where the attributes of the population are updated constantly or over yearly totals).

15.3.2. Tax and income modelling

A large number of papers in economics on microsimulation relate to work on household finance. Indeed, amongst the first applied microsimulation models was TAX, developed at the US Treasury department in the 1960s (Nelissen, 1993). Since then there have been many models built to examine the impacts on individual households of various tax or welfare changes (Bekkering, 1995; Falkingham and Lessof, 1992; Falkingham and Hills, 1995a, b; Glennerster *et al.*, 1995; Propper, 1995).

The first task in the modelling of household income is to link households with job type. Birkin and Clarke (1989) used the SYNTHESIS model to generate incomes for individuals. They used an IPF based microsimulation approach to estimate earned income at ward level for the Leeds Metropolitan District by assigning each household a job and an occupation using

information from the 'New Earnings Survey' to allocate an income variable accordingly. In addition, they estimated income from transfer payments such as the *Family Income Supplement* for each household. This was probably the first successful attempt to generate income at the small area level in the literature. Ballas and Clarke (2001a) extended this work by increasing the number of transfer or welfare payments included in the model (such as detailed work on child benefits) and also including household taxation levels. Williamson and Voas (2000) report ongoing research to provide more robust and reliable estimates of income at the small area level. They argue that income estimation at the small area level may be seen as a multilevel analysis problem where variables at individual and area levels may interact.

Work in the US has tended to extend such work to include not only household income but also household wealth. In particular, Caldwell and Keister (1996) present CORSIM, which is a dynamic microsimulation model that has been under development at Cornell University since 1986. CORSIM has been used to model wealth distribution in the United States over the historical period 1960–1995 and to forecast wealth distribution over the future (Caldwell and Keister, 1996). It is noteworthy that over 17 different national microdata files have been used to build the model, which incorporated 50 economic, demographic and social processes by means of approximately 900 stochastic equations and rule-based algorithms (*Ibid*.). Furthermore, Caldwell *et al*. (1998) review the geography of wealth in the USA and show how CORSIM has included many variables relating to assets and debts.

As mentioned in the previous section, microsimulation models can be even more powerful when they become *dynamic*. In particular, once a microsimulation database is built, dynamic microsimulation procedures can be introduced in order to update these databases. Amongst the first applied dynamic microsimulation models was DYNASIM (DYNAmic Simulation of Income Model; see Orcutt *et al*., 1961; Wertheimer *et al*., 1986), which was the base for later, more sophisticated, models such as CORSIM. One of the descendants of DYNASIM was DYNASIM2, which was developed and maintained at the Urban Institute in Washington D.C. (Wertheimer *et al*., 1986). DYNASIM2 comprised two sub-models: a *Family and Earnings History* (FEH) model and a *Jobs and Benefit History* (JBH) model (Wertheimer *et al*., 1986).

Work on income and taxation can be more focused onto particular problems. Currently in many Western countries there is a problem relating to pensions given that an ageing population will need more financial support from a declining workforce population. Notable here is the work of Hancock *et al*. (1992), who built PENSIM. This is a microsimulation model designed for the simulation of pensioners' incomes up to the year 2030. Hancock *et al*. (1992) point out that the simulation of pensions is another good example of the application of dynamic microsimulation techniques, given that pension rights accumulate over a long period of time and their estimation requires the processing of data pertaining to individuals' entire working lives. PENSIM aims at predicting aggregate income by source within certain subsets of the pensioner population under different alternative assumptions. These assumptions pertain to the rules controlling the treatment of pensioners by the social security system, pension entitlement regulations, projected demographic movements and movements in aggregate economic variables such as unemployment and inflation. Davies and Joshi (1992) also focused on modelling pensions. In particular, they employed microsimulation

modelling techniques to simulate lifetime earning and pension entitlements in Britain. They used a microsimulation model to construct illustrative individuals and examine the treatment of pensions after divorce. They also modelled lifetime earnings upon which pensions depend and they simulated dated earnings for each partner before and after dissolution of the marriage and they explored how pension entitlement varied with the duration of the marriage. Among the variables that they estimated were age, sex, age at marriage, qualifications and age at divorce. Models of income and wealth also feed significantly into models of social policy change (see section 15.3.4).

15.3.3. Modelling household activity patterns

Introduction

Wilson and Pownall (1976) provided an early example of how microsimulation models could be employed to build urban micro-analytical models based on the interdependencies between individual characteristics. In these examples they investigated the interdependency of the person and household characteristics that are listed in Table 15.4.

Table 15.4 Attributes of individual micro-unit examined by Wilson and Pownall (1976)

Person Attributes
Wage
Job location
Residential location
Journey to work costs
Housing expenditure
Shopping expenditure
Journey to shop costs
Shopping location
Other expenditure

As can be seen, this framework starts to model activity patterns of individuals or households (activity normally undertaken in more meso-scale models). Hence, there have been a number of examples of building links between these household data sets and trip making behaviour or activities. We shall explore a sample of these types of application in this section. Simulated small area micro data sets can also allow for a household demand function to be specified (likely type of supermarket, school, etc.) at the small area level given that household's socio-economic profile. This can then be fed into a household interaction model (or variant of discrete choice model) in order to add place of work, shopping destination, GP location, children's school, etc. to the household database contained within a spatial microsimulation model.

Labour and housing markets

As Table 15.4 suggests, one key link is between households and their job locations. By adding a journey to work model households can be allocated a job destination (by age, sex, occupation, social class, etc.). Ballas and Clarke (2001b) showed how it was possible to build a journey to work model for Leeds which linked individual households to particular firms. Then, the impacts of the closure of a major manufacturing firm in east Leeds could be modelled in terms of which households would be most affected and in terms of their consequent reduction in income and expenditure. This 'local' analysis showed that most of the impacts occurred within 5 miles of the firm's location – analysis in stark contrast to outputs from typical regional input–output models.

Hooimeijer (1996) suggests a geographical microsimulation framework to analyse the linkages between supply and demand in the housing market and labour market simultaneously. He argues for the modelling

of spatial mobility of households and firms in three different time sets (daily commuting, relocation, and lifetime mobility). The problem associated with this type of modelling is the order in which processes are modelled. It could be argued for example, that labour force participation is dependent on family status and attributes or that the family formation procedure is dependent on the labour market situation of each individual. As Falkingham and Lessof (1992) put it:

> ... while a woman's labour force status can depend on the number of children she has and on her marital status, it cannot also influence the probability of the woman having a child in any year. The ordering of the modules necessarily involves making assumptions about the direction of causality in relationships between variables. (Falkingham and Lessof, 1992: 9)

Their LIFEMOD model is based on the assumption that demographic variables determine labour-force participation and that labour-force participation influences health, although it is pointed out that evidence suggests causality in either direction (Falkingham and Lessof, 1992).

Transport and land-use models

Wegener and Spiekermann (1996) explore the potential of microsimulation for urban models, focusing on land-use and travel models. They argue that a new generation of travel models has emerged which requires more detailed information on household demographics and employment characteristics at the small area level. They also point out that there are new neighbourhood-scale transport policies aimed at promoting public transport, walking and cycling. These policies require detailed information on the precise location of the population and its activities. Wegener and Spiekermann (1996) also stress the need for urban models to predict not only the economic but also the environmental impacts of land-use transport policies. In order to model the environmental impacts there is a need for small-area forecasts of emissions from stationary and mobile sources as well as of emissions in terms of the affected population. After outlining the main characteristics of a micro-analytic theory of urban change, Wegener and Spiekermann (1996) report on modelling efforts carried out at the University of Dortmund to integrate microsimulation into a comprehensive urban land-use transport model (see also Veldhuisen *et al*. (2000).

The links between households, housing markets and labour markets have been explored more recently in Ballas *et al*. (2005).

Retail models

Traditional spatial interaction or discrete choice models have been used to estimate expenditure flows from households to each store. It is argued by Nakaya *et al*. (2005) that it is possible to improve the applicability of the retail interaction model, not by increasing the complexity of the model formulation, but by integrating the interaction modelling framework with spatial microsimulation. To attain a high level of predictive accuracy, models of retail interaction usually require a high degree of disaggregation (Birkin *et al*., 2002). Even if a survey of consumer behaviour is conducted by randomly distributing a questionnaire to local residents, response rates would vary by consumer type and place of residence based on people's different levels of interest and tolerance of such a survey. Consequently, survey data of this type often contain bias in the type of consumer behaviour measured, swayed towards the behaviour of individuals who least object to completing surveys. This problem of missing data tends to get worse as the spatial units used in the analysis get smaller. A solution to this problem is to generate data through spatial microsimulation which can be

used to generate estimates of expenditure on groceries by each household. These estimates can then be aggregated to any grouping including lifestyle segments and residential zones simultaneously. The end product is a retail model with a more disaggregate and useful set of demand variables and both attractiveness and distance decay parameters calibrated for different types of consumers.

15.3.4. Social policy change

From the end of the 1960s onwards microsimulation became the dominant quantitative method for forecasting the impacts of policy changes in the social welfare area in the USA (Nelissen, 1993). This is the same now in many developed countries. A good example of dynamic microsimulation modelling for economic and social policy analysis is NEDYMAS (Netherlands Dynamic Micro-Analytic Simulation model; see Nelissen, 1993). The latter is a dynamic cross-sectional microsimulation model aimed at simulating future social security benefits and contributions. In particular, NEDYMAS is a comprehensive model for the Dutch household sector and comprises three main modules: a demographic module, a labour market and income formation module, and a social security module. Demographic processes are simulated explicitly, which means that the size of the microdata base changes during the simulation period. The NEDYMAS microdatabase included 204 household attributes. Once the initial population has been determined the attributes of each individual can be updated and the micro-population can be projected into the future. First, all demographic transitions are made in the model. These include events such as birth, death, immigration, family reunification, emigration, first marriage, remarriage, cohabitation, divorce, etc. Once all the demographic transitions are simulated, the next step is to consider labour market transitions. These include education, scholarship, transitions from school, transitions from being unemployed, retirement, etc. The final step in the NEDYMAS microsimulation procedure is to simulate attributes or transitions that are related to social security. Nelissen (1993) describes how sensitivity analysis was performed to validate NEDYMAS and concludes that the model is capable of reconstructing the long-term socio-economic development at the micro level.

It is interesting to note that the LIFEMOD model described above has also been used to estimate the effects of the welfare state over the life-cycle of individuals (Falkingham and Hills, 1995a, b, Falkingham *et al*., 1995), as well as to estimate the degree to which income is redistributed between people over time, or across the life cycle (Fakingham and Hills, 1995b). It has also been used to investigate financing options for higher education (Glennerster *et al*., 1995) and to examine the dynamics of lone parenthood (Evandrou and Falkingham, 1995). Further, LIFEMOD has been used to explore the lifetime distribution of health needs and use of health services (Propper, 1995).

In the UK the work of Holly Sutherland and her colleagues has been very influential in terms of policy analysis using microsimulation (Redmond *et al*., 1998; Mitton *et al*., 2000; Hancock, 2000; Sutherland *et al*., 2003). Sutherland and Piachaud (2001) for example, developed and used a microsimulation methodology for the assessment of British government policies for the reduction of child poverty in the period 1997–2001. Their results suggest that the number of children in poverty will be reduced by approximately one-third in the short term and that there is a trend towards further reductions. However, they emphasized that there is a need for more measures in order to meet the government target of abolishing child poverty in a generation.

Another example is the research conducted by Ballas *et al*. (2005) using SIMBRITAIN. This model assumes that the initial simulation of the future population of Britain could be based on population projections (such as those of the ONS) and on the assumption that the trends in the changes to society to 2021 are similar to that of the previous decade. However, alternative projections would also be provided on the basis of *hypothetical social policy changes*. They also examined child poverty as a major application area. For example, it is possible to use a dynamic spatial microsimulation model to estimate the degree of child poverty eradication within the next 20 years under different policies and assumptions, such as the onset of a major recession or a redistribution of wealth, and the model would provide projections in order to suggest where current strategies are failing to eradicate child poverty within a generation.

Microsimulation still has to gain credibility amongst the social science community in general and social policy researchers in particular. Thus, there is currently a major challenge to build on the work described above in order to project the population into the future to predict what would happen under different macro-economic, micro-economic and social policy scenarios. This will enable an evaluation of the short and long-term impacts that various government policies are likely to have on different segments of society and different geographical areas.

15.4. THE WAY FORWARD: THE RESEARCH AGENDA

15.4.1. Towards a comprehensive spatial microsimulation of urban systems

We have seen in section 15.3 that progress has been made on adding behavioural or trip making models into microsimulation. The obvious next step is to link all these components into a more comprehensive urban model. First, more linkage is required between households and the supply-side of the economy. For example it should be possible to link all households to a retail destination (by type of good) and a destination for primary and secondary health and education. By adding more information on linkages or flows within the city it can be argued that such modelling would offer major new insights into urban deprivation or quality of life. Many households will be identified as having poor accessibility to major services. However, multiple deprivation may well exist in many areas where poor accessibility exists to all major urban services. For example, a neighbourhood may be a long way from decent retail opportunities, a hospital and a GP. In addition, although close to a secondary school, that school may be suffering from very low examination success and hence access is constrained to only a poor-performing school.

Once all the relevant demand-side and supply-side databases are constructed, the next step would be to perform *what-if* policy impact analysis. In particular, it will be possible to model what would be the impact on the quality of life of residents in different localities, under different scenarios. For instance, it would be possible to estimate what would be the socio-economic and spatial impact of a new hospital in an area, new retail facilities, new schools, etc. It will also be possible to link these activities to events taking place elsewhere in the city. For example, the impact analysis of the factory closure that has been given by Ballas and Clarke (2001b) can be extended by estimating multiplier effects and the loss of spending power in the local community. Further, it would be possible to estimate the downgrading of service facilities as businesses close or

relocate to more affluent areas. It would then be possible to determine whether this development leads to even poorer service provision for those communities affected. The possibility of individuals made redundant finding new jobs in the area, migrating or being retrained could also then be estimated. Ballas *et al*. (2006) have made a start in this direction.

The second major effort needed is to link such models more into the local and regional labour market through a framework which combines spatial microsimulation models and regional input–output models or regional econometric models. It has long been argued that the treatment of the household sector has been ignored by most input–output modellers who at best would model or aggregate variables such as household income and expenditure in aggregate form, making no distinction between the behaviour of different types of household defined in terms of socio-economic status, employment profile, skill level, etc. (Batey, 2003). It can be argued that spatial microsimulation can address this issue. For instance, the prediction of input–output models for different sectors of the local economy can be spatially disaggregated with the use of a spatial microsimulation model. Likewise, predictions of regional econometric models for the whole region can be disaggregated at the individual and household level with the use of spatial microsimulation. Jin and Wilson (1993) made some progress here but data limitations made it difficult to operationalize their models. Microsimulation potentially has the ability to provide much of that missing data.

15.4.2. Linking microsimulation and agent-based models

Microsimulation is closely linked to another type of individual level modelling: agent based models (ABM). ABM models are normally associated with the behaviour of multiple agents in a social or economic system. These agents usually interact constantly with each other and the environment they live or move within. Thus their actions are driven by certain rules. Although this methodology sounds similar to microsimulation (where agents could be the individuals within the households) Davidsson (2000) notes that ABM may offer a better framework for including behavioural rules into the actions of agents (including an element of random behaviour) and for allowing interactions between agents. There are a number of good illustrations in a geographical setting (Batty and Densham, 1996; Heppenstall *et al*., 2006). Clearly there is a research agenda to link these two complementary approaches more effectively. Microsimulation could be used to give the agents in ABM their initial characteristics and locations whilst ABM could then provide the capacity to model individual adaptive behaviours and emergence of new behaviours (see also the discussion of Boman and Holm, 2004). In addition, data from household panel surveys such as the British Household Panel Survey (BHPS) may be utilized to formulate plausible assumptions regarding these behaviours. For instance, it is possible to use panel data from surveys such as the BHPS to model the life paths of particular individuals and households who have moved into and out of work. Such data can also be combined with information from more qualitative analyses to simulate the behaviour of workers made redundant following plant closures and how they fare in adapting to the changing labour market and how long term unemployment is increased for those unable to retrain (Ballas *et al*., 2006). The findings of qualitative studies such these can provide useful insights when formulating the 'rules' that determine the likely behaviour of households

in a combined ABM/spatial microsimulation framework.

15.4.3. Spatial microsimulation and remote sensing

Another interesting research possibility is the combination of spatial microsimulation model outputs with remotely sensed data. One difficulty at present with spatial microsimulation models is that most of the probabilities are calculated from known distributions (provided by data sources such as the Census of Population) at the small area level (e.g., the Census Output Area (COA) level in the UK). That is, although estimations are made at the level of the individual household, it is not possible to know precisely where within a small area (such as the COA) a particular household is actually located. For many policy purposes that is not a major problem – it is the overall effect on the locality that is most important. However, it can be argued that for certain applications this would be a worthwhile addition – especially potential business applications.

Using remote sensing techniques it is possible to obtain a point data set of houses which would contain the housing type attribute. These point data sets can then be linked to spatially disaggregated microsimulated households in order to disaggregate the simulated population at the COA level. In other words, the task of this modelling exercise would be to populate the remotely sensed residential properties with attribute data. Table 15.5 lists the attributes that can be used as a link between the remote sensing generated database and the microsimulation output.

Further, Figure 15.2 depicts schematically, and in a simplified manner, the geographical databases that are typically generated by microsimulation models and remote sensing methodologies and how these can be linked. As can be seen, these databases can be joined on the basis of the fields that they have in common, such as the housing type and house size. However, it can be argued that all the attributes listed in Table 15.5 can be used to build an index of similarity between a remotely sensed house and a microsimulated synthetic household.

Table 15.5 Database attributes that can be used for the linkage

Spatial microsimulation output	*Remotely sensed data*
No. of residents in household (as a proxy to house size)	Land use
House type	Property size
Number of cars (as a proxy to house size)	House type
Number of rooms in household space (as a proxy to house size)	...
...	

Moreover, the linkage between the two databases can be achieved with the use of statistical matching or data fusion techniques. It should be noted that although statistical matching (also known as data fusion) has a relatively long history, its theoretical basis is somewhat narrow and there is no established, tested and widely applied methodology (Paas, 1986; Sutherland *et al.*, 2002).

The new framework would also offer further potential for calibration and for dynamic modelling. The visualization of the area being modelled would provide useful additional diagnostic information and would allow new comparisons to be made between simulated households and real households. New images obtained from remote sensing may provide a very valuable additional source of information, highlighting new construction, demolitions and major changes in land use types. The second major benefit of

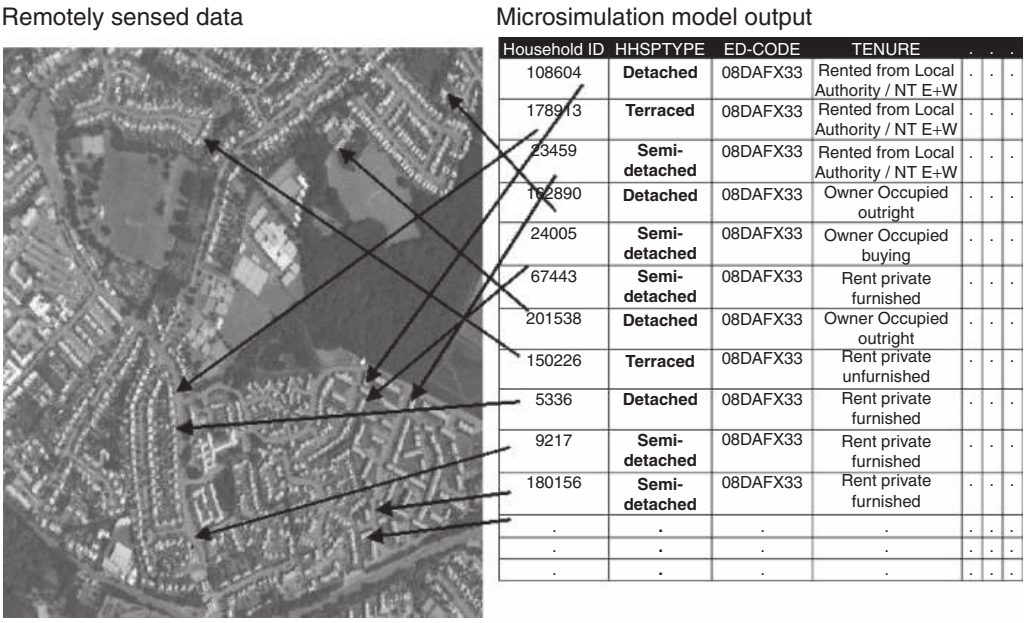


Figure 15.2 Combining spatial microsimulation and Remote sensing (Ballas *et al.*, 2000).

the new framework comes when the potential of microsimulation for business applications is considered. Given the potential to create lists of household attributes, it has long been recognized that microsimulation could be useful as a business tool. However, to date, little progress has been made in exploring this potential. In a sense, the database underpinning the microsimulation model offers the same kind of information currently in many geodemographic or lifestyle data systems. Nevertheless, microsimulation offers much greater flexibility than many standard geodemographic systems. In most cases, the geodemographic systems provide only one label for each locality. This is based on the greatest percentage of each group represented in the locality. Unless this percentage match is close to 100% there are always ecological fallacy problems: i.e., the label does not capture all consumer types resident in a particular area. This has led a number of authors to suggest 'fuzzy geodemographics', which might involve giving numerous labels to each locality. For more discussion on this see Feng and Flowerdew (1998) and See and Openshaw (2001). However, microsimulation would potentially offer another route to finding customers or consumer groups of various types. From a main database of say 100 household variables it is possible to search for distributions made up of any of these variables. The possible number of combinations is very large indeed and the user could ask for very specific combinations of variables, adding great flexibility to the task of finding customers. Second, it would be possible to provide unique classifiers for different localities. At the moment the 'underprivileged' group may be made up of key census variables clustered in many different ways to end up with this classification. A major research question is whether the underprivileged groups identified in Liverpool are the same as those identified in the East End of London. A more subtle look at the outputs of the

microsimulation could offer new insights into this issue.

Finally, the framework suggested here would add much to the potential of remotely sensed data. It would be possible to put estimations on the types of buildings in terms of housing types and characteristics of their inhabitants. Clearly, it is not possible to categorically say what types of families were in each building. However, it may be possible to give an estimation of the types of families within blocks thus giving very detailed portraits of small areas of our cities.

15.4.4. Spatial microsimulation, spatial decision support systems and virtual decision-making environments

Another area where spatial microsimulation models can play an important role is in the ongoing debates on the potential of new technologies to promote local democracy and electronic decision-making. It can be argued that spatial microsimulation models can be used not only to provide information on the possible consequences and the local multiplier effects of major policy changes but also to inform the general public about these and to enhance, in this way, the public participation in policy making procedures.

An example of work moving towards this direction is the Microsimulation Modelling and Predictive Policy Analysis System (Micro-MaPPAS) developed for the Leeds City Council by researchers at the Universities of Sheffield, Leeds and Manchester (Ballas *et al*., 2004, 2006). MicroMaPPAS is a planning support system based on the SimLeeds geographical microsimulation model mentioned above. The SimLeeds software (Ballas, 2001) has been run from a Command prompt and required the 'hard coding' of parameters and data tables together with some knowledge of Java programming – not a desirable task for the average policy or decision maker. MicroMaPPAS provides a spatial decision making interface which is much more user-friendly and suitable for decision makers who can utilize the power of the spatial microsimulation methodology. The MicroMaPPAS software also provides some basic mapping functions such as panning and zooming and symbology editing. The mapping capability in the software is provided by the GeoTools (www.geotools.org) open source Java mapping library, which has been written by a group of researchers independent of the MicroMaPPAS project. GeoTools is a versatile Java library which conforms to the Open GIS Consortium standard specifications in relation to GIS open operability. The library can be adapted to work in any Java based GUI or web-based Applet. The mapping controls allow the user to select a microsimulated variable from a query and map the results at a wide range of different geographical scales (see Ballas *et al*. (2004) and Ballas *et al*. (2006) for more details).

It can be also argued that systems such as MicroMaPPAS can have an 'e-government' dimension by allowing networking technologies including the Internet to be used by policy makers as well as the general public. In particular, these systems can be converted into web-based GIS to enhance public involvement and participation in environmental planning and decision making processes. Such systems are typically referred to in the literature as Public Participation GIS (PPGIS) and are based on the belief that by providing citizens with access to information and data in the form of maps and visualizations, they can make better informed decisions about the natural and built environment around them. It is possible to build on

the existing infrastructure and knowledge in order to combine GIS and PPGIS frameworks to enhance e-government, local democracy and public participation. In particular, GIS and spatial microsimulation models can also play a very important role in the ongoing debates on the role of the potential of new technologies to promote local democracy and electronic decision-making. It can be argued that a system such as MicroMaPPAS developed in JAVA, can be put on the World Wide Web and linked to Virtual Decision-Making Environments (VDMEs). The latter are Internet World Wide Web based systems that allow the general public to explore 'real world' problems and become more involved in the public participation processes of the planning system (Evans *et al*., 1999; Kingston *et al*., 2000).

15.4.5. New application areas

In addition to comprehensive models it is useful to highlight other areas of economic or social geography where microsimulation has been under-utilized. One such area is medical geography. A notable exception is the work of Clarke and Spowage (1984), who designed morbidity and mortality sub-models for health care planning in West Yorkshire, UK. They estimated the probabilities of being ill or dying based on age, sex, social class, ethnicity, etc. (by speciality case). Another sub-model was constructed to simulate hospital workloads and patient throughput.

Recent concerns in UK public health planning have focused on two main issues. The first has been the concern to improve health inequalities by investing more on intervention strategies. The second has been the concern to treat more patients within the community. Microsimulation lends itself well to addressing both these concerns. For intervention strategies we need to understand more about geodemographic variations in demand for health services. Of particular concern in the UK at the moment are the problems of obesity (especially childhood obesity), diabetes and smoking. The difficulty is that little is known about the prevalence of these health issues by household or neighbourhood. Given age, sex, social class, occupation, ethnicity, etc., microsimulation models can estimate the incidence of such problems (and be calibrated against any existing data). Once demand is better understood and measured, the location of community health services becomes easier in the sense of finding locations to maximize access to potential users. In addition, other *what-if* scenarios are possible. For the location of stop smoking services for example, it would also be possible to simulate the success across the city of services targeted at different geodemographic groups (young adults, heavy smokers aged 65 or over, pregnant mothers, etc.). Similarly, for diabetes, it would be possible to model the impacts of improving access to fresh fruit and vegetables and hence improving diet across households of different types. Smith *et al*. (2006) give further details on the research agenda for diabetes and food access.

15.4.6. Improving model calibration

Despite the benefits of the applications described in this chapter, it should be noted that caution is necessary when using spatial microsimulation methodologies to perform *what-if* policy analysis and evaluation. The outputs of all microsimulation models, no matter how good, are always simulations and not actual data. The validity of the simulated data will depend on the quality of the original data that are used and on the assumptions upon which the microsimulation model is based. Moreover, it will depend

on the specific microsimulation methodology that is employed. In addition, spatial microsimulation outputs generally depend on subjective judgements associated with the ordering of the conditional probability tables that are used as inputs and/or with the selection of the data sets that are used as small area constraints. As Birkin and Clarke (1995) point out, the modeller's art in microsimulation is to generate population characteristics in an appropriate order so that potential errors are minimized. These aspects should always be taken into account when using spatial microsimulation models for policy impact assessment.

However, there is the related problem of how to validate microsimulation outputs, since there are no available micro-data sets at the desired level of geographical scale (hence the need for microsimulation in the first place!). Model output validation is one of the biggest problems of microsimulation methodologies. As Williamson (1999) points out, in the United States the National Academy was commissioned to evaluate the effectiveness of microsimulation for tax-benefit analysis purposes. The National Academy found that there is a general lack of thorough validation for microsimulation models and proposed a number of validation measures such as external validity studies in which model results are compared with data from program administrative sources (Williamson, 1999). Moreover, sensitivity analysis and computer-intensive 'sample reuse technique methods' to measure the variance in model estimates were proposed.

Thus, further research is required, in order to improve the performance of spatial microsimulation models and to highlight the sources of error. For instance, as Williamson *et al*. (1998) point out, there are many ways in which combinatorial optimization methodologies can be fine-tuned, through the evaluation of the use of more or different SAS tables or by changing the model parameters (also see Voas and Williamson (2000) for a more detailed discussion and an in-depth evaluation of combinatorial optimization techniques). Further, there is a need to build on existing work on the validity and reliability of microsimulation models (such as the work of Pudney and Sutherland (1994) who investigated the role of sampling error in a tax-benefit model and the work of Voas and Williamson (2001) who present new 'goodness-of-fit' measures for synthetic microdata).

15.5. CONCLUSIONS

We hope that we have demonstrated that spatial microsimulation is a useful technique for estimating the characteristics of individuals or households which can then be used in a variety of *what-if* situations regarding policy change. The key advantage of this methodology is data fusion or linkage – a variety of data sets can be combined to provide new insights into household characteristics and, ultimately, household behaviour. Thus these models can help to solve the problem of 'missing data' such as, in the UK, household income, wealth, tax payment, water demand, health problems, crime, etc. Once built, these models can also be linked to meso or macro models (such as discrete choice models, spatial interaction models, logit models, input–output models, etc.) to show how households interact with the supply side of the economy (where they go to work, shop, visit the doctor, etc.). The ability to change these circumstances and assess the impacts of such actions is another major advantage of this methodology. Simulations can be 'run' which change either the characteristics of the households (population ageing, new job allowing greater income to be earned, change of residence, etc.) or the characteristics of the supply

side (new retail centre, closure of a major employer, new hospital, etc.). This ability to examine both household dynamics and the impacts of infrastructure change allow the analyst to explore both social policy impacts (tax or welfare changes for example) and/or area-based policy impacts (new job creation, new retail centre, etc.).

The research agenda outlined in the second half of the chapter is clearly our personal one but one that we hope other microsimulation modellers would at least partially agree with. The agenda has not been presented in any particular order of importance but the issue of how such models can support traditional spatial modelling seems a key task to address in the short term. As we noted above, a start has been made in this direction but perhaps the greatest challenge is merging microsimulation with more macro techniques such as input-output models. The latter models are excellent for modelling the interactions between key sectors of the economy but not so good at spatially disaggregating the outputs within cities and regions. A methodology which could feed individual households into the economic system at both stages of the modelling process (inputs and outputs) could be a major advantage in future policy work. We hope we can address this issue in the next few years.

REFERENCES

Ballas, D. (2001). A spatial microsimulation approach to local labour market policy analysis, unpublished PhD thesis, School of Geography, University of Leeds.

Ballas, D. and Clarke, G.P. (2000). GIS and microsimulation for local labour market policy analysis. *Computers, Environment and Urban Systems*, **24**: 305–330.

Ballas, D. and Clarke, G.P. (2001a). Modelling the local impacts of national social policies: a spatial microsimulation approach. *Environment and Planning C: Government and Policy*, **19**: 587–606.

Ballas, D. and Clarke, G.P. (2001b). Towards local implications of major job transformations in the city: a spatial microsimulation approach. *Geographical Analysis*, **33**: 291–311.

Ballas, D., Clarke, G.P. and Dewhurst, J. (2006). Modelling the socio-economic impacts of major job loss or gain at the local level: a spatial microsimulation framework. *Spatial Economic Analysis*, vol. 1(1), pp. 127–146.

Ballas, D., Clarke, G.P., Dorling, D., Eyre, H., Rossiter, D. and Thomas, B. (2003). *SimYork: Simulating Current and Future Trends in the Life of Households in York*, report to the Joseph Rowntree Foundation, May 2003.

Ballas, D., Clarke, G.P., Dorling, D., Eyre, H., Rossiter, D. and Thomas, B. (2005). SimBritain: A spatial microsimulation approach to population dynamics, *Population, Space and Place*, **11**: 13–34.

Ballas, D., Clarke, G.P., Feldman, O., Gibson, P., Jianhui, J., Simmonds, D. and Stillwell, J. (2005b). *A Spatial Microsimulation Approach to Land-use Modelling*, *CUPUM 2005 (Computers in Urban Planning and Urban Management*) Conference Proceedings, UCL, London 29 June–1 July 2005 (available on-line from: http:// 128.40.59.163/ cupum/searchPapers/papers/paper276.pdf)

Ballas, D., Clarke, G.P., Dorling, D. and Rossiter, D. (2007). Using SimBritain to Model the Geographical Impact of National Government Policies, *Geographical Analysis*, **39**(1): 44–77.

Ballas, D., Kingston, R. and Stillwell, J. (2004). Using a spatial microsimulation decision support system for policy scenario analysis. In: van Leeuwen, J. and Timmermans, H. (eds), *Recent Advances in Design and Decision Support Systems in Architecture and Urban Planning*, pp. 177–192. Dordrecht: Kluwer.

Ballas, D., Kingston, R. and Stillwell, J. and Jin, J. (2007). Building a spatial microsimulation-based planning support system for local policy making. *Environment and Planning A*, **39**(10): 2482–2499.

Ballas, D., Rossiter, D., Thomas, B., Clarke, G.P. and Dorling, D. (2005). *Geography Matters: Simulating the Local Impacts of National Social Policies*, Joseph Rowntree Foundation contemporary research issues, Joseph Rowntree Foundation, York.

Batey, P.W.J. (2003). Extended input–output modelling of regional impacts: does detail make a difference?

paper presented at the *Royal Geographical Society Annual Conference 2003* (Special session on '50 years of Regional Science or the Return of Quantitative Economic Geography'), London, 3–5 September 2003.

Batty, M. and Densham, P. (1996). Decsion support, GIS and urban planning. *Systemma Terra*, **V**(1): 72–76.

Birkin, M. and Clarke, M. (1988). SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and Planning A*, **20**: 1645–1671.

Birkin, M. and Clarke, G.P. (1995). Using microsimulation methods to synthesize census data. In: Openshaw, S. (ed.), *Census Users' Handbook*, pp. 363–387. London: GeoInformation International.

Birkin, M. and Clarke, M. (1989). The generation of individual and household incomes at the small area level using Synthesis, *Regional Studies*, **23**: 535–548.

Birkin, M., Clarke, G.P. and Clarke, M. (1996). Urban and regional modelling at the microscale. In: Clarke, G.P. (ed.), *Microsimulation for Urban and Regional Policy Analysis*, pp. 10–27. London: Pion.

Boman, M. and Holm, E. (2004). Multi-agent systems, time geography and microsimulations. In: M.O. Olsson and G. Sjostedt (eds), *Systems, Approaches and their Application*, pp. 95–118. Dordrecht: Kluwer Academic.

Caldwell, S.B. and Keister, L.A. (1996). Wealth in America: family stock ownership and accumulation, 1960–1995. In: Clarke, G.P. (ed.), *Microsimulation for Urban and Regional Policy Analysis*, pp. 88–116. London: Pion.

Caldwell, S.B., Clarke, G.P. and Keister, L.A. (1998). Modelling regional changes in US household income and wealth: a research agenda. *Environment and Planning C: Government and Policy*, **16**: pp. 707–722.

Clarke, G.P. (1996). Microsimulation: an introduction. In: Clarke, G.P. (ed.), *Microsimulation for Urban and Regional Policy Analysis*, pp. 1–9. London: Pion.

Clarke, M. and Spowage, M. (1984). Integrated models for public policy analysis: an example of the practical use of simulation models in health care planning. *Papers of the Regional Science Association*, **55**: 25–46.

Clarke, G. and Stillwell, J.C.H. (eds) (2004). *Applied GIS and Spatial Modelling*, London, Wiley.

Davidsson, P. (2000). Multi agent based simulation: beyond social simulation. In: S. Moss and P. Davidsson (eds), *Multi Agent Based Simulations*, pp. 97–100. Berlin: Springer.

Davies, H. and Joshi, H. (1992). Constructing Pensions for Model Couples, in R. Hancock and H. Sutherland (eds), *Microsimulation Models for Public Policy Analysis: New Frontiers*, Suntory-Toyota International Centre for Economics and Related Disciplines – LSE, London, 67–96.

Evans, A., Kingston, R., Carver, S. and Turton, I. (1999). Web-based GIS to enhance public democratic involvement, paper presented at the *4th International Conference on GeoComputation*, Fredericksburg, Virginia, USA, 25–28 July.

Evandrou, M. and Falkingham, J. (1995). Gender, Lone-parenthood and Lifetime Incomes, in J. Falkingham and J. Hills (eds), *The dynamic of welfare: the welfare state and the life cycle*, Prentice Hall/Harvester Wheatsheaf, New York, pp. 167–183.

Falkingham, J., Harding, A. and Lessof, C. (1995). Simulating lifetime income distribution and redistribution. In: J. Falkingham and J. Hills (eds), *The Dynamic of Welfare: the Welfare State and the Life Cycle*, pp. 62–82. New York: Prentice Hall/Harvester Wheatsheaf.

Falkingham, J. and Lessof, C. (1992). Playing God or LIFEMOD – The construction of a dynamic microsimulation model. In: R. Hancock and H. Sutherland (eds), *Microsimulation Models for Public Policy Analysis: New Frontiers*, pp. 5–32. London: Suntory-Toyota International Centre for Economics and Related Disciplines – LSE.

Falkingham, J. and Hills, J. (1995a). The effects of the welfare state over the life cycle. In: J. Falkingham and J. Hills (eds), *The Dynamic of Welfare: the Welfare State and the Life Cycle*, pp. 83–107. New York: Prentice Hall/Harvester Wheatsheaf.

Falkingham, J. and Hills, J. (1995b). Redistribution between people or across the life cycle? In: J. Falkingham and J. Hills (eds), *The Dynamic of Welfare: the Welfare State and the Life Cycle*, pp. 137–149. New York: Prentice Hall/Harvester Wheatsheaf.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis*. Sage Publications.

Glennerster, H., Falkingham, J. and Barr, N. (1995). Education funding, equity and the life cycle.

In: J. Falkingham and J. Hills (eds), *The Dynamic of Welfare: the Welfare State and the Life Cycle*, pp. 150–166. New York: Prentice Hall/Harvester Wheatsheaf.

Hägerstrand, T. (1967). *Innovation diffusion as a spatial process*, University of Chicago Press, Chicago.

Hancock, R. (2000). Changing for care in later life: an exercise in dynamic microsimulation. In: L. Mitton, H. Sutherland and M. Weeks (eds), *Microsimulation Modelling for Policy Analysis: Challenges and Innovations*, pp. 226–237. Cambridge: Cambridge University Press.

Hancock, R. and Sutherland, H. (eds) (1992). *Microsimulation Models for Public Policy Analysis: New Frontiers*. London: Suntory-Toyota International Centre for Economics and Related Disciplines – LSE.

Hancock, R., Mallender, J. and Pudney, S. (1992). Constructing a computer model for simulating the future distribution of pensioners' incomes for Great Britain. In: R. Hancock and H. Sutherland (eds), *Microsimulation Models for Public Policy Analysis: New Frontiers*. pp. 33–66. London: Suntory-Toyota International Centre for Economics and Related Disciplines – LSE.

Harding, A. (ed.) (1996). *Microsimulation and Public Policy*. Amsterdam: North Holland, Contributions to Economic Analysis 232.

Heppenstall, A.J., Evans, A.J. and Birkin, M.H. (2007). Genetic Algorithm Optimisation of a Multi-Agent System for Simulating a Retail Market. *Environment and Planning B*, **34**: 1051–1070.

Holm, E., Lindgren, U., Makila, K. and Malmberg, G. (1996). Simulating an entire nation. In: Clarke, G.P. (ed.), *Microsimulation for Urban and Regional Policy Analysis*, pp. 164–186. London: Pion.

Hooimeijer, P. (1996). A life-course approach to urban dynamics: state of the art in and research design for the Netherlands. In: Clarke, G.P. (ed.), *Microsimulation for Urban and Regional Analysis*, pp. 28–63: London: Pion.

Huang, Z. and Williamson, P. (2001). A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata, Working Paper 2001/2, Department of Geography, University of Liverpool.

Kingston, R., Carver, S., Evans, A. and Turton, I. (2000). Web-based public participation geographical information systems: an aid to local environmental decision-making, *Computers, Environment and Urban Systems*, **24**: 109.

Longley, P.A. and Batty, M. (2003). (eds), *Advanced spatial analysis: The CASA book of GIS*, Redlands, CA: ESRI Press.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds.) (1999). *Geographical Information Systems: Management Issues and Applications*. New York: Wiley.

Martin, D. (1996). *Geographic Information Systems: Socioeconomic Applications*. London: Routledge.

Mertz, J. (1991). Microsimulation – A survey of principles developments and applications, *International Journal of Forecasting*, **7**: 77–104.

Mitton, L., Sutherland, H. and Weeks, M. (eds) (2000). *Microsimulation Modelling for Policy Analysis: Challenges and Innovations*. Cambridge: Cambridge University Press.

Nakaya, T., Yano, K., Fotheringham, A.S., Ballas, D. and Clarke, G.P. (2003). Retail interaction modelling using meso and micro approaches, Paper presented at the *33rd Regional Science Association, RSAI – British and Irish Section Conference*, St. Andrews, Scotland, 20–22 August.

Nelissen, J.H.M. (1993). Labour market, income formation and social security in the microsimulation model NEDYMAS, *Economic Modelling*, **10**: 225–272.

Openshaw, S. (1995). Human systems modelling as a new grand challenge area in science. *Environment and Planning A*, **27**: 159–164.

Orcutt, G.H. (1957). A new type of socio-economic system. *The Review of Economics and Statistics*, **39**: 116–123.

Orcutt, G.H., Mertz, J. and Quinke, H. (eds) (1986). *Microanalytic Simulation Models to Support Social and Financial Policy*. Amsterdam: North-Holland.

Orcutt, G.H., Greenberger, M., Korbel, J. and Rivlin, A. (1961). *Microanalysis of Socioeconomic Systems: A Simulation Study*, Harper and Row, New York.

Paas, G. (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information, in G.H. Orcutt, J. Mertz and H. Quinke (eds), *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam, 401–421.

Propper, C. (1995). For richer, for poorer, in sickness and in health: The lifetime distribution of NHS

health care. In: J. Falkingham and J. Hills (eds), *The Dynamic of Welfare: the Welfare State and the Life Cycle*. pp. 184–203. New York: Prentice Hall/Harvester Wheatsheaf.

Pudney, S. and Sutherland, H. (1994). How reliable are microsimulation results? An analysis of the role of sampling error in a UK tax-benefit model, *Journal of Public Economics*, **53**: 327–365.

Redmond, G., Sutherland, H. and Wilson, M. (1998). *The Arithmetic of Tax and Social Security Reform: a User's Guide to Microsimulation Methods and Analysis*, Cambridge: Cambridge University Press.

See, L. and Openshaw, S. (2001). Fuzzy geodemographic targeting. In: G.P. Clarke and M. Madden (eds), *Regional Science in Business*, 269–282. Berlin: Springer-Verlag.

Smith, D., Clarke, G.P., Ransley, J. and Cade, J. (2006) Food access and health: a microsimulation framework for analysis. *Studies in Regional Science*, **35**(4), 909–927.

Sutherland, H. and Piachaud, D. (2001). Reducing child poverty in Britain: an assessment of government policy 1997–2001, *The Economic Journal*, **111**: 85–101.

Sutherland, H., Sefton, T. and Piachaud, D. (2003). *Poverty in Britain: the Impact of Government Policy since 1997*, Joseph Rowntree Foundation, York (also available on-line from: http://www.jrf.org.uk) (ISBN 1 85935 152 2).

Sutherland, H., Taylor, R. and Gomulka, J. (2002). Combining household income and expenditure data in policy simulations, *Review of Income and Wealth*, **48**(4): 75–94.

Taylor, M.F., Brice J., Buck, N., Prentice-Lane, E. (2001). *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices.* Colchester: University of Essex.

Veldhuisen, K.J., Kapoen, L.L. and Timmermans, H.J.P. (2000). RAMBLAS: A regional planning model based on the micro-simulation of daily activity patterns, *Environment and Planning A*, **31**: 427–443.

Vencatasawmy, C.P., Holm, E. and Rephann, T. *et al*. (1999). Building a spatial microsimulation model, paper presented at the 11th Theoretical and Quantitative Geography European colloquium, Durham Castle, Durham, 3–7 September, 1999.

Voas, D. and Williamson, P. (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, **6**: 349–366.

Voas, D. and Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata, *Geographical and Environmental Modelling*, **5**: 177–200.

Wegener, M. and Spiekermann, K. (1996). The potential of microsimulation for urban models, in G.P. Clarke (ed.) *Microsimulation for Urban and Regional Policy Analysis*, Pion, London, 149–163.

Wertheimer II, R., Zedlewski, S.R., Anderson, J., Moore, K. (1986). DYNASIM in comparison with other microsimulation models, in G.H. Orcutt, J. Mertz and H. Quinke (eds), *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam, 187–206.

Williamson, P. (1992). Community care policies for the elderly: a microsimulation approach. Unpublished PhD Thesis, School of Geography, University of Leeds, Leeds.

Williamson, P., Birkin, M. and Rees, P. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, **30**: 785–816.

Williamson, P. and Voas, D. (2000). Income estimates for small areas: lessons from the census rehearsal, *BURISA*, **146**: 2–10.

Williamson, P. (1999). Microsimulation: An idea whose time has come? paper presented at the *39th European Regional Science Association Congress*, University College Dublin, Dublin, Ireland, *23–27 August.*

Wilson, A. and Pownall, C.E. (1976). A new representation of the urban system for modelling and for the study of micro-level interdependence, *Area*, **8**: 246–254.

Wilson, A.G. (2000). *Complex Spatial Systems: the Modelling Foundations of Urban and Regional Analysis*. London: Prentice Hall.

16

Detection of Clustering in Spatial Data

Lance A. Waller

16.1. INTRODUCTION

It is human nature to seek pattern within any complex display of information. We organize stars into constellations, devour mystery novels, and even give detailed descriptions of ink stains to analysts. This innate desire for order within chaos applies spatially as well. Given a map of a set of locations of an event, say, residences of cases of a particular type of disease or the locations of a particular type of crime, we seek patterns that might reveal something about the underlying process generating the events, be that a common environmental exposure or the habits of a particular criminal. In short, our hope is that arranging what we know spatially might reveal something about how the events arise in the first place.

In this chapter, we review analytic methods for detecting 'clusters' or 'hot spots' in spatially-referenced data. We begin with a discussion of what we mean conceptually, geographically, and mathematically by the term 'cluster', then discuss and illustrate many standard approaches proposed and applied in the literature within a variety of scientific fields. Many analytic approaches for detecting clusters have been summarized in several texts (Elliott *et al*., 1992, 1999; Cressie, 1993; Bailey and Gatrell, 1995; Goldsmith *et al*., 2000; Lawson, 2001; Lawson and Denison, 2002; Diggle, 2003; Waller and Gotway, 2004; Eck *et al*., 2005), so we do not attempt a complete review here. Rather, we focus on developed and developing conceptual and theoretical constructs behind many of the methods while contrasting the underlying questions of

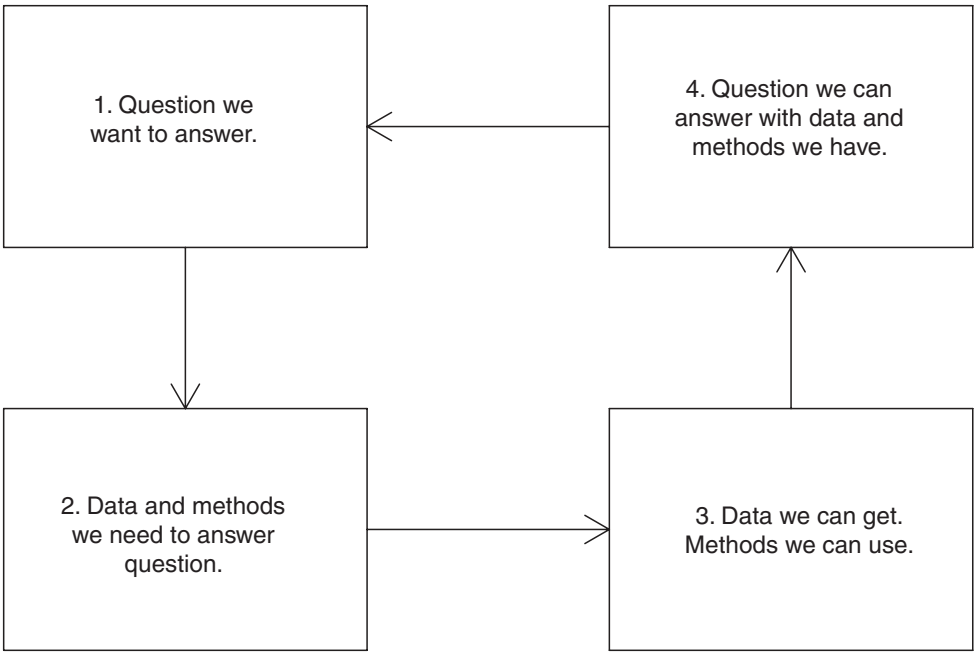


Figure 16.1 The 'whirling vortex' of spatial data analysis.

interest driving different families of analytic approaches.

To set the stage conceptually, Figure 16.1 provides a starting point for developing and evaluating analytic methods for detecting clusters and clustering. We begin with Step 1 with a question we wish to answer (for example, 'Are disease risks elevated for individuals living near a source of pollution?'). The question of interest defines the sorts of data and methods we require to answer the question (for example, individual-level case status and individual exposure histories). However, the data required often are unavailable for reasons varying from cost to privacy and we often settle for related data we can obtain within budget and satisfying availability constraints (for example, present residential location of cases and proximity to known sources of pollution). Similarly, available methods may only address part of the question or may be particularly susceptible to data shortcomings (for example, missing data or location inaccuracy). This situation is particularly relevant in the analysis of spatial data with the increasing sophistication and data holdings of geographic information systems (GISs). One is increasingly faced with the ease of including 'found' data collected by others that seems to fit the bill for the data one would really like to have. After obtaining the data we can retrieve, we conduct analysis on these available data, often without explicitly acknowledging that our analyses may be addressing slightly different questions (e.g., in our conceptual example, we have moved from a question involving associations between disease and a particular exposure, to associations between disease and present proximity to a known or suspected exposure source). As a final step, we should carefully examine how closely the questions we do answer mirror those we originally intended to answer. All too often, this last step is ignored.

While we can consider the steps shown in Figure 16.1 as a linear set of steps (1, 2, 3, 4), it is often a loop where the answers obtained on the available observational data in Step 4 inform on refinements to the questions

asked in Step 1 and suggest limitations arising due to the data compromises between Steps 2 and 3.

16.2. WHAT ARE WE LOOKING FOR?

It is appropriate to begin by considering the very basic question: What exactly do we hope to find? Besag and Newell (1991) provide several important observations relevant to the search for clusters. The first key distinction is between detection of 'clusters' and the detection of 'clustering'. A cluster represents an unusual collection of events while clustering represents a general tendency for events to occur nearer other events than one might expect.

These definitions of 'cluster' and 'clustering' differ from those found in 'cluster analysis', a set of analytical classification methods designed to group observations into 'clusters' wherein observations within the same cluster are more alike than those from different clusters. The overlap in terminology can be confusing when reviewing the literature, especially since some spatial methods to detect clusters and/or clustering utilize concepts and methods from cluster analysis (Knorr-Held and Raßer, 2000; Denison and Holmes, 2001). As illustrated in Figure 16.1, it remains critical to clearly identify goals and conclusions in the context of both the questions addressed and the methods used to address them.

In the discussion below, we follow Besag and Newell (1991) and take the term 'cluster' to define an anomaly, an interesting collection of spatial locations that appears to be inconsistent with some background conceptual model of how events arise. For instance, a cancer registry might report six new cases of childhood leukemia in a small neighborhood in a particular year, when only one new case would be expected if the national annual incidence rate applied directly to all individuals in the study area. That is, the aggregation of six cases appears to be unlikely under a simple model of all children experiencing equal risk. Contrast this example with that of clustering where we observe multiple pockets of higher incidence than expected from national rates, perhaps separated by areas of lower-than-expected local rates.

Besag and Newell (1991) also note the difference between seeking clusters or clustering anywhere versus around particular locations of interest. They denote the former as 'general' methods and the later as 'focused' methods, also referred to as 'global' and 'local' methods, respectively, in the geography literature by Anselin (1995) and in the disease clustering literature by Kulldorff *et al*. (2003).

As suggested by Figure 16.1, seeking general or focused clusters or clustering defines different questions of interest and, as a result, methods appropriate for seeking individual clusters might not be the best approach to measure clustering and vice versa. We will explore this in more depth in the examples below.

The general ideas of clusters and clustering arise in many different disciplines. However, each discipline often brings its own particular sets of questions of interest, assumptions regarding data availability, and familiar statistical methods. For example, the fields of epidemiology and criminology both exhibit interest in the detection of clustering within geographically referenced data sets. However, the sets of techniques appearing in their respective literatures are largely distinct and cross-references between the fields are rare. This situation is unfortunate since both fields could draw from the experiences and ideas of the other. Figure 16.1 provides a general context for comparison and we express and compare ideas from recent surveys in both fields in the sections below.

The remainder of the chapter addresses the typical types of data available for cluster detection; some basic analytic concepts, assumptions, and complications; an illustrative data set from archaeology; an overview of some different approaches for detecting clusters and/or clustering in point-referenced data with application to the data set; and general conclusions. As a result, the chapter represents more of a review of the questions one should ask in performing a search for clusters or clustering than an exhaustive collection of methods.

16.3. WHAT DATA DO WE HAVE?

As one might expect, the typical data for cluster detection consist of locations on a map. These may be point locations of events or may represent counts of events occurring within a set of zones partitioning the study area into non-overlapping pieces. Examples of the latter include census enumeration districts, postal zones, or other administrative regions. Regional counts may arise to preserve individual confidentiality or simply due to the relative ease of obtaining records sorted by political district, mailing addresses, or other identifier. We concentrate on point-referenced data in the development below noting that methodologically we typically assume a latent, unobserved set of points behind regional counts and many of the analytic tools used for points provide the basis for similar tools for counts (Waller and Gotway, 2004, Chapters 6–7).

In addition to the point locations or regional counts of events, it is often very important to have access to data defining the spatial heterogeneity of the population from which our events are drawn. These may be potential crime victims, individuals susceptible to the disease of interest, or simply the population sizes for each area. The background information is critically important in the interpretation of any detected clusters since it defines the amount of clustering we would expect under some null model of event occurrence. This null model defines the patterns we would expect in the absence of anomalies. A common null model is one of *constant risk* where each individual in the study area experiences an identical probability of experiencing the event under study. To illustrate the importance of the background information, consider as a contrived example a collection of six childhood leukemia cases in one neighborhood which would seem very unusual if only six children reside there but not as unusual if 600,000 children live there. The background data coupled with the null model provide our statistical point of reference for detecting clustering and clusters.

We also may have spatially-referenced covariate information providing information regarding the spatial distribution of factors impacting the local probability of the events of interest. For instance, the incidence of most cancers increases dramatically with age. As a result, we would tend to expect more cases in neighborhoods with higher numbers of older residents. The covariate information may include both endogenous and exogenous variables. In some sense, the covariate information is collected to define 'uninteresting' clustering, that is, clustering for reasons we already know or suspect. In most cases, cluster detection builds from a desire to identify areas where the observed pattern of events doesn't match our general expectations.

16.4. WHAT ANALYTIC TOOLS CAN WE USE?

Most methods to detect clusters and clustering build from probability models

operationalizing the null model mentioned above. As a result, most tools aim to define some measure of the 'unusualness' of a cluster, then determine the distribution of this quantity under the null (uninteresting) model, and compare the quantity based on the observed data to this null distribution (Waller and Jaquez, 1995). In a statistical hypothesis setting, the null hypothesis is defined conceptually as the absence of clusters/clustering, and operationally as the expected distribution of our measure (statistic) under the null model.

As a result, the analytic tools required for statistical inference are a definition of our statistic and its null distribution. In the sections below, we will illustrate several types of statistics and contrast the underlying questions addressed by each.

Before defining particular methods, we offer a brief review of some basic probabilistic elements for point-referenced event locations driving many of the methods illustrated below. The first is the definition of complete spatial randomness (CSR). A set of events arising from CSR has the following properties: first, the total number of events observed in the study area follows a Poisson distribution; second, given the observed number of events, event locations occur independently of one another and the expected number of events per unit area is a constant, denoted λ, across the entire study area. CSR corresponds to a spatial Poisson point process yielding the following features: the number of events observed in a region A within the study area follows a Poisson distribution with mean $\lambda|A|$ where $|A|$ denotes the area of A, the number of observed events in non-overlapping areas are independent of one another, and, given the observed number of events, events are uniformly distributed across the study area (and any region within it). For clarity we follow Diggle (2003) and distinguish between an *event location* where an observed event did occur, and a *point location* where an event could occur. A typical data set consists of a set of event locations and we often compare the value of our statistic based on events to the distribution of values associated with randomly selected events.

While CSR represents a complete lack of clustering, data generated by CSR nonetheless visually exhibits some 'clumping' and 'gapping' due to the inherent randomness, and one purpose of a statistical test is to determine whether the observed patterns in our data are more extreme than the amount of clumping and gapping expected under CSR. Figure 16.2 illustrates three realizations of CSR with 100 events uniformly distributed across a square. It is worth noting that the

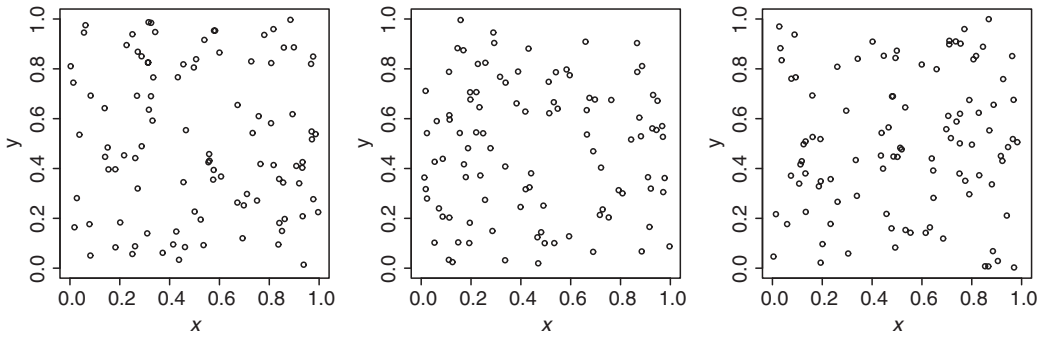


Figure 16.2 Three examples of complete spatial randomness (CSR).

uniform distribution of event locations represents a uniform probability of occurrence, not an 'evenly spaced' set of events.

CSR represents a convenient null model and many tests of CSR exist in the literature (Cressie, 1993, p. 604), but CSR may not be the appropriate reference pattern for applications where the population at risk is spatially heterogeneous. A common adjustment is the use of a heterogeneous Poisson process where the number of events expected per unit area is allowed to vary with location. If we define $\lambda(x)$ as the expected number of events per unit area at location (point) x, we refer to $\lambda(x)$ as the *intensity function* of the process. We adjust the Poisson process properties as follows: first, the number of events observed in any region still follows a Poisson distribution but now with the mean defined as the integral of $\lambda(x)$ over that region, counts from non-overlapping regions remain statistically independent, and events are distributed according to a (spatial) probability density function proportional to the intensity function. That is, more events are expected in locations where the intensity function is high, and fewer events are expected in locations where the intensity function is low.

The heterogeneous Poisson process offers a flexible model of the spatial distribution of point-locations of events, and its properties regarding counts for non-overlapping regions define the distributional basis for several commonly-used models for regional count data. However, the assumed independence of counts raises some eyebrows, especially among geographers for whom spatial autocorrelation is often a fundamental assumption in any spatial analysis (Tobler's First Law of Geography; Tobler, 1970). The distinction between a process defined by independent events with spatially patterned means versus a process defined by spatially correlated counts with identical means represents a lurking issue in the analysis of spatial pattern in general, and specifically in the detection of clusters. Bartlett (1964) showed that, without additional information, a pattern of independent events arising from a process with spatially varying intensity is mathematically indistinguishable from a process of spatially dependent events arising from a process with spatially constant intensity, let alone from patterns based on spatial variations in both correlation and intensity. The 'additional information' allowing one to separate the intensity and correlation effects could be based on temporal ordering of events to see if the location of past events influences future events (for example, with infectious diseases or diffusion of new technologies), or replicated observations of the same process over time to see if a suspected cluster remains in the same location (for example, near a putative source of increased risk) or if one observes similar patterning but in different locations for each time period. When contrasting methods based on independent or dependent events, it is important to recognize that both approaches represent an idealization of reality: neither approach is right, both are useful, but each answers our questions of interest in slightly different ways.

The basic probability models described above also provide a recipe for simulating sets of events following a given null model, thereby providing a powerful tool for Monte Carlo-based statistical inference. Recall that in frequency-based statistical hypothesis testing, one often considers the p-value, the probability under the null hypothesis of observing a more extreme value of the test statistic than one observes in the data set. Monte Carlo hypothesis testing (Barnard, 1963; Waller and Gotway, 2004; Chapter 5) uses simulation to estimate this probability by generating multiple data sets under the

null model, calculating the test statistic for each, constructing a histogram of these values as an approximation to the null distribution of the test statistic, and calculating the proportion of test statistic values associated with null simulations exceeding the value of the test statistic associated with the observed data. Note that the accuracy of the estimated p-value is a function of the number of simulations, not the sample size of the observed data, thereby putting the level of accuracy into the analyst's hands. This is not to say that sample size is unimportant. Sample size impacts the variation of the statistic under the null and alternative hypotheses, while the number of simulations controls the accuracy of the simulation-based tail probability (p-value) estimates. In some cases, theoretical derivations of proposed test statistics exist, but often these are based on particular distributional assumptions (for example, Gaussian or normally-distributed observations) and it is not always immediately clear whether the results apply in settings having different structures. In contrast, as long as one can simulate data under a reasonable null model, the Monte Carlo approach yields accurate inference.

Two general null models are worth mention in our discussion of Monte Carlo techniques for the detection of clusters/clustering. The first, mentioned above, is that of constant risk, that is, an assumed constant probability of the event outcome for each individual under study. If one has either point locations or regional counts reflecting a census, one can estimate the overall global risk of the event through the overall observed proportion of individuals experiencing the event. Then, one may randomly assign the observed number of events to the population at risk to obtain each simulated data set. The constant risk null model can also adjust for local covariate effects by using the covariates to define the local probability of an event. *Random labelling* provides a second null model, similar to the first, but designed when one has a sample of event locations and a sample of non-event or 'control' locations (individuals sampled from the population at risk of events) (Diggle, 2003; Waller and Gotway, 2004, Chapter 6) wherein we condition on the observed locations and randomly assign the event status ('label') among the full set of locations. That is, if we observe 30 events and have locations for 70 individuals not experiencing an event (controls), we keep the set of 100 locations, and randomly assign 30 of these to be 'events' in each simulated data set. Note that random labelling assumes a constant probability of event assignment, based on the observed ratio of events to non-events. At first glance, this seems identical to the constant risk assumption but two subtle differences remain. First, the random labelling hypothesis is conditional on the set of locations (both event and non-event) so random labelling simulations will not place events in any other locations. Second, constant risk simulations could be based on an event probability estimated from the observed data or could be based on an externally reported probability (for example, national disease or crime rates). If the study takes place in an area different from that providing the basis for the external probability, it is possible that the local probability is sufficiently higher or lower than the external probability so the observed data will seem inconsistent with simulated values based on the external value for no other reason than the discrepancy in the background probability and not due to spatial clusters or clustering within the data set.

Again referring to Figure 16.1, each of these steps represents a decision that may, subtly or not, impact the question addressed in the analysis. In the development, implementation, and review of specific

spatial analyses, it is important to design, report, and understand the type of null model driving the simulations in order to place results within the proper context and to connect Steps 4 and 1 in Figure 16.1.

Finally, it is worth noting that there are many more advanced computational and mathematical methods of statistical analysis of point patterns under current development. Such models allow one to define parametric models of clustering of event locations (Lawson and Denison, 2002), assign random measurements (often referred to as 'marks') to event locations, or allow interactions between multiple point processes observed over the same spatial study area (see Møller and Waagepetersen (2002) for detailed technical development). Many of these make use of computationally intensive Markov chain Monte Carlo (MCMC) methods for likelihood or Bayesian inference for parametric models of point processes. However, the non-parametric Monte Carlo approaches presented below represent exploratory techniques for detecting the presence of clusters and/or clustering without explicitly modeling the type of clustering. The approaches illustrated here offer robust statistical inference and a good starting place for analysis.

16.5. ILLUSTRATIVE DATA SET: ANASAZI SITES ON BLACK MESA, ARIZONA

To illustrate these concepts and to provide an illustration of the methods below, we consider a data set from the field of archaeology. The Peabody Coal Company leased land on the Black Mesa in northeastern Arizona, USA for coal mining. As part of the lease, the company contracted with archaeologists to conduct a detailed survey of archaeological sites in the lease area. The Black Mesa Archaeology Project conducted field research in the area between 1967 and 1987 leading to a body of research summarized in texts by Gumerman (1970), Gumerman *et al*. (1972), Plog (1986), and Powell and Smiley (2002). The study is relatively unique in its careful survey of a large tract of land and detailed mapping of the location of every site discovered on the surface. For our illustrative purposes, we make the simplifying assumption of a constant probability of detection of surface sites regardless of age or location. Figure 16.3 represents data locations abstracted from maps presented in Plog (1996). The 100 open circles represent sites dated to the time period 950–1049 CE and the 390 filled circles represent sites dated to the time period 1050–1150 CE. The later period represents a time of great expansion of the Anasazi culture (as represented by the increased number of settlement sites), but ends coincident with a time of large-scale abandonment of sites by the Anasazi throughout the southwestern United States *c*. 1100–1150 CE.

To illustrate the methods described below, we will compare spatial patterns between the 'early' and 'late' sites represented in the data set, seeking both clusters and clustering within the data sets.

16.6. DETECTING CLUSTERING

We begin with a general examination of clustering, the overall tendency for events to occur near other events. In the Anasazi data, possible questions of interest are: 'Do we observe clustering among all sites?' and 'Do we observe different types of clustering among the early and the late sites?' We focus on the latter question but, in the spirit of Figure 16.1, consider how it differs from the former in discussions below.

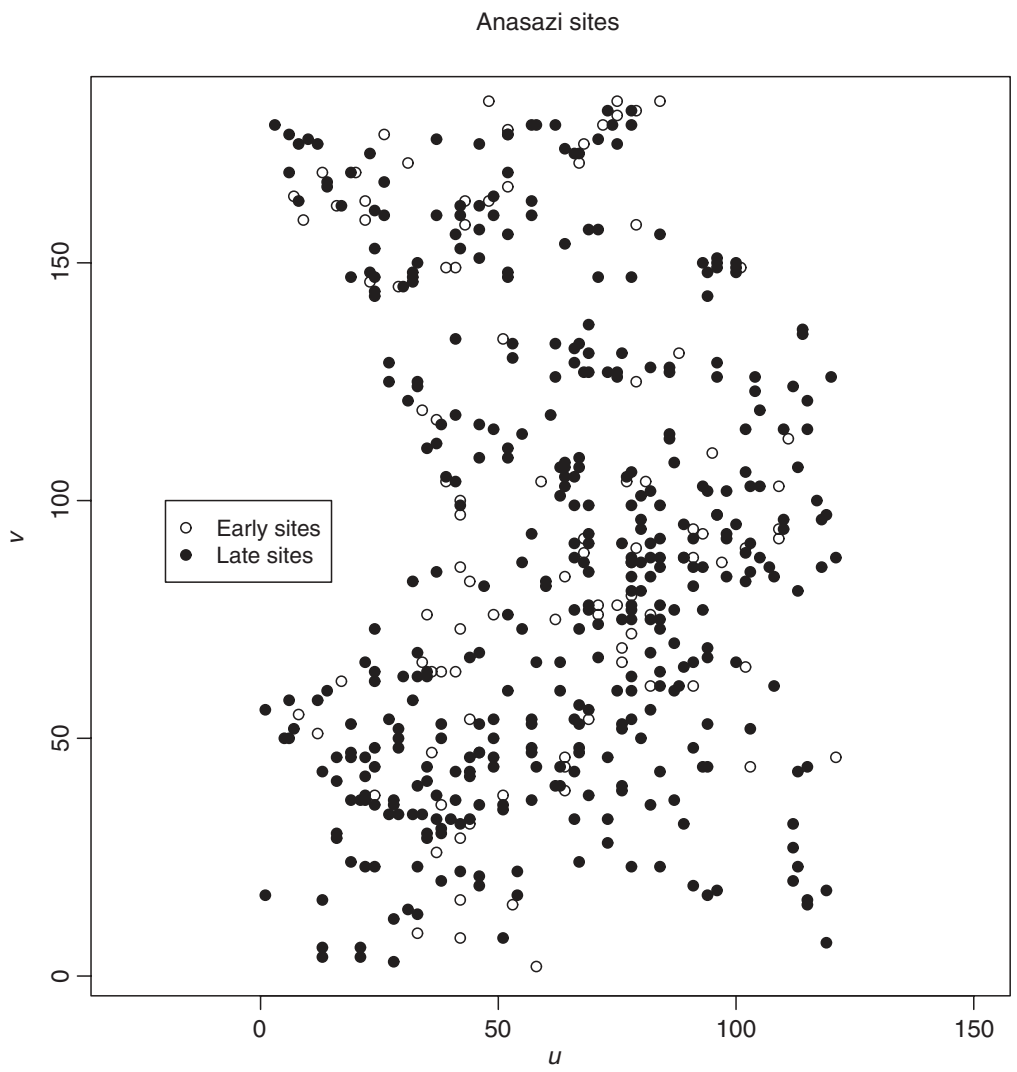


Figure 16.3 The Anasazi data set from the Peabody Coal Eastern Lease on Black Mesa in northeastern Arizona. Empty circles represent early sites (dated 950–1149 CE) and filled circles represent locations of later sites (dated 1050–1150 CE).

16.6.1. Who is my neighbor? Nearest neighbor analysis

First, suppose we observe two types of events in the same study area. In our data example, these correspond to early and late sites and the question of interest becomes: 'Does the pattern of clustering in late sites differ from that in early sites?' Note that this question explores the relative degree of clustering within the set of early and late sites, not whether either set of sites exhibits clustering or not.

There is a long tradition of exploring nearest neighbor patterns in spatial data (Cliff and Ord, 1973) and Cuzick and Edwards (1990) propose a test of clustering of one type of event within a set of two types of events in the same area. The test statistic is defined for a fixed number (k) of nearest neighbors and is, intuitively, the total number of late sites observed within the k nearest neighbors of other late sites. More formally, suppose we observe N events of which n_{late} are late sites. If we define the matrix B to

have elements $B_{k,ij} = 1$ if event j is in the k nearest neighbors of event i and if we define $\delta_i = 1$ if the ith event is late and $\delta_i = 0$ otherwise, then the test statistic becomes:

$$T_k = \sum_{i=1}^{N} \sum_{j=1}^{N} B_{k,ij}\delta_i\delta_j.$$

If late sites exhibit more clustering than early sites, we should observe more late sites near other late sites than we would expect under a random assignment of late sites to the observed locations of either type of event. Cuzick and Edwards (1990) derive an asymptotic normal distribution for the test statistic under the null hypothesis, but Monte Carlo tests under random labelling are applicable for any sample size.

Figure 16.4 illustrates the observed test statistic, a histogram approximation to the null distribution and the associated Monte Carlo p-value based on 999 simulations for odd numbers of nearest neighbors $k = 3, 5, 7, 9, 11,$ and 13. None of the sets of nearest neighbors considered suggest any statistically significant clustering of late sites among the set of early and late sites combined.

In the spirit of our discussion of Figure 16.1, the lack of statistically significant clustering of one type of events among its nearest neighbors does not necessarily preclude the existence of a more general definition of clustering among sites. In addition, since clustering represents a feature averaged over the entire data set, non-significant clustering also does not preclude the existence of a few isolated clusters within the data set. We next consider both options with other analytic approaches.

16.6.2. Second-order measures and spatial scale

The nearest neighbor approach above explores clustering of event types among the sets of nearest neighbors but ignores inter-event distances. Statistical estimation of evidence for clustering as a function of distance provides an approach that addresses the question of clustering in a slightly different manner.

The most commonly used distance-based approach for assessing clustering among a single set of events is the so-called K function originally due to Ripley (1977).

The K function is a function, $K(d)$, of distance d defined as the average number of additional events observed within distance d of a randomly chosen event, scaled by the overall intensity (average number of events per unit area). As a result, we could estimate the K function via:

$$\widehat{K}(d) = \widehat{\lambda}^{-1} \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \delta(d(i,j) < d) \quad (16.1)$$

where N represents the number of observed events, $\widehat{\lambda}$ is an estimate the overall intensity of events, $d(i, j)$ denotes the distance between events i and j, and $\delta(d(i, j) < d) = 1$ if $d(i, j) < d$ and 0 otherwise. Note that the intensity λ is assumed to be constant so that any pattern in the events will be described within the K function rather than as a spatially heterogeneous intensity function. In practice, we should make some adjustment for events observed near the edge of the study area since events occurring nearby but outside of the study area will not

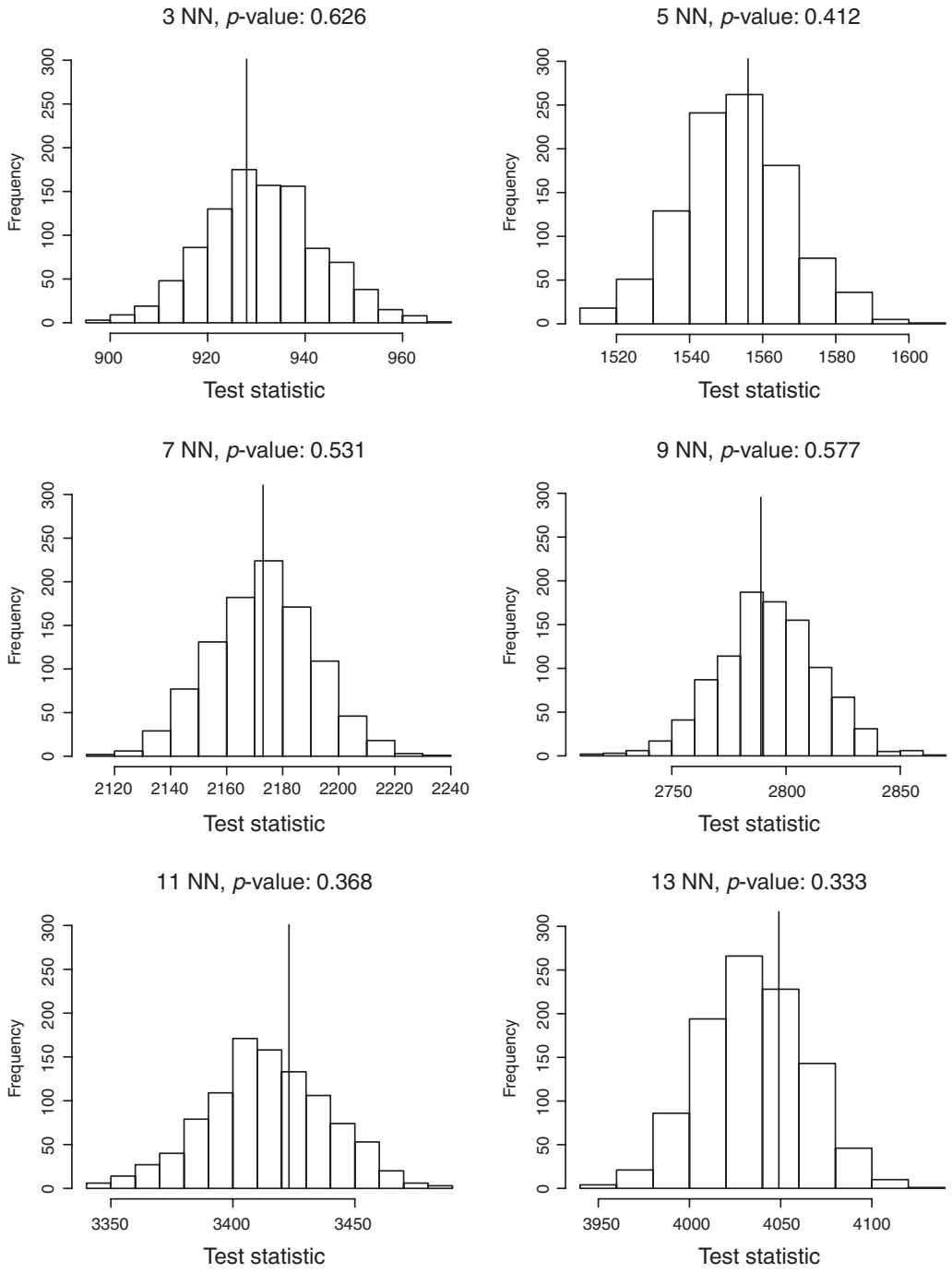


Figure 16.4 Histograms and associated *p*-values of the cumulative number of late events among the nearest neighbors of early events based on 999 random labelling simulations for the Anasazi data set.

be observed. An 'edge corrected' (ec) version is provided by

$$\widehat{K}_{\text{ec}}(d) = \widehat{\lambda}^{-1} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (w_{ij})^{-1} \delta(d(i, j) < d) \tag{16.2}$$

where the average is replaced by a weighted average with weight w_{ij} defined as the proportion of the circumference of the circle centered at event i with radius $d(i, j)$ which lies within the study area. With a constant intensity, w_{ij} denotes the conditional probability of an event occurring at distance $d(i, j)$ from event i falling within the study area, given the location of event i. Note that $w_{ij} = 1$ if the distance between events i and j is less than the distance between event i and the edge of the study area.

Under CSR, $K(d) = \pi d^2$ (the area of a circle with radius d and patterns exhibit clustering for $K(d) > \pi d^2$. To simplify the graphical expression of the K function, Besag (1977) proposed a transformation:

$$\widehat{L}(d) - d = \left(\frac{\widehat{K}_{\text{ec}}(d)}{\pi} \right)^{1/2} - d$$

where the first term on the right-hand side equals d under CSR, so subtracting d yields a CSR-associated reference value of zero. Plotting d versus $\widehat{L}(d) - d$ allows us to quickly identify distances at which patterns exhibit clustering ($\widehat{L}(d) - d > 0$) and those at which patterns appear too evenly spaced to be consistent with CSR ($\widehat{L}(d) - d < 0$).

The thick line in Figure 16.5 provides a graph of $\widehat{L}(d)$ for the late Anasazi sites. The transformed K function is well above the CSR reference value of zero indicating more clustering that we would expect under CSR. However, the question of interest is not 'Do the late sites appear consistent with CSR?' but rather 'Do the late sites exhibit more clustering than the early sites?' We can use a random labelling Monte Carlo approach to address this question by repeatedly sampling 390 sites from the set of early and late sites combined, estimating the K function and exploring the variability of these estimates. Figure 16.5 illustrates the pointwise median, 2.5th and 97.5th percentiles of estimates of $\widehat{L}(d) - d$, based on 999 random labelling samples. We note that the estimate based on the data falls well within the band of values likely under the random labelling hypothesis so that the observed set of late sites does not differ from the patterns expected under random labelling in a statistically significantly way.

At this point, the pattern of the late sites does not appear to differ significantly from the pattern of the early sites either in its observed nearest neighbor relationships or its distance-based associations. However, both approaches applied so far explore clustering and we next consider approaches to evaluate the possible existence of clusters within the late sites.

16.7. DETECTING CLUSTERS

We consider two conceptual approaches for detecting clusters, namely, the detection of the most unusual collection of events, and the comparison of the distribution of event locations experiencing the phenomenon of interest (e.g., a disease case or a crime), to that of locations not experiencing the phenomenon (controls). These two approaches cover many but not all examples and we refer the interested reader to texts by Lawson (2001), Elliott *et al*. (1992, 1999), and Waller and Gotway (2004) for additional approaches and techniques.

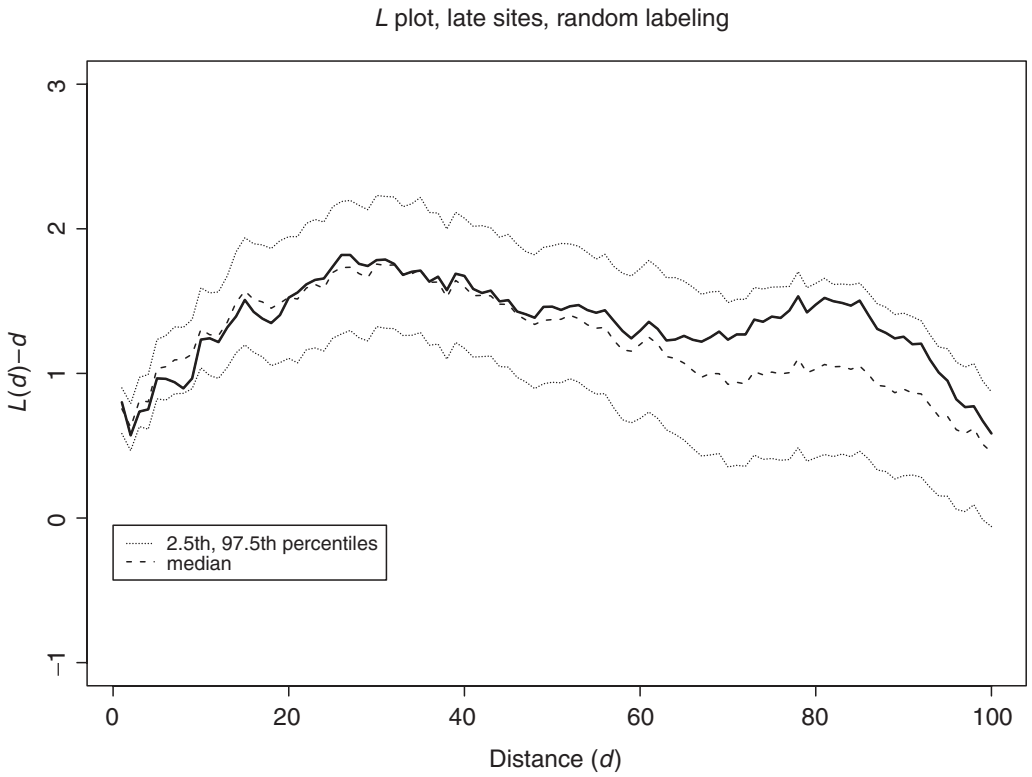


Figure 16.5 The estimate of the standardized *K* function ($\hat{L}(d)$) for the late Anasazi sites (solid line) compared to the median (dashed) and 95 percent tolerance bands based on 999 random labeling simulations.

16.7.1. Finding the oddest ball in the urn: Scan statistics

If we consider a cluster to be defined by an 'unusual' collection of events, then an initial place to start is with methods designed to detect the most unusual collection (or collections) of events observed within the data set. Such methods define a (large) set of 'potential clusters', collections of events each of which we might define as a cluster if the collection appears unusual enough (discrepant from the null model of interest), then identify the most unusual of these.

This general idea motivated the 'geographical analysis machine' (GAM) of Openshaw *et al*. (1988) where potential clusters were defined as collections of events falling within circular buffers of varying radii. The buffers were centered at each point in a fine grid covering the study area and the GAM approach mapped any circle whose collection of events was detected as unusual, e.g., those circles where the number of events exceeded the 99.8th percentile of a Poisson distribution with mean defined by the population size within the buffer multiplied by the overall disease risk. (The use of the 99.8th percentile was an attempt to adjust for the extremely high number of hypothesis tests conducted, one for each potential cluster.)

The GAM received a fair amount of attention, both in applications and in criticisms of the relatively *ad hoc* statistical inference associated with it. Subsequent methods proposed by Besag and Newell (1991) and Turnbull *et al*. (1990) revised the basic idea in more statistically-based ways, but the most widely-used variant of this general idea is the spatial scan

statistic originally proposed by Kulldorff (1997) and freely available in the software package SaTScan (Kulldorff and Information Management Services Inc., 2002).

The spatial scan statistic works as follows. The set of potential clusters consists of all circular collections of cases centered at observed cases or controls, and radii ranging from the minimum observed inter-event distance to radii containing approximately one-half of the study area. For each potential cluster, we measure its 'unusualness' via a local likelihood ratio statistic comparing a null hypothesis that events arise within the potential cluster with the same probability as they do outside of the potential cluster to an alternative hypothesis where events arise within the potential cluster with a higher probability than outside of the potential cluster. If we assume events follow a Poisson process within and without the potential cluster, we are simply testing the null hypothesis of equal intensities within and without the potential cluster versus the alternative hypothesis of a greater intensity within the potential cluster. In this case, the local likelihood ratio statistic becomes:

$$\left(\frac{N_{1,in}}{N_{in}}\right)^{N_{1,in}}\left(\frac{N_{1,out}}{N_{out}}\right)^{N_{1,out}} I\left(\frac{N_{1,in}}{N_{in}} > \frac{N_{1,out}}{N_{out}}\right) \tag{16.3}$$

where $N_{1,in}$ and $N_{in} = \left(N_{0,in} + N_{1,in}\right)$ denote the number of event locations and persons at risk (number of event *and* control locations) within the potential cluster, respectively, and $N_{1,out}$ and $N_{out} = \left(N_{0,out} + N_{1,out}\right)$ for outside of the potential cluster. By extending the statistic with the inclusion of the indicator function $I(\cdot)$ we can limit attention to only windows where the observed rate inside the window exceeds that outside the window, rather than including 'cold spots' where the rate inside the window is less than that outside the window.

At this point, we have a value measuring the unusualness of each potential cluster. Next, we identify the potential cluster with the highest local likelihood ratio statistic as the 'most likely cluster' among the set of potential clusters considered.

Next, we determine the statistical significance of this most likely cluster, an important step since there will always be a most likely cluster, i.e., the most unusual collection of events considered. The relevant question is: How unusual is this most unusual collection of events? Kulldorff (1997) addresses this question in a clever way using Monte Carlo hypothesis testing. Given the total number and locations of events of both types (those experiencing the phenomenon and those not), we randomly assign 'events' among the set of all locations, find the most likely cluster and save its associated likelihood ratio statistic. We repeat this exercise many times and construct a histogram of the maximum local likelihood ratio statistic for each random allocation. We estimate the statistical significance of the most likely cluster detected in our data set by the proportion of simulated maximized local likelihood ratio test statistics exceeding that of the observed data (i.e., the proportion, under the random labelling null hypothesis, of measures of unusualness that are more unusual than observed in the data).

This approach avoids the multiple testing problem encountered in Openshaw *et al.*'s (1988) GAM in the following way. The key lies in comparing the measure of unusualness of the most likely cluster in the observed data (the maximized local likelihood ratio statistic) to the same value from each of a large number of data sets simulated under the null hypothesis. Each simulated assignment generates its own most likely cluster and associated local likelihood ratio statistic. These values are independent of one another since the simulated data sets are independent of one another, so the collection of maximum

local likelihood ratio statistics represents an independent sample under the null hypothesis and its histogram provides an estimate of the null distribution of the maximized local likelihood ratio statistic. Note that this approach compares the maximized likelihood statistic regardless of where it occurs rather than comparing the measure of unusualness at its observed location to the measures of unusualness *at that same location*.

We can contrast these two approaches by considering the questions answered by each. By comparing the observed measure of unusualness to the measure observed anywhere in the simulated data sets we answer 'How unusual does our most likely cluster appear compared to how unusual the most likely cluster appears under the null hypothesis?' If we compare the observed measure at a particular location to the observed measure at that location in each of the simulated data sets, we answer 'How unusual does our most likely cluster appear compared to any other cluster at this location?' The first question represents a single question particular to the most likely cluster but the second is particular to a location and radius. Openshaw *et al*.'s (1988) GAM and similar methods essentially ask the second question for each location and radius which generates multiple hypothesis tests and complicates inference, again illustrating the importance of Figure 16.1.

To illustrate the spatial scan statistic, Figure 16.6 shows the most likely cluster of late sites in the Anasazi data by the thick, dark circle and the most likely cluster of early sites by the thin, light circle. Neither is statistically significant. Even though the most likely cluster of late sites consists of only one early site (on the edge), the late sites outnumber the early sites in the data set so this is not a particularly unusual grouping of events.

A few items merit mention. First, note that seeking the most likely cluster of late sites is a different exercise than seeking the most likely cluster of early sites. In some applications it is clear which events one wishes to find a cluster of (e.g., cases versus non-case controls in epidemiology); in others it is not as obvious and both questions are of interest. Second, we must interpret the results in light of the set of potential clusters considered. Here, we only consider circular clusters and may miss more oblong or sinuous clusters, perhaps following rivers. The most recent version of SaTScan incorporates elliptical potential clusters and recent methodological work by Assunção (2006) and Patil and Tallie (2004) further expand the set of potential clusters at increasing computational cost. The impact of expanding the set of potential clusters on the statistical power of detection for subsets of this class remains to be studied in detail. For instance, it is not known to what extent including both circular and elliptical clusters might reduce the power to detect only elements of the subset of circular clusters.

16.7.2. Finding peaks and valleys: Estimating the spatial intensity

The spatial scan statistic is appealing, but is limited to the set of potential clusters. A more general approach involves estimation of the intensity function associated with a set of observed event locations. The conceptual framework of a spatial point process views the set of observed locations as a realization of a random distribution in space. The next step involves estimating the local probability of an event occurrence and defining clusters as areas where events appear to be most likely.

Kernel estimation is a popular approach for estimating probability distributions and has seen broad use in spatial analysis as well (Bailey and Gatrell, 1995;

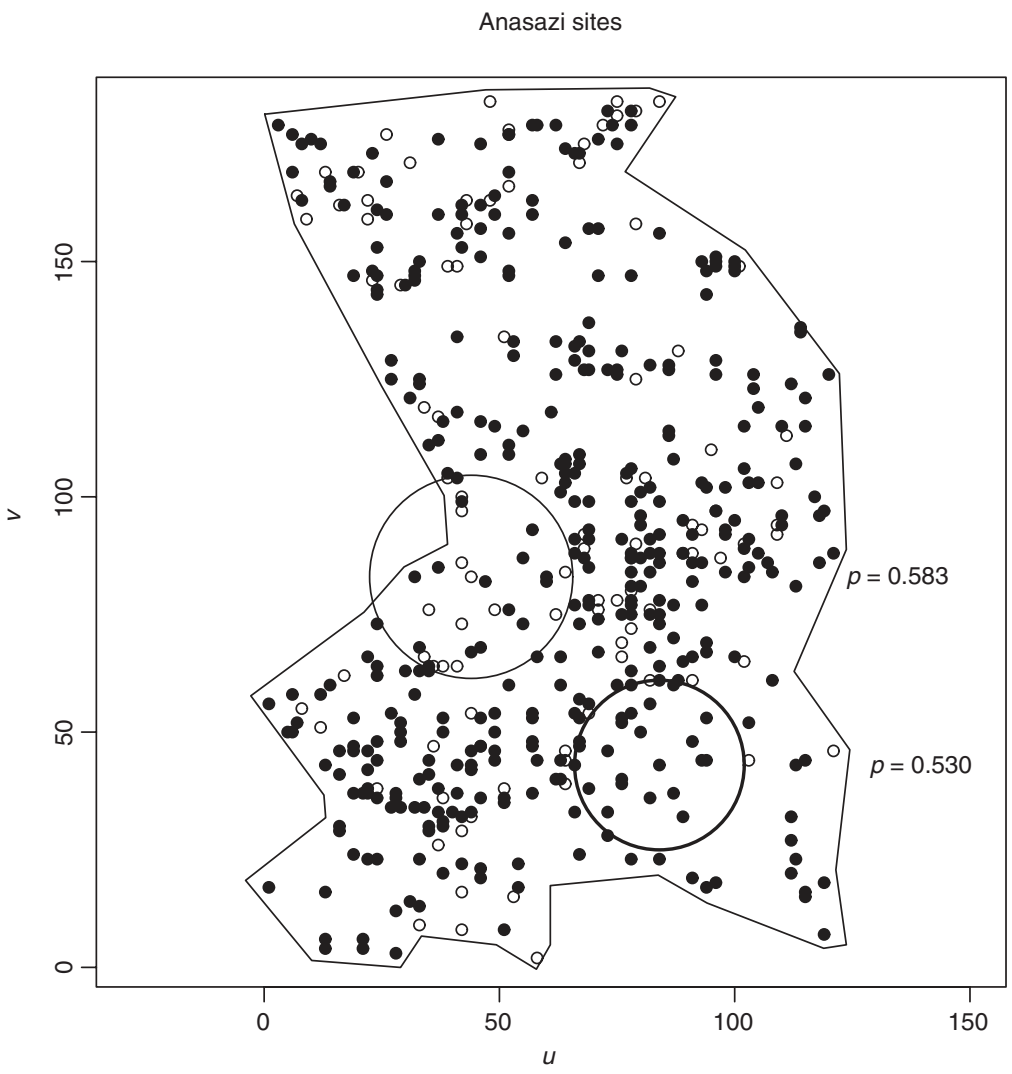


Figure 16.6 SaTScan results for the Anasazi data set. Thick, dark circles and *p*-values correspond to the most likely clusters of late sites, and thin, light circles and *p*-values correspond to the most likely clusters of early sites.

McLafferty *et al*., 2000; Diggle, 2003; Eck *et al*., 2005). Conceptually, suppose we place an equal amount of soft modeling clay over each event location on our map. These will overlap for events near each other and the resulting height of the entire surface represents our estimate of the spatial intensity, higher in areas with many observed events, lower in areas with few observed events. More precisely, we place a smooth, symmetric function (the 'kernel') over events, typically a probability density function such as a bivariate Gaussian density or other function which integrates to one. At each of a fine grid of points, we sum the kernel values associated with each observed event, yielding a smooth surface estimating the unknown intensity function. The 'bandwidth' (spatial extent) of each kernel controls the overall amount of smoothness in the estimated intensity surface with larger bandwidths corresponding to smoother surfaces. Essentially, the kernel takes each

observation and 'spreads' its influence over a local area corresponding to the kernel function.

Mathematically, suppose x denotes the vector location of N events $(x_1, x_2, \ldots, x_N)$, and x denotes any location within our study area A. The kernel estimate of the intensity $\lambda(x)$ is:

$$\tilde{\lambda}(x) = \frac{1}{|A|b} \sum_{i=1}^{N} \text{Kern}\left(\frac{x - x_i}{b}\right) \quad (16.4)$$

where $|A|$ denotes the geographic area of our study area A, Kern$(\cdot)$ is a kernel function satisfying:

$$\int_A \text{Kern}(x)\,\mathrm{d}x = 1$$

and b denotes the kernel's bandwidth.

Figure 16.7 represents the two intensity estimates for the Anasazi site data for a bandwidth of 15 distance units. Visually, we observe some differences between the two intensity estimates, such as a more distinct gap between site intensity for the late period (right-hand plot) in the northern third of the study area, and perhaps an additional mode for the early period (left-hand plot) in the southwestern section of the study area. Such conclusions must be interpreted with caution however, since they are dependent upon the bandwidth used for estimation. In this illustration we use the same bandwidth in both plots to facilitate numerical comparisons between them in the next subsection, even though the two time periods contain different sample sizes.

16.7.3. Comparing maps: Contouring relative risk

Intensity estimates provide a descriptive view of local variations in the probability of event occurrence. However, as mentioned above, the interpretation of clustering depends on the

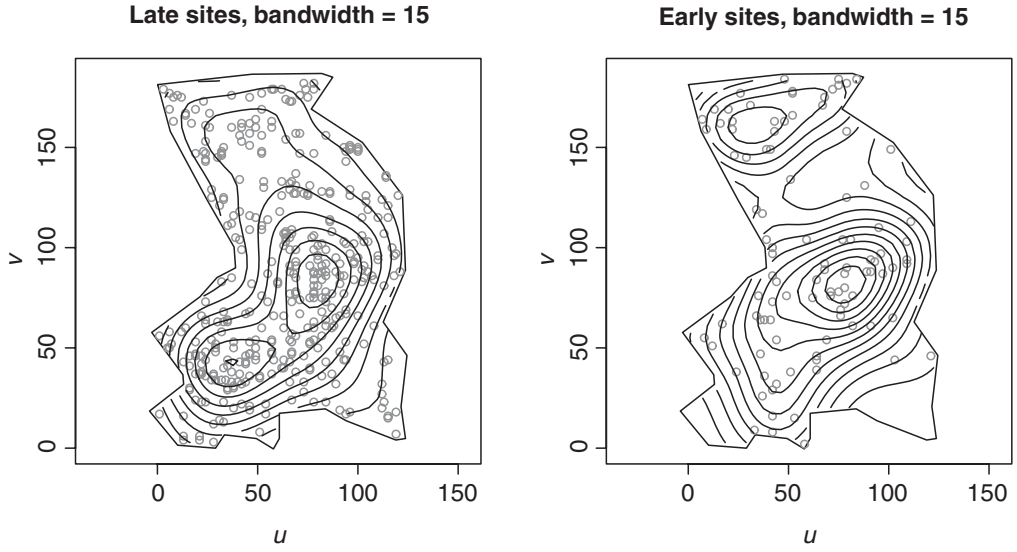


Figure 16.7 Kernel estimates of the intensity functions for the patterns of late (left) and early (right) sites for the Anasazi site data based on a bandwidth of 15 distance units.

(often spatially-varying) population at risk of an event. That is, we are often more interested in spatial variations in the risk (probability) of an event rather than in spatial variations in the actual numbers of events. For crime data, we often do not have point-level population data or samples of the locations of 'control' individuals not experiencing the crime under study, and intensity analysis concludes with interpretation of the intensity function of events (Eck *et al.*, 2005). In other fields, such comparison patterns are more readily available, and we next consider statistical identification of clusters via comparisons between two estimated intensity functions.

Suppose we have two types of events (events and controls, early or late sites, etc.). Bithell (1990), Lawson and Williams (1993), and Kelsall and Diggle (1995) propose approaches for comparing kernel estimates from each type of event, say $\tilde{\lambda}_0$ and $\tilde{\lambda}_1$. Kelsall and Diggle (1995) examine the surface generated by the natural logarithm of the ratio of the two intensity functions:

$$r(x) = \log \frac{\tilde{\lambda}_1(x)}{\tilde{\lambda}_0(x)}$$

for any location x in our study area A. To borrow a term from epidemiology, the ratio of the two intensity functions reflects the *relative risk*, and the log transformation places the ratio on a more symmetric scale around its null value (0.0 on the log scale). Kelsall and Diggle (1995) point out technical and practical reasons for using the same bandwidth for both kernel estimates, primarily to avoid confounding the smoothness of the $r(x)$ surface by differences in the underlying smoothness of the two intensity estimates.

The log relative risk surface $r(x)$ illustrates areas where events of each type are more or less likely than the other. In order to use this approach to detect clusters, we seek peaks or valleys in the surface. To assess statistical significance, the next step is to decide whether the peaks and valleys are more extreme than one would expect to observe under a null hypothesis. Kelsall and Diggle (1995) propose using random labeling simulations to determine local clusters. Suppose we have n_0 type 0 events and n_1 type 1 events. Conditional on the complete set of observations of both types of events, we randomly assign n_0 of the events to be type 0, the rest to be type 1, and calculate $r(g)$ for a grid of locations $\boldsymbol{g} = (g_1, g_2, \ldots, g_G)$. We repeat the random labeling a large number of times providing a large number of $r(g_i)$ values for each g_i in our grid, under the random labeling null hypothesis. If the value of $r(g_i)$ based on the observed data is more extreme than the 2.5th or 97.5th percentiles of the values based on the simulation, we mark the location on the map. We note that this approach provides *pointwise* inference, not overall inference due to the multitude of grid points and the correlation between values of $r(g)$ induced by the kernel function (nearby estimates share the same data).

Figure 16.1 provides a basis for comparison between the spatial scan statistic and the log relative risk surface. The scan statistic addresses the question 'Where is the most unusual collection of cases and how unusual is it compared to what would be expected of the most unusual collection under the null hypothesis?' The log relative risk surface addresses: 'Where are different types of events more or less likely than others and how do these differences compare to what we would expect under the null hypothesis?' One important distinction between these two questions is the emphasis on a single cluster in the first and the emphasis on the entire log relative risk surface in the second. For instance, a focus on a single

cluster ignores the size, number, and location of other local peaks and valleys across the surface. Also, if we were to use the pointwise interval inference to identify a single most likely cluster from the log relative risk surface we would fall into the same multiple inference problem as discussed above for GAM-type methods. Instead, we should think of the collection of pointwise intervals as a general guide to describe the variability (under the null hypothesis) of the estimated log relative risk surface across the study area, and draw attention to locations where the estimated log relative risk surface wanders outside of these bounds. Leong (2005) recently proposed and compared several approaches to move from pointwise to simultaneous intervals around such log relative risk functions in one dimension and extensions to higher dimensions would provide a stronger basis for inference.

To illustrate the approach, Figure 16.8 illustrates the log relative risk of late versus early sites based on the kernel intensity estimates shown in Figure 16.7. On the contour plot, we indicate grid points with local relative risk estimates falling above and below the 95 percent tolerance intervals (defined by random labeling) by '+' and '−' symbols, respectively. We see locally statistically significant increases in the relative probability of late versus early sites in the north-central area mentioned in our discussion of Figure 16.7.

How can we reconcile the locally significant cluster shown in Figure 16.8 with the non-significant most likely cluster found by the spatial scan statistic in Figure 16.6? Closer examination of Figure 16.6 reveals that the collection of late sites (filled circles) driving the cluster identified in the log relative risk plot is an oblong concentration of late sites in the north central portion of

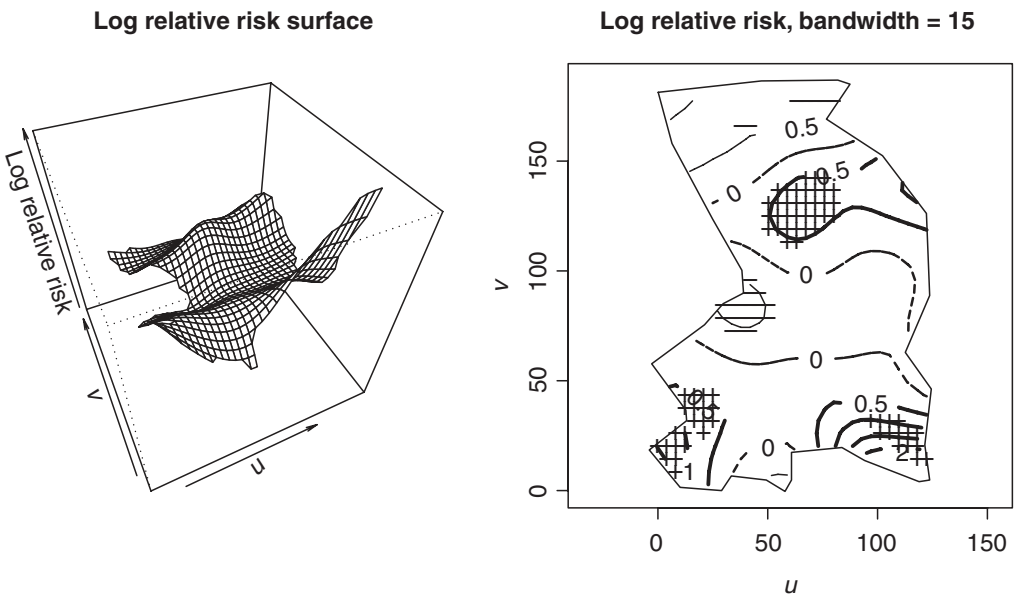


Figure 16.8 Log relative risk surface comparing the probabilities of late sites versus that of early sites for the Anasazi site data based on a bandwidth of 15 distance units. On the contour plot, '+' denotes a point exceeding the upper 95 percent pointwise tolerance limits and '−' a point exceeding the lower 95 percent limit (see text).

the study area. This concentration would not be considered among the circular potential clusters we used in our application of the spatial scan statistic. The example illustrates the importance of understanding the types of clusters evaluated by a particular method when comparing results between different approaches. In addition, the most likely clusters identified by the spatial scan statistic do not appear as unusual peaks in the log relative risk surface since (as with the scan statistic) there is not a strong excess of early or late sites in these locations.

16.8. CONCLUSIONS

The sections above illustrate the importance of understanding what sort of spatial patterns statistical approaches investigate in studies to detect clusters and/or clustering. The data set provides an interesting example where we observe no significant clustering but a significant cluster, provided we examine a broad enough class of potential clusters. Figure 16.1 illustrates that the example is not simply a situation of applying multiple methods until we get the answer we desire, but rather an example of the sorts of patterns *not* considered by many common summaries of spatial pattern, and how some potentially interesting patterns may be missed by some methods.

ACKNOWLEDGMENTS

Thanks to John Richardson, a toxicologist for US EPA Region IV, who provided the initial sketch that became Figure 16.1. In a simple diagram, he provided a summary of many important issues relating to the cluster/clustering detection problem.

This work is supported in part by grant NIEHS R01 ES007750. The opinions expressed herein are solely those of the author and may not reflect those of the National Institutes of Health or the National Institute of Environmental Health Sciences.

REFERENCES

Anselin, L. (1995). Local indicators of spatial association: LISA. *Geographical Analysis*, **27**: 93–116.

Assunção, R., Costa, M., Tavares, A. and Ferreira, S. (2006). Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, **25**: 723–742.

Bailey, T.C. and Gatrell, A.C. (1995). *Interactive Spatial Data Analysis*. New York: John Wiley and Sons.

Barnard, G.A. (1963). Contribution to the discussion of Professor Bartlett's paper. *Journal of the Royal Statistical Society, Series B*, **25**: 294.

Bartlett, M.S. (1964). The spectral analysis of two-dimensional point processes. *Biometrika*, **51**: 299–311.

Besag, J. (1977). Discussion of 'Modeling spatial patterns' by B.D. Ripley. *Journal of the Royal Statstical Society, Series B*, **39**: 192–225.

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, **154**: 327–333.

Bithell, J. (1990). An application of density estimation to geographical epidemiology. *Statistics in Medicine*, **9**: 691–701.

Chainey, S. (2005). Methods and techniques for understanding crime hot spots. In: Mapping Crime: Understanding Hot Spots. Eck, J.E., Chainey, S., Cameron, J.G., Leitner, M. and Wilson, R.E. (eds.), National Institute of Justice Report NCJ 209393. Washington DC: United States Department of Justice, Office of Justice Programs, pp. 15–34.

Cliff, A.D. and Ord, J.K. (1973). *Spatial Autocorrelation*. London: Pion.

Cressie, N.A.C. (1993). *Statistics for Spatial Data,* Revised Edition. New York: John Wiley and Sons.

Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations (with discussion). *Journal of the Royal Statistical Society, Series B*, **52**: 73–104.

Denison, D. and Holmes, C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics*, **57**: 143–147.

Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns,* Second Edition. New York: Oxford University Press.

Eck, J.E., Chainey, S., Cameron, J.G., Leitner, M. and Wilson, R.E. (2005). *Mapping Crime: Understanding Hot Spots*. National Institute of Justice Report NCJ 209393. Washington DC: United States Department of Justice, Office of Justice Programs.

Elliott, P., Cuzick, J., English, D. and Stern, R. (1992). *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*. Oxford: Oxford University Press.

Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. (1999). *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford University Press.

Goldsmith, V., McGuire, P.G., Mollenkopf, J.H. and Ross, T.A. (2000). *Analyzing Crime Patterns: Frontiers of Practice*. Thousand Oaks, CA: Sage Publications, Inc.

Gumerman, G.J. (1970). *Black Mesa: Survey and Excavation in Northeastern Arizona, 1968*. Prescott College Press.

Gumerman, G.J., Westfall, D. and Weed, C.S. (1972). *Archaeological Investigations on Black Mesa: The 1969–1970 Seasons*. Prescott College Press.

Kelsall, J. and Diggle, P.J. (1995). Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine*, **14**: 2335–2342.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, **56**: 13–21.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**: 1487–1496.

Kulldorff, M. and Information Management Services, Inc. (2002). *SaTScan v. 3.0: Software for the Spatial and Space-time Scan Statistics*. Bethesda, MD: National Cancer Institute.

Kulldorff, M., Tango, T. and Park, P.J. (2003). Power comparisons for disease clustering tests. *Statistics in Medicine*, **42**: 665–684.

Langworthy, R.H. and Jefferis, E.S. (2000). The utility of standard deviation ellipses for evaluating hot spots. In: *Analyzing Crime Patterns: Frontiers of Practice.* Goldsmith, V., McGuire, P.G., Mollenkopf, J.H. and Ross, T.A. (eds.), Thousand Oaks, CA: Sage Publications, Inc.

Lawson, A.B. (2001). *Statistical Methods in Spatial Epidemiology*. Chichester: John Wiley & Sons.

Lawson, A.B. and Denison, D.G.T. (2002). *Spatial Cluster Modelling*. Boca Raton FL: Chapman & Hall/CRC.

Lawson, A.B. and Williams, F.L.R. (1993). Applications of extraction mapping in environmental epidemiology. *Statistics in Medicine*, **12**: 1249–1258.

Leong, T. (2005). First- and second-order properties of spatial point processes in biostatistics. Unpublished Ph.D. dissertation, Department of Biostatistics, Rollins School of Public Health, Emory University. Atlanta, GA.

McLafferty, S., Williamson, D. and McGuire, P.G. (2000). Identifying crime hot spots using kernel smoothing, In: *Analyzing Crime Patterns: Frontiers of Practice.* Goldsmith, V., McGuire, P.G., Mollenkopf, J.H. and Ross, T.A. (eds.), Thousand Oaks, CA: Sage Publications, Inc.

Møller, J. and Waagepetersen, R. (2002). *Statistical Inference and Simulation for Spatial Point Patterns*. Boca Raton, FL: Chapman & Hall/CRC.

Openshaw, S., Craft, A.W., Charlton, M. and Birch, J.M. (1988). Investigation of leukaemia clusters by use of a geographical analysis machine. *Lancet*, **1** (8580): 272–273.

Patil, G.P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, **11**: 183–197.

Plog, S. (ed.) (1986). *Spatial Organization and Exchange: Archaeological Survey on Northern Black Mesa*. Southern Illinois University Press.

Powell, S. and Smiley, F.E. (2002). *Prehistoric Culture Change on the Colorado Plateau: Ten Thousand Years on Black Mesa*. Tucson AZ: The University of Arizona Press.

Ripley, B.D. (1977). Modeling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**: 172–212.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46**: 234–240.

Turnbull, B.W., Iwano, E.J., Burnett, W.S., Howe, H.L. and Clark, L.C. (1990). Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. *American Journal of Epidemiology*, **132**, supplement: S136–S143.

Waller, L.A. and Jacquez, G.M. (1995). Disease models implicit in statistical tests of disease clustering. *Epidemiology*, **6**: 584–590.

Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Analysis of Public Health Data*. Hoboken NJ: John Wiley & Sons.

17

Bayesian Spatial Analysis

Andrew B. Lawson and Sudipto Banerjee

17.1. INTRODUCTION

Spatially referenced data occur in diverse scientific disciplines including geological and environmental sciences (Webster and Oliver, 2001), ecological systems (Scheiner and Gurevitch, 2001), disease mapping (Lawson, 2006) and in broader public health contexts (Waller and Gotway, 2004). Very often, such data will be referenced over a fixed set of locations in a region of study. These locations can be with regions or areas with well-defined neighbors (such as pixels in a lattice, counties in a map, etc.), whence they are called *areally referenced* or *lattice* data. Alternatively, they may be simply points with coordinates (latitude–longitude, Easting–Northing etc.), in which case they are called *point referenced* or *geostatistical*. Statistical theory and methods to model and analyze such data depend upon these configurations and has enjoyed significant developments over the last decade; see, for example, the books by Cressie (1993), Chilés and Delfiner (1999), Móller and Waagpetersen (2004), Schabenberger and Gotway (2004), and Banerjee *et al*. (2004) for a variety of methods and applications.

With recent advances in computational methods (particularly in the area of Monte Carlo algorithms), it is now commonplace to be able to incorporate spatial correlation as an important modeling ingredient. It is now feasible to fit routinely linear models with a variety of features within a modeling hierarchy. With the implementation of fast algorithms such as Markov Chain Monte Carlo (MCMC), sophisticated models that were previously inaccessible are now within reach allowing us to move beyond the simpler, and often inadequate, descriptive measures for analyzing spatial structure.

Spatial analysis can be viewed in a number of ways. For the statistician, there are two basic approaches to statistical modeling and inference: frequentist or likelihood based

inference, and Bayesian inference. Here we focus on the latter approach. Bayesian inference and modeling can be seen as an extension of likelihood methods, but it also has a fundamentally different view of the inferential process.

17.2. NOTATION

The following notation will be used throughout this chapter. A random variate is denoted y_i, for an item in a vector. The vector of these items is $\mathbf{y}$. Often $\mathbf{y}$ will be related to independent variables (such as in a linear model). In that case the matrix of such variables can be defined as X. A linear model can be defined, for a single independent variable $\mathbf{x}_1$ as:

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i.$$

In general, the matrix formulation of the model, where $i = 1, \ldots, n$ will be:

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e} \tag{17.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of the dependent variable, X is an $n \times p$ matrix of p independent predictors (or covariates), $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector of the corresponding slopes and $\mathbf{e}$ is an $n \times 1$ vector of the errors. Often we make distributional assumptions, such as $\mathbf{e} \sim N(\mathbf{0}, \Sigma)$ These expressions imply that the errors are normally distributed with a zero-vector, $\mathbf{0}$, as the mean and a covariance matrix $\boldsymbol{\Sigma}$.

17.2.1. Point-referenced spatial data notation

As we will be dealing with spatial data, we will require some notation specific to such settings. When the referencing is done using coordinates (latitude–longitude, Easting–Northing, etc.) over a domain $\mathcal{D}$, we denote it as $s \in \mathcal{D}$; for instance in two-dimensional domains we have $s \equiv (s_x, s_y)$. The most frequently encountered scenario observes a spatial field measured at a finite set of locations, say $\mathcal{S} = \{s_1, \ldots, s_n\}$. We usually name this a random field, which we denote as $\{w(s) : s \in \mathcal{D}\}$ or simply as $w(s)$ in short. A realization of this random field will be a vector $\mathbf{w} = (w(s_1), \ldots, w(s_n))$.

17.2.2. Health data notation

For health data discussed in this chapter we will confine ourselves (mostly) to examining count data arising within small arbitrary administrative areas (such as census tracts, zip codes, postcodes, counties). Define y_i as the count of disease within the ith small area. Assume that $i = 1, \ldots, m$. For this we need to define a relative risk for the ith region: θ_i. We usually want to make inferences about the relative risk, in any study.

We also usually have available an expected rate for the ith region: e_i. Often the count within the regions will have a Poisson distribution, i.e., $y_i \sim Pois(e_i\theta_i)$.

17.3. LIKELIHOOD AND BAYESIAN MODELS

17.3.1. Likelihood

A random variable X is usually associated with a distribution which governs its behavior. We denote this distribution as $f(x \mid \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a parameter. In general, θ could be a vector of parameters and so is denoted $\boldsymbol{\theta}$. In this case we have $f(x \mid \boldsymbol{\theta})$. When a random sample of values of X are taken $\{x_i, i = 1, \ldots, n\}$ then the likelihood is

defined as the joint distribution of the sample values:

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}). \qquad (17.2)$$

It is assumed that conditional on $\boldsymbol{\theta}$ the sample values are independent. If this were not so, then we would require to take the product of conditional distributions in equation (17.2). When using the frequentist inferential process it is important to base decisions about parameters (estimation of parameter values or confidence intervals) on the likelihood function. Maximum likelihood estimation seeks point estimates of the parameters in $\boldsymbol{\theta}$ by maximising $f(\mathbf{x} \mid \boldsymbol{\theta})$ or $\log f(\mathbf{x} \mid \boldsymbol{\theta})$. Testing and interval estimation is often based on likelihood ratios derived for different values of $\boldsymbol{\theta}$ under different hypotheses. Inference for quantities such as confidence intervals is based on the concept of repeated experimentation, in that probability statements are derived based on properties of repeated sequences of experiments.

17.4. BAYESIAN INFERENCE

Fundamental philosophical differences with the frequentist approach are found when a Bayesian perspective is assumed. First of all, parameters within Bayesian models are assumed to be random variables and hence are governed by distributions themselves. Hence, there is no longer a fixed (true) value for a given parameter. Instead an expected value or other functional of a distribution can be defined. Because parameters have distributions then the likelihood previously defined must be extended to accommodate these distributions.

By modeling both the observed data and any unknown parameter or other unobserved effects as random variables, the hierarchical Bayesian approach to statistical analysis provides a cohesive framework for combining complex data models and external knowledge or expert opinion (e.g., Berger, 1985; Carlin and Louis, 2000; Robert, 2001; Gelman *et al.*, 2004; Lee, 2005) In this approach, in addition to specifying the distributional model $f(\mathbf{y} \mid \boldsymbol{\theta})$ for the observed data $\mathbf{y} = (y_1, \ldots, y_n)$ given a vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, we suppose that $\boldsymbol{\theta}$ is a random quantity sampled from a *prior* distribution $p(\boldsymbol{\theta} \mid \lambda)$, where λ is a vector of hyperparameters. Inference concerning $\boldsymbol{\theta}$ is then based on its *posterior* distribution:

$$p(\boldsymbol{\theta} \mid \mathbf{y}, \lambda) = \frac{p(\mathbf{y}, \boldsymbol{\theta} \mid \lambda)}{p(\mathbf{y} \mid \lambda)} = \frac{p(\mathbf{y}, \boldsymbol{\theta} \mid \lambda)}{\int p(\mathbf{y}, \boldsymbol{\theta} \mid \lambda)\, \mathrm{d}\boldsymbol{\theta}}$$

$$= \frac{f(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \lambda)}{\int f(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \lambda)\, \mathrm{d}\boldsymbol{\theta}}. \qquad (17.3)$$

Notice the contribution of both the data (in the form of the likelihood $f(\mathbf{y} \mid \boldsymbol{\theta})$) and the external knowledge or opinion (in the form of the prior $p(\boldsymbol{\theta} \mid \lambda)$) to the posterior. If λ is known, this posterior distribution is fully specified; if not, a second-stage prior distribution (called a *hyper-prior*) may be specified for it, leading to a *fully Bayesian* analysis. Alternatively, we might simply replace λ by an estimate $\hat{\lambda}$ obtained as the value which maximizes the marginal distribution $p(\mathbf{y} \mid \lambda)$ viewed as a function of λ. Inference proceeds based on the estimated posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y}, \hat{\lambda})$, obtained by plugging $\hat{\lambda}$ into equation (17.3). This is called an *empirical Bayes* analysis and is closer to maximum likelihood estimation techniques.

The Bayesian decision-making paradigm improves on the classical approaches to statistical analysis in its more philosophically sound foundation, its unified approach to data analysis, and its ability to formally incorporate prior opinion or external

empirical evidence into the results via the prior distribution. Statisticians, formerly reluctant to adopt the Bayesian approach due to general skepticism concerning its philosophy and a lack of necessary computational tools, are now turning to it with increasing regularity as classical methods emerge as both theoretically and practically inadequate. Modeling the θ_is as random (instead of fixed) effects allows us to induce specific (e.g., spatial, temporal or more general) correlation structures among them, hence among the observed data y_i as well. Hierarchical Bayesian methods now enjoy broad application in the analysis of complex systems, where it is natural to pool information across different sources e.g., Gelman *et al*. (2004).

Modern Bayesian methods seek complete evaluation of the posterior distribution using simulation methods that draw samples from the posterior distribution. This sampling-based paradigm enables *exact* inference free of unverifiable asymptotic assumptions on sample sizes and other regularity conditions. A computational challenge in applying Bayesian methods is that for many complex systems, the simulations required to do inference under equation (17.3) generally involve distributions that are intractable in closed form, and thus one needs more sophisticated algorithms to sample from the posterior. Forms for the prior distributions (called *conjugate* forms) may often be found which enable at least partial analytic evaluation of these distributions, but in the presence of nuisance parameters (typically unknown variances), some intractable distributions remain. Here the emergence of inexpensive, high-speed computing equipment and software comes to the rescue, enabling the application of recently developed MCMC integration methods, such as the Metropolis–Hastings algorithm (Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984; Robert and Casella, 2005). Univariate MCMC algorithms are particularly attractive for general purpose implementation, since all that is required is the ability to sample easily from each parameter's complete conditional distribution, namely $p(\theta_i \mid \mathbf{y}, \theta_{j\neq i})$, $i = 1, \ldots, k$. The recently developed `WinBUGS` language (`www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`) and the `R` statistical platform (`www.r-project.org`) with its Bayesian packages are promising steps towards a general purpose software package for hierarchical modeling, though it may be insufficiently general in some advanced analysis settings, and in any case more work is needed before it is suitable for routine use by statistical support staff.

Statistical prediction in Bayesian settings is particularly elegant and intuitive. Let $\mathbf{y}_{\text{pred}}$ denote the random variables (they can be a collection) we seek to predict. Then, we simply treat $\mathbf{y}_{\text{pred}}$ as a random variable whose *prior*, conditional upon the parameters, is the data likelihood $f(\mathbf{y} \mid \boldsymbol{\theta})$. Then, all predictions will be summarized in the *posterior predictive* distribution:

$$p(\mathbf{y}_{\text{pred}} \mid \mathbf{y}) = \int f(\mathbf{y}_{\text{pred}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}) \, \mathrm{d}\boldsymbol{\theta}.$$

Once the posterior samples are available from $p(\boldsymbol{\theta} \mid \mathbf{y})$, it is routine to draw samples from $p(\mathbf{y}_{\text{pred}} \mid \mathbf{y})$ using the principle of *composition*: for each posterior draw of $\boldsymbol{\theta}$, we draw $\mathbf{y}_{\text{pred}}$ from $f(\mathbf{y}_{\text{pred}} \mid \boldsymbol{\theta})$. Details of such methods are particularly well explained in the texts by Carlin and Louis (2000) and Gelman *et al*. (2004).

17.4.1. Posterior sampling methods

Practical Bayesian modeling relies upon efficient computation of the posterior

distribution of the parameters. As mentioned above, the main computational challenge lies in evaluating the integral in the denominator of equation (17.3). This is especially compounded when $\boldsymbol{\theta}$ is multi-dimensional. Hence, instead of designing multi-dimensional integration routines, even the best of which can easily prove inadequate for several practical settings, we focus upon *sampling* from the posterior distribution, also known as *simulating* the posterior distribution. Once a posterior sample is obtained, all inference summaries (e.g., point estimates and credible intervals) are calculated using the sample. In principle, this strategy works equally well for simpler models where the posterior distribution is a standard family as well as for very complex hierarchical models where the posterior distribution is highly complex. Depending upon the complexity of the posterior distribution, the sampling strategies will vary: with a standard family we can directly draw a random sample, while with complex families more elaborate MCMC algorithms (see below) may be required.

Since the posterior distribution now describes the behavior of the parameters once the data are observed, we work with this distribution for estimation and inference. To obtain estimates of parameters this distribution must be summarized.

A simple example of this type of model in disease mapping is where the data likelihood is Poisson and there is a common relative risk parameter with a single gamma prior distribution:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\mathbf{y} \mid \theta) g(\theta)$$

where $g(\theta)$ is a gamma distribution with parameters α, β, i.e., $G(\alpha, \beta)$, and $L(\mathbf{y} \mid \theta) = \prod_{i=1}^{m} \{(e_i\theta)^{y_i} \exp(-e_i\theta)\}$ bar a constant only dependent on the data. A compact notation for this model is:

$$y_i \mid \theta \sim Pois(e_i\theta)$$

$$\theta \sim G(\alpha, \beta).$$

Here, the posterior distribution is again a Gamma and one can sample from it by simply employing a Gamma random number generator.

Another useful mechanism for posterior simulations when the posterior distribution is not a standard family arises from the principle of *composition*. This essentially observes that the joint posterior distribution of two arbitrary parameter vectors, say $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can be expressed as $P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \mathbf{y}) = P(\boldsymbol{\theta}_1 \mid \mathbf{y}) P(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \mathbf{y})$. To obtain samples from the above joint posterior distribution, we first sample $\boldsymbol{\theta}_1^{(j)}$ from the *marginal posterior* distribution $P(\boldsymbol{\theta}_1 \mid \mathbf{y})$ and then sample a $\boldsymbol{\theta}_2^{(j)}$ from the *conditional posterior* distribution $P(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(j)}, \mathbf{y})$). Repeating this for $j = 1, \ldots, M$ results in a joint posterior sample $(\boldsymbol{\theta}_1^{j}, \boldsymbol{\theta}_2^{(j)})_{j=1}^{M}$ of size M. We illustrate this principle below using the linear regression model mentioned in equation (17.1) from a Bayesian perspective. Several other examples can be found in the texts by Carlin and Louis (2000) and Gelman *et al.* (2004).

Let us suppose that we have data y_i from n experimental units, which forms our dependent variable. Suppose also that we have observed p covariates, $x_{1i}, \ldots, x_{pi}$, on the ith individual. Using matrix notations, we write:

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}; \quad \mathbf{e} \sim N(0, \sigma^2 I)$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations, X is a $n \times p$ matrix of independent

predictors with full column rank (we assume independent columns – so that covariates are not collinear), $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\mathbf{e}$ is the $n \times 1$ vector of uncorrelated normally distributed errors with common variance σ^2.

To construct a Bayesian framework, we will need to assign a prior distribution for $(\boldsymbol{\beta}, \sigma^2)$ in the above model. For illustration, consider the non-informative or *reference* prior distribution for $(\boldsymbol{\beta}, \sigma^2)$:

$$P(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}.$$

This is equivalent to a flat or Uniform prior on $(\boldsymbol{\beta}, \sigma^2)$. In hierarchical language we write the Bayesian linear regression model as:

$$\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{0}, \sigma^2 I)$$

$$\boldsymbol{\beta}, \sigma^2 \sim P(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Simple computations (see, e.g., Gelman *et al.*, 2004, Section 14.2) reveal that the marginal distribution $p(\sigma^2 \mid \mathbf{y})$ is a scaled Inv-$\chi^2(n-p, s^2)$ distribution, which is the same as the Inverse-Gamma distribution $IG((n-p)/2, (n-p)s^2/2)$ where:

$$s^2 = \frac{1}{n-p}(\mathbf{y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}})$$

with $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$ being the usual least-squares estimate (also the MLE). The distribution $P(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y})$ is $N(\hat{\boldsymbol{\beta}}, \sigma^2(X^TX)^{-1})$. In fact, here the marginal posterior distribution for $P(\boldsymbol{\beta} \mid \mathbf{y})$ can be derived in closed form as a multivariate-t distribution (see, e.g., Robert, 2001) but we outline the sampling-based perspective.

Following the principle of composition sampling, we draw, say for $j = 1, \ldots, M$, $\sigma^{2(j)} \sim IG(n-p/2, (n-p)s^2)$ followed by $\boldsymbol{\beta}^{(j)} \sim N(\hat{\boldsymbol{\beta}}, \sigma^{2j}(X^TX)^{-1})$. This yields our desired posterior sample $(\boldsymbol{\beta}^{(j)}, \sigma^{2(j)})$ with $j = 1, 2, \ldots, M$. Posterior confidence intervals and all inference will again be carried out using these samples.

17.5. HIERARCHICAL MODELS

The idea that the values of parameters could arise from distributions is a fundamental feature of Bayesian methodology and leads naturally to the use of models where parameters arise within hierarchies. In the Poisson-gamma example there is a two level hierarchy: θ has a $G(\alpha, \beta)$ distribution at the first level of the hierarchy and α will have a hyperprior distribution (h_α) as will $\beta(h_\beta)$, at the second level of the hierarchy. This can be written as:

$$y_i \mid \theta \sim Pois(e_i\theta)$$

$$\theta \mid \alpha, \beta \sim G(\alpha, \beta)$$

$$\alpha \mid \nu \sim h_\alpha(\nu)$$

$$\beta \mid \rho \sim h_\beta(\rho).$$

Clearly it is important to terminate a hierarchy at an appropriate place, otherwise one could always assume an infinite hierarchy of parameters. Usually the cut-off point is chosen to lie where further variation in parameters will not affect the lowest level model. At this point the parameters are assumed to be fixed. For example, in the gamma-Poisson model if you assume α and β were fixed then the Gamma prior would be fixed and the choice of α and β would be uninformed. The data would not inform about

the distribution at all. However, by allowing a higher level of variation i.e., hyperpriors for α, β, then we can fix the values of ν and ρ without heavily influencing the lower level variation. This allows the data to inform more about the different parameters in the lower levels of the hierarchy.

17.6. MARKOV CHAIN MONTE CARLO METHODS

Markov chain Monte Carlo (MCMC) methods are a set of methods which use iterative simulation of parameter values within a Markov chain. The convergence of this chain to a stationary distribution, which is assumed to be the posterior distribution, must be assessed.

Prior distributions for the p components of θ are defined as $g_i(\theta_i)$ for $i = 1, \ldots, p$. The posterior distribution of $\boldsymbol{\theta}$ and $\mathbf{y}$ is defined as:

$$P(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\mathbf{y} \mid \boldsymbol{\theta}) \prod_i g_i(\theta_i). \qquad (17.4)$$

The aim is to generate a sample from the posterior distribution $P(\boldsymbol{\theta} \mid \mathbf{y})$. Suppose we can construct a Markov chain with state space $\boldsymbol{\theta}_c$, where $\boldsymbol{\theta} \in \boldsymbol{\theta}_c \subset \Re^k$. The chain is constructed so that the equilibrium distribution is $P(\boldsymbol{\theta} \mid \mathbf{y})$, and the chain should be easy to simulate from. If the chain is run over a long period, then it should be possible to reconstruct features of $P(\boldsymbol{\theta} \mid \mathbf{y})$ from the realized chain values. This forms the basis of the MCMC method, and algorithms are required for the construction of such chains. A selection of recent literature on this area is found in Ripley (1987), Besag and Green (1993), Gelman *et al.* (2004), Gamerman (2000) and Robert and Casella (2005).

The basic algorithms used for this construction are:

1 the Metropolis and its extension, the Metropolis–Hastings algorithm;

2 the Gibbs Sampler algorithm.

17.6.1. Metropolis and Metropolis–Hastings algorithms

In all MCMC algorithms, it is important to be able to construct the correct *transition probabilities* for a chain which has $P(\boldsymbol{\theta} \mid \mathbf{y})$ as its equilibrium distribution. A Markov chain consisting of $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^t$ with state space Θ and equilibrium distribution $P(\boldsymbol{\theta} \mid \mathbf{y})$ has transitions defined as follows.

Define $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ as a transition probability function, such that, if $\boldsymbol{\theta}^t = \boldsymbol{\theta}$, the vector $\boldsymbol{\theta}^t$ drawn from $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is regarded as a proposed possible value for $\boldsymbol{\theta}^{t+1}$.

17.6.2. Metropolis and Metropolis–Hastings updates

In this case choose a symmetric proposal $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and define the transition probability as:

$$p(\boldsymbol{\theta}, \boldsymbol{\theta}') = \begin{cases} \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') q(\boldsymbol{\theta}, \boldsymbol{\theta}') & \text{if } \boldsymbol{\theta}' \neq \boldsymbol{\theta} \\ 1 - \sum_{\boldsymbol{\theta}''} q(\boldsymbol{\theta}, \boldsymbol{\theta}'') \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'') & \text{if } \boldsymbol{\theta}' = \boldsymbol{\theta} \end{cases}$$

where $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, P(\boldsymbol{\theta}' \mid \mathbf{y}) / P(\boldsymbol{\theta} \mid \mathbf{y})\right\}$.

In this algorithm a proposal is generated from $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and is accepted with probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')$. The acceptance probability is a simple function of the ratio of posterior distributions as a function of $\boldsymbol{\theta}$ values.

The proposal function $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ can be defined to have a variety of forms but must be an irreducible and aperiodic transition function.

Metropolis–Hastings (M–H) is an extension to the Metropolis algorithm where the proposal function is not confined to symmetry and:

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{P(\boldsymbol{\theta}' \mid \mathbf{y})\, q(\boldsymbol{\theta}', \boldsymbol{\theta})}{P(\boldsymbol{\theta} \mid \mathbf{y})\, q(\boldsymbol{\theta}, \boldsymbol{\theta}')}\right\}.$$

Some special cases of chains are found when $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ has special forms. For example, if $q(\boldsymbol{\theta}, \boldsymbol{\theta}') = q(\boldsymbol{\theta}', \boldsymbol{\theta})$ then the original Metropolis method arises and further, with $q(\boldsymbol{\theta}, \boldsymbol{\theta}') = q(\boldsymbol{\theta}')$ (i.e., when no dependence on the previous value is assumed) then:

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{w(\boldsymbol{\theta}')}{w(\boldsymbol{\theta})}\right\}$$

where $w(\boldsymbol{\theta}) = P(\boldsymbol{\theta} \mid \mathbf{y})/q(\boldsymbol{\theta})$ and $w(.)$ are importance weights. One simple example of the method is $q(\boldsymbol{\theta}') \sim$ Uniform $(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$ and $g_i(\theta_i) \sim$ Uniform $(\boldsymbol{\theta}_{ia}, \boldsymbol{\theta}_{ib})\, \forall i$; this leads to an acceptance criterion based on a likelihood ratio. Hence the original Metropolis algorithm with uniform proposals and prior distributions leads to a stochastic exploration of a likelihood surface. This, in effect, leads to the use of prior distributions as proposals. However, in general, when the $g_i(\theta_i)$ are not uniform this leads to inefficient sampling. The definition of $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ can be quite general in this algorithm and, in addition, the posterior distribution only appears within a ratio as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Hence, the distribution is only required to be known up to proportionality.

17.6.3. Gibbs updates

The Gibbs Sampler has gained considerable popularity, particularly in applications in medicine, where hierarchical Bayesian models are commonly applied (see, e.g., Gilks *et al.* (1993)). This popularity is mirrored in the availability of software that allows its application in a variety of problems (e.g., WinBUGS, MLWin, BACC). This sampler is a special case of the Metropolis–Hastings algorithm where the proposal is generated from the conditional distribution of θ_i given all other $\boldsymbol{\theta}$s, and the resulting proposal value is accepted with probability 1.

More formally, define:

$$q(\theta_j, \theta_j') = \begin{cases} p(\theta_j^* \mid \theta_{-j}^{t-1}) & \text{if } \theta_{-j}^* = \theta_{-j}^{t-1} \\ 0 & \text{otherwise} \end{cases}$$

where $p(\theta_j^* \mid \theta_{-j}^{t-1})$ is the conditional distribution of θ_j given all other $\boldsymbol{\theta}$ values (θ_{-j}) at time $t-1$. Using this definition it is straightforward to show that:

$$\frac{q(\boldsymbol{\theta}, \boldsymbol{\theta}')}{q(\boldsymbol{\theta}', \boldsymbol{\theta})} = \frac{P(\boldsymbol{\theta}' \mid \mathbf{y})}{P(\boldsymbol{\theta} \mid \mathbf{y})}$$

and hence $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1$.

17.6.4. M–H versus Gibbs algorithms

There are advantages and disadvantages to M–H and Gibbs methods. The Gibbs Sampler provides a *single* new value for each $\boldsymbol{\theta}$ at each iteration, but requires the evaluation of a conditional distribution. On the other hand the M–H step does not require evaluation of a conditional distribution but does not guarantee the acceptance of a new value. In addition, block updates of parameters are available in M–H, but not usually in Gibbs steps (unless joint

conditional distributions are available). If conditional distributions are difficult to obtain or computationally expensive, then M–H can be used and is usually available.

In summary, the Gibbs Sampler may provide faster convergence of the chain if the computation of the conditional distributions at each iteration are not time consuming. The M–H step will usually be faster at each iteration, but will not necessarily guarantee exploration. In straightforward hierarchical models where conditional distributions are easily obtained and simulated from, then the Gibbs Sampler is likely to be favored. In more complex problems, such as many arising in spatial statistics, resort may be required to the M–H algorithm.

17.6.5. Special methods

Alternative methods exist for posterior sampling when the basic Gibbs or M–H updates are not feasible or appropriate. For example, if the range of the parameters is restricted then slice sampling can be used (Robert and Casella, 2005, Ch. 7; Neal, 2003). When exact conditional distributions are not available but the posterior is log-concave then adaptive rejection sampling algorithms can be used. The most general of these algorithms (ARS algorithm; Robert and Casella, 2005, pp. 57–59) has wide applicability for continuous distributions, although they may not be efficient for specific cases. Block updating can also be used to effect in some situations. When generalized linear model components are included then block updating of the covariate parameters can be effected via multivariate updating.

17.6.6. Convergence

MCMC methods require the use of diagnostics to assess whether the iterative simulations have reached the equilibrium distribution of the Markov chain. There are a wide variety of methods now available to assess convergence of chains within MCMC. algorithms (ARS algorithm; Robert and Casella, 2005, pp. 57–59) provide recent reviews. The available methods are largely based on checking the distributional properties of samples from the chains.

17.7. MODEL GOF MEASURES

It is inevitable that our statistical analysis will entail the fitting and comparison of a variety of models. For this purpose, we will need to attend to issues concerning model adequacy and model comparison. To compare between the different models and perhaps help us choose those that provide better fits, we will use the Deviance Information Criteria (DIC) (Spiegelhalter *et al*., 2002) as a measure of model choice. The DIC has nice theoretical properties for a very wide class of likelihoods since it provides an estimate of goodness-of-fit and for model complexity and is particularly convenient to compute from posterior samples. This criterion is the sum of the Bayesian deviance (a measure of model fit) and the (effective) number of parameters (a penalty for model complexity). It rewards better fitting models through the first term and penalizes more complex models through the second term, with lower values indicating favorable models for the data. The deviance, up to an additive quantity not depending upon the parameters $\boldsymbol{\theta}$, is simply minus twice the log-likelihood, $D(\boldsymbol{\theta}) = -2 \log f(\mathbf{y} \mid \boldsymbol{\theta})$, where $f(\mathbf{y} \mid \boldsymbol{\theta})$ is the first stage likelihood for the respective model. The Bayesian deviance is the posterior mean, $\overline{D(\boldsymbol{\theta})} = E_{\boldsymbol{\theta} \mid \mathbf{y}}[D(\boldsymbol{\theta})]$, while the effective number of parameters is given by $p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$. The DIC is then given by $\overline{D(\boldsymbol{\theta})} + p_D$ and is easily computed from the posterior samples.

We also often use predictive fits to assess model performance using the posterior predictive distributions. We will employ the posterior predictive loss approach (Gelfand and Ghosh, 1998) to identify models providing the best fit. The actual computations are very similar to the predictive paradigm discussed towards the end of Section 17.2. Here, for any given model, if $\boldsymbol{\theta}$ is the set of parameters, the posterior predictive distribution of a *replicated* data set is given by:

$$P(\mathbf{y}_{\text{rep}} \mid \mathbf{y}) = \int P(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta} \mid \mathbf{y})\, \mathrm{d}\boldsymbol{\theta}$$

where $P(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})$ has the same distribution as the data likelihood. Replicated data sets from the above distribution are easily obtained by simulating a replicated data set from the above distribution. Preferred models will perform well under a decision-theoretic *balanced loss function* that penalizes both departure from corresponding observed values (lack of fit), as well as from what we expect the replicates to be (variation in replicates). Measures for these two criteria are evaluated as $G = (\mathbf{y} - \boldsymbol{\mu}_{\text{rep}})^T(\mathbf{y} - \boldsymbol{\mu}_{\text{rep}})$ and $P = \text{tr}\,(\text{Var}\,(\mathbf{y}_{\text{rep}}) \mid \mathbf{y})$, where $\boldsymbol{\mu}_{\text{rep}} = E[\mathbf{y}_{\text{rep}} \mid \mathbf{y}]$ is the posterior predictive mean for the replicated data points, and P is the trace of the posterior predictive dispersion matrix for the replicated data; both of these are easily computed from the samples drawn. Gelfand and Ghosh (1998) suggest using the score $D = G + P$ as a model selection criterion, with lower values of D indicating better models.

Using these formal statistical methods, we will be able to enhance the accuracy of the outputs of computer models, compare between them to validate an underlying scientific hypothesis and provide predictions of complex systems.

17.8. UNIVARIATE SPATIAL PROCESS MODELS

17.8.1. Ingredients of a Gaussian process

As briefly mentioned in the Introduction, modeling of point-referenced spatial data typically proceeds from a spatial random field $\{w(s) : s \in \mathcal{D}\}$, where $\mathcal{D}$ is typically an open subset of $\Re^d$ where d is the dimension; in most practical settings $d = 2$ or $d = 3$. We say that a random field is a *valid* spatial process if for an any finite collection of sites $\mathcal{S} = \{s_1, \ldots, s_n\}$ of arbitrary size, the vector $\mathbf{w} = (w(s_1), \ldots, w(s_n))$ follows a well-defined joint probability distribution.

For the practical spatial modeller, the most common specification is a *Gaussian Random Field* (GRF) or a *Gaussian Process* (GP), which additionally specifies that $\mathbf{w}$ follows a multivariate normal distribution. To be more specific, we write $w(s) \sim GP(\mu(s), C(\cdot))$ which is a Gaussian Process with a mean function $\mu(s)$, i.e., $E[w(s)] = \mu(s)$, and a *covariance function* $\text{Cov}(w(s), w(s')) = C(s, s')$. This specifies the joint distribution for a collection of sites $s_1, \ldots, s_n$ as $\mathbf{w} \sim N(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu(s_i))_{i=1}^n$ is the corresponding $n \times 1$ mean vector and $\Sigma_{\mathbf{w}} = [C(s_i, s_j)]$ is the $n \times n$ covariance matrix with (i, j)th element given by $C(s_i, s_j)$.

Clearly the covariance function cannot be just any function: it needs to ensure that the resulting $\Sigma_{\mathbf{w}}$ matrix is symmetric and positive definite. Symmetry is guaranteed as long as $C(s, s')$ is symmetric in its arguments, while functions that ensure the positive-definiteness are known as positive definite functions. The important characterization of such functions, at least from a modeler's perspective, says that a real-valued function is a valid covariance function if and only if it is the characteristic function of a symmetric random variable

(this is derived from a famous theorem due to Bochner). Further technical details about positive definite functions can be found in Cressie (1993), Chilés and Delfiner (1999) and Banerjee *et al*. (2004).

Since it is common for spatial data to consist of single observations from a site, we often need to assume *stationary* or *isotropic* processes for ensuring estimable models. Stationarity, in spatial modeling contexts, refers to the setting when $C(s, s') = C(s - s')$; that is, the covariance function depends upon the separation of the sites. Isotropy goes further and specifies $C(s, s') = C(\|s - s'\|)$, where $\|s - s'\|$ is the distance between the sites. Furthermore, we will parametrize the covariance function as $C(s - s') = \sigma^2\rho(s - s')$, where $\rho(s - s')$ is called a *correlation function* and σ^2 is a spatial variance parameter. In particular, we will use the the isotropic exponential correlation function $\rho(d, \phi) = \exp(-\phi d)$, with $d = \|s - s'\|$.

17.8.2. Bayesian spatial regression and kriging

There is an expanding literature on modeling point-referenced spatial data. The most common setting assumes a response or dependent variable $Y(s)$ observed at a generic location s, referenced by a coordinate system (e.g., UTM or lat–long), along with a vector of covariates $\mathbf{x}(s)$. One seeks to model the dependent variable in a spatial regression setting such as:

$$Y(s) = \mathbf{x}^T(s)\boldsymbol{\beta} + w(s) + \varepsilon(s). \qquad (17.5)$$

The residual is partitioned into a spatial process, $w(s)$, capturing residual spatial association, and an independent process, $\varepsilon(s)$, also known as the *nugget* effect, modeling pure errors that are independently and identically distributed as $N(0, \tau^2)$, where τ^2 is a measurement error variance or micro-scale variance. The key to incorporating spatial association is by modeling $w(s)$ as a Gaussian Process with spatial variance σ^2 and a valid correlation function $\rho(\cdot, \boldsymbol{\xi})$ with ξ representing parameters that quantify correlation decay and smoothness of the resulting spatial surface.

When we have observations, $\mathbf{y} = (Y(s_1), \ldots, Y(s_n))$, from n locations, we treat the data as a partial realization of a spatial process, modeled through $w(s)$. Hence, $w(s) \sim GP(0, \sigma^2\rho(\cdot, \phi))$, is a zero-centered Gaussian Process with variance σ^2 and a valid correlation function $\rho(d, \phi)$, which depends upon inter-site distances ($d_{ij} = \|s_i - s_j\|$) and a parameter ϕ quantifying correlation decay. Also, we assume $\varepsilon(s)$ are i.i.d. $N(0, \tau^2)$. Inferential goals include estimation of regression coefficients, spatial and nugget variances, and the strength of spatial association through distances. Likelihood-based inference proceeds from the distribution of the data, $\mathbf{y} \sim N(X\boldsymbol{\beta}, \Sigma)$, with $\Sigma = \sigma^2 R(\phi) + \tau^2 I$, where X is the covariance matrix and $R(\phi)$ is the correlation matrix with $R_{ij} = \rho(d_{ij}, \phi)$. See Cressie (1993) for details, including maximum-likelihood and restricted maximum-likelihood methods, and Banerjee *et al*. (2004) for Bayesian estimation.

Statistical prediction (kriging) at a new location s_0 proceeds from the conditional distribution of $Y(s_0)$ given the data $\mathbf{y}$. Collecting all the model parameters into $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \phi, \nu)$, we note that

$$E[Y(s_0) \mid \mathbf{y}] = \mathbf{x}(s_0)^T\boldsymbol{\beta} + \boldsymbol{\gamma}^T\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) \qquad (17.6)$$

$$\mathrm{Var}[Y(s_0) \mid \mathbf{y}] = \sigma^2 + \tau^2 - \boldsymbol{\gamma}^T\Sigma^{-1}\boldsymbol{\gamma} \qquad (17.7)$$

where $\boldsymbol{\gamma} = (\sigma^2\rho(\phi; d_{01}), \ldots, \sigma^2\rho(\phi; d_{0n}))$ and $d_{0j} = \|s_0 - s_j\|$. Classical prediction computes the BLUP (Best Linear Unbiased Predictor) by substituting maximum-likelihood estimates for the above parameters. A Bayesian solution first computes a posterior distribution $P(\boldsymbol{\theta} \mid \mathbf{y})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\xi})$ is the collection of all model parameters and then computes the posterior predictive distribution $P(Y(s_0) \mid \mathbf{y})$ by marginalizing over (averaging over) the posterior distribution, $\int P(Y(s_0) \mid \mathbf{y}, \boldsymbol{\theta})\, P(\boldsymbol{\theta} \mid \mathbf{y})$.

A Bayesian framework is convenient here, driving inference assisted by proper and moderately informative priors on the weakly identified correlation function parameters. For example, for the smoothness parameter in the Matérn covariance, ν, we can follow Stein (1999) in assuming that the data cannot distinguish $\nu = 2$ and $\nu > 2$, which suggests placing a Unif$(0, 2)$ prior on ν. Usually a MCMC algorithm is required to obtain the joint posterior distribution of the parameters, but again there are different strategies to opt for. For example, we may work with the marginalized likelihood as above, $\mathbf{y} \mid \boldsymbol{\theta} \sim N(X\boldsymbol{\beta}, \sigma^2 H(\phi) + \tau^2 I)$, or we may add a hierarchy with spatial random effects, $\mathbf{w} = (w(s_1), \ldots, w(s_n))$:

$$\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{w} \sim N(X\boldsymbol{\beta} + \mathbf{w}, \tau^2 I)$$

$$\mathbf{w} \sim N(\mathbf{0}, \sigma^2 R(\phi)).$$

In either framework, a Gibbs sampler may be designed, with embedded Metropolis or slice-sampling steps, to obtain the marginal posterior distribution (see, e.g., Banerjee *et al.*, 2004). Much more complex hierarchical models have been discussed extensively in the spatial literature but, irrespective of their complexity, they mostly fit into the template we outlined above.

When we want to capture spatial and temporal associations, modeling is accomplished by envisioning a spatial process evolving through time. The literature in spatiotemporal models is quite rich (see, e.g., Cressie, 1993; Banerjee *et al.*, 2004, and the references therein). Essentially, modeling proceeds from a spatiotemporal process $w(s, t)$ in the above context, where s denotes the location, and t denotes time. Of course, appropriate assumptions on the covariance function associated with $w(s, t)$ have to be made. A popular covariance specification for spatiotemporal models is separability, which models spatiotemporal correlation functions as a product of a purely spatial and a purely temporal covariance function. These and other more general specifications may be found in Banerjee and Johnson (2006).

17.8.3. Illustration

Interest lies in predicting the relative density of eastern hemlock across the Bartlett Experimental Forest. Basal area per hectare[1] of all tree species was estimated at each of 438 forest inventory plots distributed across the domain of interest. The response variable is the fraction of estimated eastern hemlock basal area per hectare. Covariates include elevation and six spring and fall Tasseled Cap spectral components that were derived from Landsat satellite images (Kauth and Thomas, 1976).

A spatial regression model (as in equation (17.5)) was fitted to the data. We employed flat priors for the regression estimates $\boldsymbol{\beta}$ and, based on estimates from initial descriptive analyses including variograms (see, e.g., Banerjee *et al.*, 2004), we used inverted-gamma $IG(2, 0.01)$ for both the spatial variance σ^2 and the measurement error variance τ^2. The maximum distance between inventory plots is 4834.81 meters, so a uniform prior on ϕ was set so that the

effective range was less than 3000 meters. Using these priors an MCMC algorithm was devised to obtain posterior samples. Gibbs updates were used for the regression parameters $\boldsymbol{\beta}$ while Metropolis updates were employed for spatial variance components (σ^2, τ^2) and the spatial range parameter ϕ.

The CODA package in R (`www.r-project.org`) was used to diagnose convergence by monitoring mixing, Gelman–Rubin diagnostics, autocorrelations, and cross-correlations. Analysis was based on three chains of 11,000 samples each. The first 1,000 samples were discarded from each chain as a part of burn-in. Subsequent parameter estimation and analysis used the remaining 30,000 (10,000 × 3) samples.

Table 17.1 presents the 95% central credible intervals for the parameter estimates based upon the posterior samples. All six covariates are significant and perhaps explain some of the spatial variation in the data, as is indicated by the spatial variance σ^2 being smaller than the measurement error variance τ^2. The spatial range is calculated as the distance beyond which the correlation function drops below 0.05; for the exponential correlation function this is approximately $3/\phi$. Finally Figure 17.1 displays an image plot of the estimated response surface overlaid with contours of the estimated spatial random effects (the $w(s)$s). The random effects serve to offset the spatially varying density of the response surface.

17.9. BAYESIAN MODELS FOR DISEASE MAPPING

In previous sections we have alluded to a simple Poisson model for disease counts. In fact, this is the basic model often assumed for small area counts of disease (in tracts, zip codes, counties, etc.). We consider two data resolutions here. First we consider case event data where, within a suitable study region (W), realization of cases arises. The locations of cases are usually residential addresses. These form a spatial point process. Often data is not available at this level of spatial resolution and aggregation to larger spatial units occurs. Aggregated counts of disease are often more readily available (e.g., from

Table 17.1 Parameter estimates for the model covariates elevation and spring and fall Tasseled Cap spectral components. Lower table provides parameter estimates for error terms σ^2 and τ^2, spatial range ϕ, and associated effective range

Parameter	*Estimate: 50% (2.5%, 97.5%)*
Intercept	−0.262 (−0.954, 0.387)
ELEV	−0.002 (−0.002, −0.001)
SPR-TC1	0.007 (0.001, 0.013)
SPR-TC2	−0.007 (−0.011, −0.003)
SPR-TC3	0.011 (0.006, 0.015)
FALL-TC1	−0.007 (−0.011, −0.003)
FALL-TC2	0.008 (0.004, 0.011)
FALL-TC3	−0.004 (−0.008, −0.001)
σ^2	0.009 (0.005, 0.016)
τ^2	0.014 (0.012, 0.018)
ϕ	0.002546 (0.001325, 0.005099)
Effective range (meters)	1178.448 (588.301, 2264.629)

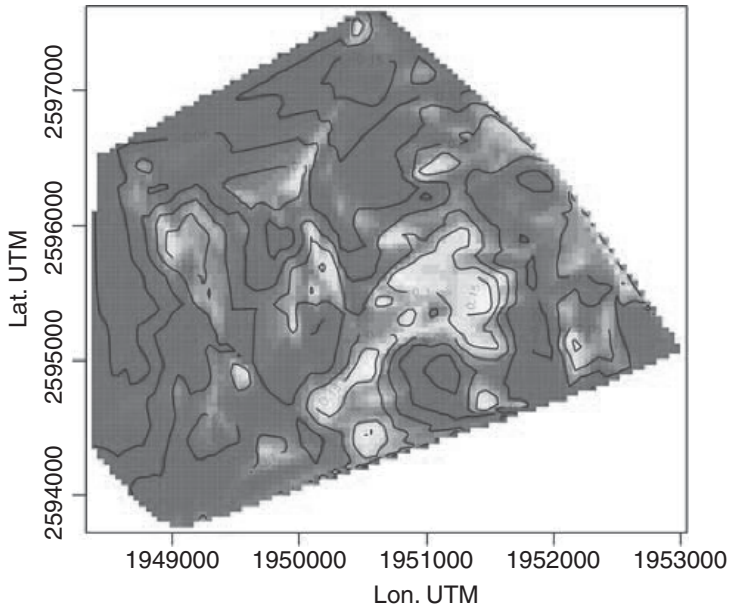


Figure 17.1 Contour lines of estimated spatial random effects overlayed on an image plot of estimated relative density of eastern hemlock. Note, the random effects serve to offset the spatially varying density of eastern hemlock.

official government sources). Hence, the second common data type is disease count data within small areas. These small areas are arbitrary with respect to the disease process (such as census tracts, counties, postcodes) and form a sub-division of the study region. In what follows we will briefly consider case event data, but will concentrate discussion on the more commonly available count data type.

17.9.1. Case event data

Assume we observe within a study region (W), a set of m cases, with residential addresses given as $\{s_i\}, i = 1, \ldots, m$. Figure 17.2 displays an example of such data: larynx cancer incident case addresses for a fixed time period (see Lawson, 2006, Ch 1 for discussion). Here the random variable is the *spatial location*, and so we must employ models that can describe the distribution of locations. Often the natural likelihood model for such data is a heterogeneous Poisson Process (PP). In this model, the distribution of the cases (points) is governed by a first-order intensity function. This function, $\lambda(s)$ say, describes the variation across space of the intensity (density) of cases. This function is the basis for modeling the spatial distribution of cases. we denote this model as:

$$\mathbf{s} \sim \mathbf{PP}(\lambda(\mathbf{s})).$$

The likelihood associated with this model is given by:

$$L = \prod_{i=1}^{m} \lambda(s_i) \exp\{-\int_W \lambda(\mathbf{u})\, d\mathbf{u}\}$$

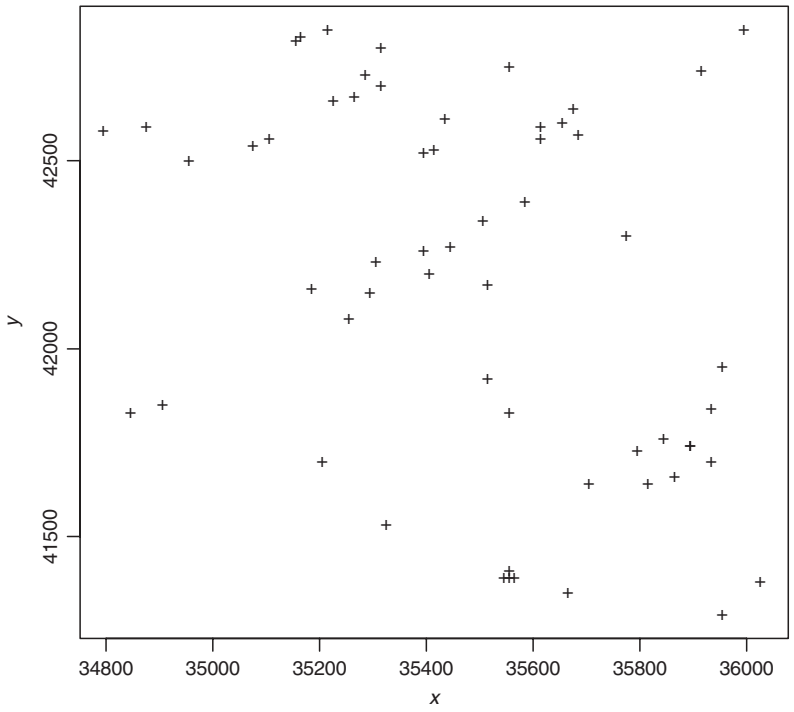


Figure 17.2 Larynx cancer incident case address locations in NW England (1974–1983).

where $\lambda(s_i)$ is the first-order intensity evaluated at the sample locations $\{s_i\}$. This likelihood involves an integral of $\lambda(\mathbf{u})$ over the study region.

In disease mapping studies, usually the variation in disease relates closely to the underlying population that is *at risk* for the disease in question. This is known as the *at risk* background. Hence any definition of the intensity of cases must make allowance for this effect. Any areas where there are lots of *at risk* people are more likely to yield cases and so we must adjust for this effect. Often the intensity is specified with a multiplicative link between these components:

$$\lambda(s) = \lambda_0(s)\lambda_1(s \mid \boldsymbol{\theta}).$$

Here the *at risk* background is represented by $\lambda_0(s)$ while the modeled excess risk of the disease is defined to be $\lambda_1(s \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters. In modeling we usually specify a parametric form for $\lambda_1(s \mid \boldsymbol{\theta})$ and treat $\lambda_0(s)$ as a nuisance effect that must be included. Usually some external data is used to estimate $\lambda_0(s)$ nonparametrically (leading to profile likelihood). This data relates to the local population density. Alternatively, if the spatial distribution of a *control disease* is available (see Lawson and Cressie (2000) for more details), then the problem can be reformulated as a binary logistic regression where $\lambda_0(s)$ drops out of the likelihood. Denote the control disease locations as $\{s_j\}, j = m+1, \ldots, m+n$, and with $N = n + m$, a binary indicator function can be defined:

$$y_i = \begin{cases} 1 & \text{if } i \in 1, \ldots, m \\ 0 & \text{otherwise} \end{cases}$$

$$\forall i, i = 1, \ldots, N$$

and the resulting likelihood is just given by:

$$L(\mathbf{s} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{[\lambda_1(s_i)]^{y_i}}{1 + \lambda_1(s_i)}.$$

By conditioning of the joint set of cases and controls the population effect is removed and does not require estimation.

17.9.2. Parametric forms

Often we can define a suitable model for excess risk within $\lambda_1(s)$. In the case where we want to relate the excess risk to a known location (e.g., a putative source of pollution) then a distance-based definition might be considered. For example:

$$\lambda_1(s) = \rho \exp\{\mathbf{F}(s)\boldsymbol{\alpha} + \gamma d_s\} \qquad (17.8)$$

where ρ is an overall rate parameter, d_s is a distance measured from s to a fixed location (source) and γ is a regression parameter, $\mathbf{F}(s)$ is a design vector with columns representing spatially-varying covariates, and $\boldsymbol{\alpha}$ is a parameter vector. The variables in $\mathbf{F}(s)$ could be site-specific or could be measures on the individual (age, gender, etc.). In addition this definition could be extended to include other effects. For example we could have:

$$\lambda_1(s) = \rho \exp\{\mathbf{F}(s)\boldsymbol{\alpha} + \eta\nu(s) + \gamma d_s\} \qquad (17.9)$$

where $\nu(s)$ is a spatial process, and η is a parameter. This process can be regarded as a random component and can include within its specification spatial correlation between sites. One common assumption concerning $\nu(s)$ is that it is a random field defined to be a spatial Gaussian process.

In the intensity (17.8), all the variables can be estimated using maximum likelihood. However when a Bayesian approach is assumed then all parameters have prior probability distributions and so we would need to consider sampling the posterior distribution given by:

$$P_1(\boldsymbol{\alpha}, \eta, \gamma \mid \mathbf{s}) \propto L(\mathbf{s} \mid \boldsymbol{\alpha}, \eta, \gamma) \cdot P_0(\boldsymbol{\alpha}, \eta, \gamma)$$

where $P_0(\boldsymbol{\alpha}, \eta, \gamma)$ is the joint prior distribution of the parameters. Assuming independent prior distributions for each parameter component, i.e., $P_0(\boldsymbol{\alpha}, \eta, \gamma) = g_{\alpha_1}(\alpha_1) \cdot g_{\alpha_2}(\alpha_2) \cdot g_{\alpha_3}(\alpha_3) \ldots g_{\eta}(\eta) \cdot g_{\gamma}(\gamma)$, this model can be sampled via standard MCMC algorithms. In intensity (17.9), the spatial component $\nu(s)$ would have a spatially correlated prior distribution and so a Bayesian approach would be natural.

17.9.3. Count data

Often only count data is available within a set of small areas. Denote y_i as the count of disease within the ith small area where $i = 1, \ldots, p$. As in the case of case event data we need to allow for the at risk population in our models. This can usually be easily achieved for count data since *expected rates* or *counts* can be obtained or calculated for small areas. For example, age × sex standardized rates for census tracts, postal zones, or zip codes are often available from government sources. Denote these rates as e_i, $i = 1, \ldots, p$. Also, in our model we want to model the *relative risk* of disease via the parameter θ_i, $i = 1, \ldots, p$. The relative risk will be the focus of modeling and it is usually assumed that the $\{e_i\}$ are fixed.

The simplest model for such data is a Poisson log linear model where:

$$y_i \sim Poiss(e_i\theta_i).$$

In addtion the relative risk θ_i is usually modeled with a log link for positivity. A simple example could be:

$$\log \theta_i = \alpha_0,$$

a constant. This model represents constant area-wide risk and often the null hypothesis aasumed by many researchers is that $\alpha_0 = 0$, so that $\theta_i = 1$. This represents the situation where the underlying rate or count generates the risk directly (i.e., $y_i \sim Poiss(e_i)$). This would be applicable if there were no excess risk in the study area. Of course this is seldom reality and it is the alternative hypotheses where θ_i have some spatilal structure that is of interest in modeling.

Some examples of models currently adopted for different applications can be instructive:

Putative health hazard assessment

Usually in these applications some measure of the association between small area counts and a fixed location or locations is to be made. This association could be via distance or directional measures. For example, define the distance from the *i*th small area centroid to the source as d_i and the angle as ψ_i. A log linear model for risk related to a source might be of the form:

$$\log \theta_i = \alpha_0 + \alpha_1 d_i + \alpha_2 \cos(\psi_i - \mu_0) + \alpha_3 \sin(\psi_i - \mu_0) + \Gamma_i.$$

Here, the directional component is summarized by the cosine and sine terms in relation to a mean angle parameter (μ_0), while the distance component is assumed to be log-linearly related to risk. The final term Γ_i is meant to repesent unattributed extra variation in risk. This could include random effect terms, such as:

$$\Gamma_i = u_i + v_i$$

where each term could represent different aspects of the extra variation. For example, u_i is often defined to have a correlated prior distribution (and is called correlated or structured heterogeneity (CH)), whereas v_i is often assumed to represent uncorrelated heterogeneity (UH). The prior distributions assumed for these terms are commonly:

$$v_i \sim N(0, \tau_v)$$

$$(u_i \mid \cdots) \propto \frac{1}{\sqrt{\beta}} \exp\left\{ -\sum_{j \in \partial_i} w_{ij}(u_i - u_j)^2 \right\}$$

where $w_{ij} = 1/2\beta \ \forall i, j$. The neighborhood ∂_i is assumed to be the areas with common boundary with the *i*th area. The second of these prior distributions assumes dependence between neighboring areas. This distribution is termed a conditional autoregressive (CAR) prior distribution. It is an example of a Markov random field. Note that in this definition the parameter β controls the spatial smoothness (or correlation) of the component.

The posterior distribution can be specified as follows:

$$P(\mathbf{u}, \mathbf{v}, \beta, \tau_v, \boldsymbol{\alpha} \mid \mathbf{y}) \propto \mathbf{L}(\mathbf{y} \mid \boldsymbol{\theta}) \times \mathbf{f}_1(\mathbf{u})\mathbf{f}_2(\mathbf{v})\mathbf{f}_3(\boldsymbol{\alpha})f(\beta)f(\tau_v)$$

where $\mathbf{f}_1(\mathbf{u})$ is the CAR prior distribution, $\mathbf{f}_2(\mathbf{v})$ is a zero mean normal distribution,

$\mathbf{f}_3(\boldsymbol{\alpha})$ is the joint prior distribution for the regression parameters, $f(\beta)$ and $f(\tau_v)$ are prior distributions for the remaining parameters. Note that β and τ_v are hyperparameters and they have prior distributions as could any hyperparameters within the other prior distributions ($\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$). The prior distributions for regression parameters are often assumed to be independent and each parameter is often assumed to have a zero mean normal prior distribution.

Disease map reconstruction

Often the main aim of modeling disease incidence is simply to provide a good estimate of disease risk. This can be specified as the relative risk within each region (θ_i). Hence the aim is to provide an accurate estimate of the true underlying risk within the map. Much recent work has been focussed on this area of concern, and many models and approaches have been developed (see, e.g., Banerjee *et al*., 2004, section 5.4; Lawson, 2006, Chapter 8.0, Lawson (2008)). Typically a log linear model with random effects is defined:

$$\log \theta_i = \alpha_0 + \Gamma_i \quad \text{where} \quad \Gamma_i = u_i + v_i.$$

Here the u_i, v_i terms are CH and UH defined as above. This is often called the convolution model and was originally proposed by Besag *et al*. (1991). This model has proved to be very robust against mis-specification of the risk, although it can also over-smooth rates. Lawson *et al*. (2000), Best *et al*. (2005) and Hossain and Lawson (2006) have provided recent simulation-based evaluations of a range of methods in this area.

Ecological analysis

This area of focus arises when the risk within a small area is to be related to a covariate or covariates usually measured at the aggregate level. Often the main issue relates to making individual level inference from aggregate data. Aggregation or averaging induces biases in estimation of parameters for models (see, e.g., Wakefield, 2004). The *modifiable areal unit problem* (MAUP) is an example of an aggregation-related inference problem. Another problem that can arise is the *misaligned data problem (*MIDP*)*. This arises when the spatial resolution of covariates is different from the outcome variable. The classic example of this would be modeling cancer outcomes at zip code level and relating these to groundwater uranium measured at point locations (wells). A fuller discussion of these issues can be found in Banerjee *et al*. (2004). In general the type of model assumed is often of the form:

$$\log \theta_i = x_i^T \beta + z_i^T \xi$$

where x_i^T is a row vector of fixed covariate values for the ith small area and β is a corresponding parameter vector, and z_i^T is a row vector of random effects and ξ a unit vector.

Surveillance

With recent concerns over bioterrorism (Fienberg and Shmueli, 2005; Sosin, 2003; Lawson and Kleinman, 2005), the focus of disease surveillance has become important. Essentially this focus concerns the monitoring of disease incidence with a view to detecting aberrations or unusual incidence events. This often requires the monitoring of large scale databases of health information. In addition, the focus of the monitoring could be a range of effects. There could be a need to find clusters of disease on maps or change points in time series or some mixture of these effects in space–time. Detection of change in multiple time and spatial series is the focus. This is a challenging area that requires

the use of fast computational algorithms and novel spatial-sequential inference. In essence, a range of models found in equations (17.1)–(17.3) above may need to be examined simultaneously in this analysis.

17.9.4. Example

Here we examine briefly an example of relative risk estimation. The example consists of the South Carolina incidence of congenital anomalies deaths by county for 1990. This has also been examined in Chapter 6 of Lawson *et al*. (2003). Figure 17.3 diplays the standardised mortality ratio for this disease for 1990. We are concerned to estimate the true relative risk underlying these county rates. To achieve this we propose a log linear model for the risk in each area. Hence we assume the likelihood:

$$y_i \sim Poiss(e_i\theta_i)$$

and then a log linear model of the form

$$\log\,\theta_i = \alpha_0 + \Gamma_i \quad \text{where} \quad \Gamma_i = u_i + v_i.$$

The two effects have the following prior distributions:

$$u_i \sim CAR(\overline{u}_{\delta_i}, \tau/n_{\delta_i})$$

where δ_i is the neighborhood of the ith area, $\overline{u}_{\delta_i}$ is the mean of u_i in the neighborhood, and n_{δ_i} is the number of neighbors, τ is the variance, and

$$v_i \sim N(0, \kappa)$$

where κ is the variance. Now α_0 is assumed to have a uniform prior distribution on a large range, while the τ and κ are variances and their inverses (precisions: $1/\tau, 1/\kappa$) have gamma prior distributions with fixed parameters (shape: 0.5, scale: 0.0005). There is some debate currently about how informative such hyperprior distributions are (see, e.g., Gelman, 2005). In fact it is always recommended that sensitivity to prior assumptions be examined in any application. The Bayes estimate of the relative risk is the posterior expected value of relative risk for

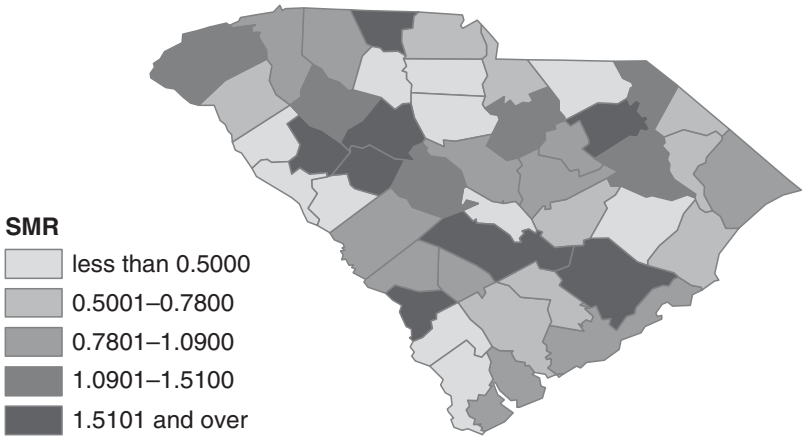


Figure 17.3 Congenital anomalies deaths, standardized mortality ratio, South Carolina, 1990.

each region. This can be obtained from a posterior sample by averaging the converged sample output. The estimates of the relative risk for the congential abnormalities data are displayed in Figure 17.4. The posterior probability of $\theta_i > 1$ over the whole map is shown in Figure 17.5 Note that this quantity can be used to assess whether ther are any areas of 'significant' risk elevation on the map. For more details of this example see Lawson *et al*. (2003: chapter 6).

17.10. SOFTWARE FOR BAYESIAN MODELING

Posterior sampling is the commonest approach to Bayesian inference. There is now a range of software that can peform this task. The best known of these is the free software WinBUGS (downloadable from www.mrc-bsu.cam.ac.uk/bugs/). This package employs both Gibbs sampling and Metropolis–Hastings updating methods for a

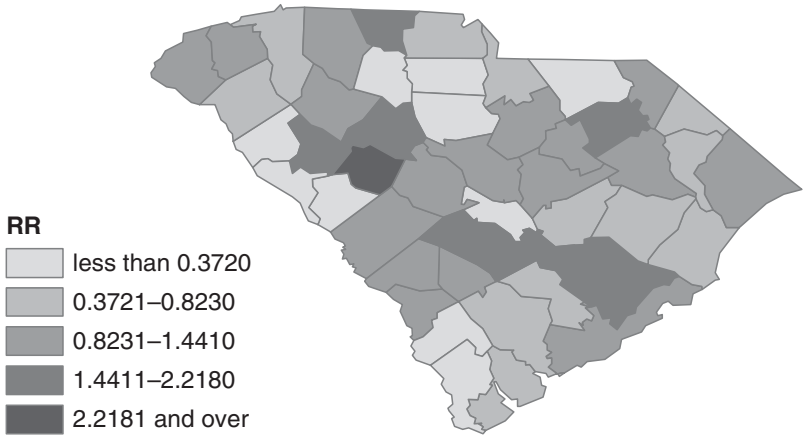


Figure 17.4 Posterior expected relative risk estimates for the congenital abnormalities data for South Carolina, 1990.

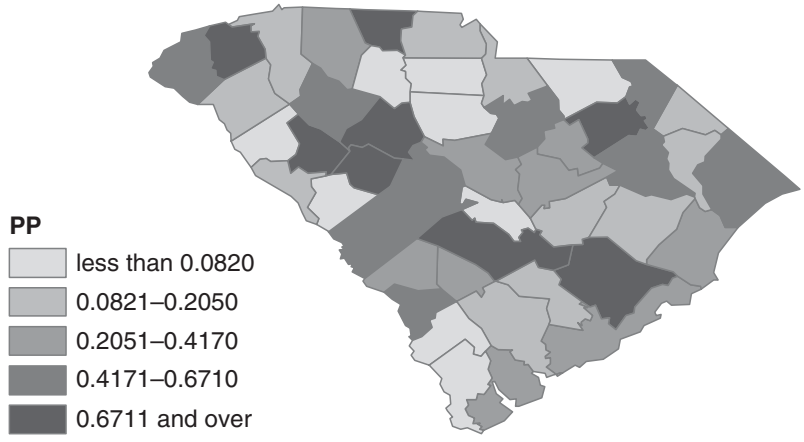


Figure 17.5 Posterior probability of exceedance ($P_r(\theta_i > 1)$) for the South Carolina congenital abnormalities data.

wide range of models. The package also has a wide range of online runnable examples and has a GIS tool called GeoBUGS that allows mapping of small area data and parameter estimates, as well as spatial modeling of various kinds. Bayesian Kriging and both CAR and multivariate CAR models can be fitted using this package. Facilities also exist within R (e.g. packages such as bayesm, geoR, geoRglm, MCMCpack, mCmC, spBayes etc.) and MATLAB (spatial statistics toolbox) to perform MCMC computations for Bayesian spatial models.

ACKNOWLEDGMENTS

Portions of this research were based upon data generated in long-term research studies on the Bartlett Experimental Forest, Bartlett, NH, funded by the U.S. Department of Agriculture, Forest Service, Northeastern Research Station. The authors would especially like to thank Marie-Louise Smith in the USDA Forest Service Northeastern Research Station for sharing a data set and Andrew Finley in the Department of Forest Resources at the University of Minnesota for help with the statistical computations.

NOTE

1 Basal area is the cross-sectional area of a tree at 1.37 meters from the ground. Basal area per hectare is the sum of all the basal area per tree in the hectare.

REFERENCES

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. London: Chapman and Hall/CRC Press.

Banerjee, S. and Johnson, G.A. (2006). Coregionalized Single- and Multi-resolution Spatially-varying Growth Curve Modelling with Applications to Weed Growth. *Biometrics*, 61, 617–625.

Berger, J.O. (1985). *Bayesian Decision Theory*. New York: Springer Verlag.

Besag, J. and Green, P.J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B*, **55**: 25–37.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**: 1–59.

Best, N., Richardson, S. and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, **14**: 35–59.

Carlin, B.P. and Louis, T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. London: Chapman and Hall/CRC Press.

Chen, M., Shao, Q. and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer Verlag.

Chilés and Delfiner (1999). *Geostatistics: Modelling Spatial Uncertainty*, p. 43. New York: Wiley.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*, revised edition. New York: Wiley.

Fienberg, S. and Shmueli, G. (2005). Statistical issues and challenges associated with rapid detection of bio-terrorist attacks. *Statistics in Medicine*, **24**: 513–529.

Gamerman, D. (2000). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. New York: CRC Press.

Gelfand, A. and Ghosh, S. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, **85**: 1–11.

Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**: 1–19.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D. (2004). *Bayesian Data Analysis*. London: Chapman and Hall/CRC Press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. PAMI*, **6**: 721–741.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D. and Kirby, A.J.

(1993). Modelling complexity: Applications of Gibbs sampling in medicine. *Journal of the Royal Statistical Society B*, **55**: 39–52.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**: 97–109. 44

Hossain, M. and Lawson, A.B. (2006). Cluster detection diagnostics for small area health data: with reference to evaluation of local likelihood models. *Statistics in Medicine*, **25**: 771–786.

Kauth, R.J. and Thomas, G.S. (1976). The tasseled cap – a graphic description of the spectral-temporal development of agricultural crops as seen by landsat. In: *Proceedings of the Symposium on Machine Processing of Remotely Sensed Data*, pp. 41–51. West Lafayett: Purdue University.

Lawson, A.B. (2006). *Statistical Methods in Spatial Epidemiology*, 2nd edn. New York: Wiley.

Lawson, A. B. (2008) *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. London: Chapman and Hall/CRC Press.

Lawson, A.B., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P. and Divino, F. (2000). Disease mapping models: an empirical evaluation. *Statistics in Medicine*, **19**: 2217–2242. Special issue: Disease mapping with emphasis on evaluation of methods.

Lawson, A.B., Browne, W.J. and Vidal-Rodiero, C.L. (2003). *Disease Mapping with WinBUGS and MLwiN*. New York: Wiley.

Lawson, A.B. and Cressie, N. (2000). Spatial statistical methods for environmental epidemiology. In: Rao, C.R. and Sen, P.K. (eds), *Handbook of Statistics: Bio-Environmental and Public Health Statistics*, volume 18, pp. 357–396. Amsterdam: Elsevier.

Lawson, A.B. and Kleinman, K. (eds) (2005). *Spatial and Syndromic Surveillance for Public Health*, p. 45. New York: Wiley.

Lee, P. (2005). *Bayesian Statistics*, 4th edn. London: Arnold.

Móller, J. and Waagpetersen, R. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. New York: CRC/Chapman and Hall.

Neal, R.M. (2003). Slice sampling. *Annals of Statistics*, **31**: 1–34.

Ripley, B.D. (1987). *Stochastic Simulation*. New York: Wiley.

Robert, C. (2001). *The Bayesian Choice: A Decision-theoretic Motivation*. New York: Springer Verlag.

Robert, C. and Casella, G. (2005). *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer.

Schabenberger, O. and Gotway, C. (2004). *Statistical Methods For Spatial Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press.

Scheiner, S.M. and Gurevitch, J. (2001). *Design and Analysis of Ecological Experiments*, 2nd edn. London: Oxford University Press.

Sosin, D. (2003). Draft framework for evaluating syndromic surveillance systems. *Journal of Urban Health*, **80**: i8–i13. supplement.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex models. *Journal of the Royal Statistical Society*, **64**: 583–640.

Stein, M. (1999). *Statistical Interpolation of Spatial Data: Some Theory for Kriging*, p. 46. New York: Springer Verlag.

Wakefield, J. (2004). A critique of statistical aspects of ecological studies in spatial epidemiology. *Environmental and Ecological Statistics*, **11**: 31–54.

Waller, L. and Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*. New York: Wiley.

Webster, R. and Oliver, M. (2001). *Geostatistics for Environmental Scientists*. New York: Wiley.

18

Monitoring Changes in Spatial Patterns

Peter A. Rogerson

18.1. INTRODUCTION

The tools of spatial analysis have long been used to study the characteristics of geographic patterns. Central to this effort has been the application of statistical tools to test the null hypothesis of spatial randomness. Interest in the spatial distribution of species within the field of ecology gave rise to some of the earliest approaches, including the nearest neighbor statistic for use with point data (Skellam, 1952; Clark and Evans, 1954) and the quadrat approach for use with counts of events lying within predefined subregions (see Blackman (1935) for an early application).

As the field of spatial analysis has developed, other statistical measures and tests for geographic pattern have become popular. Moran's I (1950) is of special note, owing to its widespread use and popularity. The use of K-functions to assess the nature of point patterns over a range of spatial scales (Ripley, 1976) and kernel density methods to visualize the spatially-varying intensity of variables are now in common use (see Bailey and Gatrell (1995) and Waller and Gotway (2004) for reviews).

While the majority of early approaches relied upon a single, global statistic to evaluate the null hypothesis of spatial randomness, there has been more recent interest in local statistics; these are the location-specific components of global statistics that allow one to test whether spatial association exists in the vicinity of a particular location (see, e.g., Anselin, 1995; Getis and Ord, 1992; Ord and Getis, 1995).

Many of the more recent developments in the statistical analysis of spatial patterns have

taken place within the field of epidemiology, where there is interest in the detection of geographic clusters. Besag and Newell (1991) suggest three categories for these statistical approaches. In addition to the global and local statistics outlined above (referred to by Besag and Newell as general and focused tests, respectively), they note that there is a separate category for *tests for the detection of clustering*. While global tests lead to acceptance or rejection of a specified null hypothesis (perhaps one of spatial randomness, but more realistically, one where the observed spatial distribution of cases is compared with an expected distribution based upon population distribution and possibly other covariates), they do not indicate the size and/or location of geographic clusters. Similarly, local tests are limited in the sense that they evaluate only one location. A test for the detection of clustering may essentially be viewed as a set of local tests (where one or more specifications of potential cluster size are made for many locations within the study area). Scan statistics (Kulldorff and Nagarwalla, 1994), and the maximum of smoothed Gaussian random fields (Rogerson, 2001) fall into this category, where the extreme local statistic is assessed, and the multiple hypothesis testing associated with carrying out many local tests is accounted for.

Like other subfields of spatial analysis, interest in the statistical analysis of spatial patterns and the development of statistical methods for cluster detection has grown rapidly in the last decade. Waller (Chapter 16 in this volume) provides a review of many of these developments and related issues.

Spatial statistical tests of null hypotheses are almost always carried out on a single set of data; the hypothesis is accepted or rejected, and ideally the size and location of significant geographic clustering is revealed. However, there are many situations where repeated tests of this type are required. Imagine the crime analyst, who, each month, receives a new map of burglaries, or the epidemiologist who maps the locations of new cancer cases each year. A market researcher may wish to assess the degree to which customers cluster around a store, and it may be of particular interest to monitor this each month, based upon new sales data. If statistical tests are simply carried out each time a new map is available, the multiplicity of tests will increase the likelihood that a false declaration of significance is made. For instance, if 20 tests are carried out using a Type I error probability of 0.05, we can expect to find on average one false rejection of the null hypothesis among the 20 tests.

In this chapter we describe and review the use of statistical approaches designed for carrying out repeated tests concerned with the evaluation of spatial patterns. The common objective of such repeated tests is the quick detection of geographic change (where most commonly the goal is to find new, emergent clusters as quickly as possible). It can be noted that this objective of prospective, quick detection of temporal change in spatial pattern differs from that of retrospectively finding space–time interaction in a set of data using a single test such as those outlined by say Knox (1964), Mantel (1967), or Raubertas (1988).

The development of methods for the surveillance or monitoring of spatial patterns has received much of its impetus during the last few years from intense interest in surveillance for bioterrorism, and following from that, interest in public health surveillance. The recent reviews of outbreak detection algorithms (Buckeridge *et al.*, 2005) and control charts for public health surveillance (Woodall, 2006) include discussions of spatial considerations in surveillance and summarize the many recent advances in this area. In addition, Chapter 9 of Lawson (2001) and the more recent collection of contributions edited by Kleinman and Lawson (2005) also attest to the growing importance of this field.

The remainder of the chapter is structured as follows: Section 18.2 describes and reviews the use of the methods of statistical process control; these methods have been developed primarily within an industrial context for the quick detection of change in industrial processes and product quality. They are appropriate for monitoring an outcome variable for a single region, and they lie at the core of many approaches to spatial surveillance. The focus is upon cumulative sum methods in particular, due to both their optimality properties and their widespread use both in temporal public health surveillance and in many early attempts at spatial surveillance. Section 18.3 gives a very brief history of the recent development of interest in the methods of spatial surveillance. The intent here is to indicate some of the early approaches and some of the general perspectives taken in various attempts to monitor geographic patterns; no attempt is made to be comprehensive. Section 18.4 describes how these statistical process control methods have been adopted for use with spatial statistics to carry out surveillance for potential changes in geographic patterns.

18.2. STATISTICAL PROCESS CONTROL FOR TEMPORAL SEQUENCES OF OBSERVATIONS

18.2.1. Shewhart charts

The majority of statistical approaches for spatial surveillance have their methodological roots in the field of statistical process control. Industrial processes are often monitored so that various process parameters stay within tolerable limits, and so that manufactured products maintain acceptable quality. Control charts for such purposes were developed by Shewhart in 1924. In a straightforward application of Shewhart charts, sequential observations are plotted on a chart that has both a centerline corresponding to the assumed process mean, and upper and lower control limits, usually corresponding to plus and minus three standard deviations from the mean, respectively. If data come from a standard normal distribution, an observation outside of the control limits of plus or minus three would be observed once every 370 observations on average, since the tail area of a normal distribution lying outside three standard deviations is approximately equal to 1/370. One possible rule for declaring a process to be out-of-control, therefore, could be to do so when an observation is observed outside of the control limits; the average run length (i.e., number of observations) until an alarm is declared, when the process is in control (designated ARL_0), would be 370. Since this procedure would declare alarms for single, outlying observations, various alternative rules are also commonly implemented – for example, some users advocate declaring an out-of-control alarm if there are nine consecutive observations on one side of the mean (see, e.g., Nelson (1984) for this and other suggested rules). For normally distributed observations, the control limits for a Shewhart chart can easily be redefined to be consistent with a desired value of ARL_0. For example, suppose that false alarms were desired only once every 700 observations. The standard normal score associated with a two-tail area of 1/700 (i.e., an area of 1/1,400 in each tail) is found to be 3.19, and so upper and lower control limits would be set at ± 3.19 standard deviations.

18.2.2. Cumulative sum (CUSUM) charts

Although Shewhart charts are straightforward to employ, and they are good at detecting

large changes from the process mean, they are not as sensitive as other methods at detecting smaller and therefore more subtle deviations from the baseline process. The cumulative sum (CUSUM) chart was introduced by Page (1954); the approach consists of maintaining the cumulative sum of deviations between observed and expected values. Cumulative sum methods are covered in detail by Hawkins and Olwell (1998). For the particular example of standardized, independent, normally distributed observations (z_t), the one-sided cumulative sum at time t, S_t, is:

$$S_t = \max(0, S_{t-1} + z_t - k)$$

where k is a parameter chosen to be equal to one-half the size of the deviation that is expected when the process goes out of control. In this example, the expected value of each observation is equal to zero (since observations have been standardized), and it is easy to see that the cusum is, more precisely, the cumulative sum of deviations for observations that exceed their expectation by more than k standard deviations. The parameter k is almost always chosen in this case to be equal to ½; this choice minimizes the time it will take to detect a one standard deviation increase in the mean of the process. An alarm indicating an increase in the underlying mean of the process is declared when the cumulative sum exceeds some predefined threshold, h (i.e., $S_t > h$). The threshold is chosen in conjunction with a desired value of ARL_0; for the case of $k = ½$, Rogerson (2006) provides the following formula:

$$h \approx \frac{\mathrm{ARL}_0 + 4}{\mathrm{ARL}_0 + 2} \ln\left(\frac{\mathrm{ARL}_0}{2} + 1\right) - 1.166. \tag{18.1}$$

For other choices of k in the range $1/\sqrt{\mathrm{ARL}_0} \leq k \leq 1$ one can use the more general

$$h \approx \frac{2k^2\mathrm{ARL}_0 + 2}{2k^2\mathrm{ARL}_0 + 1} \ln\left(\frac{2k^2\mathrm{ARL}_0}{2k} + 1\right) - 1.166. \tag{18.2}$$

The Shewhart chart is a special case of the cusum chart, where k is equal to the Shewhart control limit and $h = 0$.

There is a tradeoff between the rate of false alarms and the ability to detect change when it actually occurs; the higher the value of ARL_0 (and hence the lower the false alarm rate), the greater will be the time until true change is detected (as signified by ARL_1, the average number of observations until an alarm is signaled, once change has occurred). Moustakides (1986) shows, and Frisen and Sonesson (2005) note, that the cusum approach minimizes the maximum expected delay until an alarm is sounded, for a particular changepoint.

Cusums for Poisson data

Regional data to be used for monitoring are often not normally distributed. For example, counts of disease or crime incidents are often taken to have a Poisson distribution. Lucas (1985) gives the Poisson cusum as:

$$S_t = \max(0, S_{t-1} + y_t - k)$$

where y_t is the count at time t. If the expected count is constant and equal to λ_0, the value of k is:

$$k = \frac{\lambda_1 - \lambda_0}{\ln \lambda_1 - \ln \lambda_0}$$

where it is desired to detect an increase in the Poisson parameter from λ_0 to λ_1 as

quickly as possible. Lucas gives tables for the threshold h, which is determined from both k and the analyst's choice of the in-control average run length, ARL_0.

An alternative approach is to attempt to transform the Poisson counts to normality. Rossi *et al.* (1999) find that the following transformation converts the data, approximately, to a standard normal distribution:

$$z_t = \frac{y_t - 3\lambda_0 + 2\sqrt{\lambda_0 y_t}}{2\sqrt{\lambda_0}}.$$

Rogerson and Yamada (2004) give examples however showing that this transformation may be unreliable when $\lambda_0 < 2$. In addition, Hawkins and Olwell (1998) note that detection times are shorter when cusums designed for the distribution are employed, in comparison with cusums based upon transformations to normality.

Situations where the expected count remains constant over time are unusual; disease counts might be expected to vary seasonally, or exhibit other temporal trends. The use of transformations to normality allows such temporally-varying expectations to be easily handled. Alternatively, the Poisson cusum can itself be generalized to handle changing expectations (see Hawkins and Olwell, 1998; Rogerson and Yamada, 2004).

Cusums for exponential data

Quicker detection of increases in the rate of rare events can often be achieved by monitoring the times between events (Wolter, 1987; Gan, 1994). For a random process, the times between events are exponentially distributed:

$$f(x) = \theta \exp(-\theta x)$$

where $1/\theta$ is the mean time between events.

To detect a decrease in the mean time between events (and hence an increase in θ from, say, θ_0 to θ_1), one can use the exponential cusum:

$$S_t = \max(0, S_{t-1} - x_t + k)$$

where x_t is the time between events, and:

$$k = \frac{\theta_1 - \theta_0}{\ln(\theta_1 \theta_0)}.$$

Rogerson (2005) derives the approximate threshold associated with a desired ARL_0 by first transforming the problem into one having an in-control parameter of $\theta = 1$; this is achieved by dividing each observation by θ_0. The normalized out-of-control parameter is then $\tilde{\theta}_1 = \theta_1/\theta_0$. The threshold is then given by:

$$h \approx \frac{(q+2)\ln(q+1)}{(q+1)\ln(\tilde{\theta}_1)} - 1.33$$

where:

$$q = ARL_0 \ln(\tilde{\theta}_1)|1 - k|.$$

18.2.3. Other methods for temporal surveillance

The exponentially weighted moving average (EWMA) chart was introduced by Roberts (1959) and is discussed further by Hunter (1986) and by Lucas and Saccucci (1990); it is based upon the quantities:

$$z_t = (1 - \lambda)z_{t-1} + \lambda x_t$$

where x_t is the observation at time t and λ is a parameter that dictates the importance of dated information. An alarm is signaled at the first time when the value of z_t exceeds a time-varying threshold that over time reaches an asymptotic limit. In the special case of $\lambda = 1$, only current information is used, and the method is identical to the Shewhart chart.

The Shiryaev–Roberts method, based upon contributions from Shiryaev (1963) and Roberts (1966), can be derived as a special case of a likelihood ratio method with a noninformative prior distribution on the time of the changepoint (Frisen and Sonesson, 2005). This approach minimizes the expected time until an alarm following a change.

Many other approaches to temporal surveillance exist; these range from simple calculations of historical limits that are empirically based upon recent data, to sophisticated use of time series analysis – these are reviewed by Farrington and Beale (1998), and more recently by Le Strat (2005).

18.3. SPATIAL SURVEILLANCE

18.3.1. *Brief overview of the development of methods for spatial surveillance*

Like recent developments in spatial cluster detection, many of the recent developments in the monitoring of spatial patterns have occurred within the field of public health. Raubertas (1989) was one of the first to outline how statistical approaches to spatial surveillance could be developed, and he did so in the context of disease surveillance.

Raubertas employed cumulative sum methods to suggest how disease monitoring for a particular region within a study area could be carried out. Monitoring is based upon forming a weighted sum of the number of cases occurring both in the region of interest and in the surrounding regions. The weights define the spatial structure of the alternative, and should be matched as closely as possible with the definition of any presumed cluster. The weights for example might decline as the distance from the region of interest increases. For each time period, the weighted sum of observations is compared with expectations, and deviations are cumulated; if these deviations exceed a pre-specified threshold, an alarm signaling a possible increase in disease in the vicinity of the region of interest is sounded. Raubertas notes some of the complications that arise when one wishes to monitor several regions simultaneously, since there will be correlation in the monitoring statistics obtained for regions that are close to one another (since they will have shared neighborhoods).

Statistical process control approaches to spatial surveillance may be categorized into those that maintain separate, local charts for each region (where, like Raubertas', the regional chart may possibly include information from a defined neighborhood around the region), and those that monitor a single, global spatial statistic.

As an example of the latter category, Rogerson (1997) also uses cumulative sum methods to monitor temporal changes in a global spatial statistic (specifically, Tango's 1995 statistic). Each time a new case is observed, Tango's statistic is updated and the resulting statistic is then compared with the expectation of the statistic (conditional upon the previous value of the statistic, before the new case was observed) under the null hypothesis of no raised incidence in any subregion. An alarm is sounded, indicating a significant change in the global statistic, if deviations between observed and expected statistics cumulate sufficiently.

Kulldorff (2001) has extended his spatial scan statistic to the case of prospective disease surveillance, by considering the

likelihood of the observed number of events in space-time cylinders (where the vertical axis represents time, and the horizontal plane represents a region and its surrounding neighborhood), under the null hypothesis. The spatial scan statistic (Kulldorff and Nagarwalla 1994) is based upon the likelihood ratio associated with the number of events inside and outside of a circular scanning window. The numerator of the ratio is associated with the hypothesis that the rates inside and outside of the rate are different, and the denominator of the ratio is associated with the null hypothesis that the rates inside and outside of the window are the same. Likelihood ratios are found using circular scanning windows of various sizes, and the window moves, to scan over space. The most unusual window under the null hypothesis is the one displaying the maximum likelihood ratio. This maximum observed ratio is compared with ratios that are simulated by assuming the null hypothesis to be true; if for example the maximum observed ratio is greater than 95% of the simulated ratios, the cluster is said to be significant using $\alpha = 0.05$.

For disease surveillance, the circular scanning windows become cylinders with time on the vertical axis, where the top of the cylinder represents the most recent time period. To find space-time clusters as the cylinders grow vertically with the progression of time, the maximum likelihood ratio concept is simply generalized. At each time period, the likelihood of the most interesting cylinder (i.e., the one with the highest likelihood ratio) is compared with the likelihood of the most interesting cylinder generated from many simulations of the null hypothesis. The popularity of the method has been aided by freely available software (SatScan), available at www.satscan.org.

Kleinman *et al.* (2004) model the count of cases in a small region using covariates in a generalized linear mixed model (Breslow and Clayton, 1993) for an historical period. In particular, they use a logistic equation to model the probability that a particular individual is a case. Next, they use the coefficient estimates to derive the expected probability that an individual becomes a case during the next time period. Statistical significance is achieved if the observed count of cases is unlikely to have occurred using a binomial distribution based upon the number of individuals and the predicted probability resulting from the model.

Other approaches to spatial surveillance include distance-based methods (see, e.g., Forsberg *et al.*, 2005), and perspectives that adopt more of a model-based than a statistical hypothesis testing perspective (Lawson, 2005).

18.3.2. Spatial issues in spatial surveillance

One way to monitor variables for a set of regional subunits is to maintain a separate cusum chart for each subunit. An immediate issue that arises in the context of monitoring across a set of regional subunits is how to properly account for the multiple testing across spatial units. If cusum control charts are kept for each region, the average run length between false alarms will be less than that implied by the threshold derived for each chart (which is based upon the desired ARL). Thus if thresholds for each chart are chosen using a desired ARL_0 of 100, the mean time until the first alarm on at least one of the charts will be less than 100. More precisely, the average run length between false alarms for a set of m charts (one for each region), ARL_0^*, will be

$$\mathrm{ARL}_0^* = \frac{1}{1 - (1 - 1/\mathrm{ARL}_0)^m}.$$

This is based upon the fact that the time between false alarms has an exponential distribution (Page, 1954), and hence the probability that any single observation leads to a false alarm is $1/\text{ARL}_0$. Alternatively stated, the ARL to use on each chart is given by:

$$\text{ARL}_0 = \left[1 - \left(1 - \frac{1}{\text{ARL}_0^*}\right)^{1/m}\right]^{-1} \quad (18.3)$$

where, again, ARL_0^* is the desired time between alarm investigations. A computationally simpler way to account for the simultaneous monitoring of the m charts is to use a Bonferroni-type adjustment; instead of using Equation (18.3) to determine the threshold for each chart, the quantity:

$$\text{ARL}_0 = m\,\text{ARL}_0^* \quad (18.4)$$

is used. Thus if there are $m = 10$ regional units and a desired time between false alarms of $\text{ARL}_0^* = 100$, the threshold for each chart is found using $\text{ARL}_0 = 10\,(100) = 1000$, together with equation (18.1) or (18.2).

This type of adjustment is appropriate and will yield the desired ARL when (a) no spatial autocorrelation in the regional variables exists, (b) when all regions are in control, and (c) there is a desire to monitor individual regions, and not neighborhoods around regions. However, Equations (18.3) and (18.4) will often lead to thresholds that are too conservative (i.e., thresholds that are too high). One reason for this is that not all m regions may be 'in-control'; we only require a threshold and false alarm rate that have been adjusted for the number of in-control regions (which is unknown, but is less than or equal to m). When a region goes out of control, other (e.g., surrounding regions) may simultaneously go out of control. This suggests that the adjustments for multiple testing may be too severe, and recent developments in the area of multiple testing can be used to lower the thresholds (for a review, see Castro and Singer, 2006). A second reason that equations (18.3) and (18.4) can be conservative is that they assume that the m regional charts are independent. More commonly, regional charts may exhibit spatial dependence; a cusum chart for one region may look a lot like a chart for a nearby region. Finally, if emergent clusters might exceed the size of regional subunits, this will provide a rationale for monitoring local statistics for neighborhoods around regions.

Maintaining separate charts for each region is a *directional* scheme; the approach will work very well when the actual change occurs in one of the regions (and not, for example, combinations of regions), but can lose considerable power in detecting change quickly when changes in other directions occur. If, for example, an increase occurs in a neighborhood containing several regions (corresponding to several charts), this approach will not be as effective and can yield longer times to detection than other methods.

In the next section, we examine some alternative approaches to multiplicity adjustment.

Monitoring a single local statistic

Suppose that there is no spatial autocorrelation in the regional values being monitored, and that we suspect that when change occurs, it will occur in the form of increases in a subset of regions comprising a neighborhood. There are at least two ways forward if our objective is to detect this increase quickly:

1 Keep a single chart for the variable consisting of a weighted sum of the regional values (similar to the suggestion of Raubertas).

2 Use the approach of Healy (1987), which is optimal for quick detection of change in a single, hypothesized direction.

While these approaches should give identical results under the conditions specified, Healy's approach is more general, since it can also handle the situation where the underlying variables are correlated. Specifically, when the variance–covariance matrix associated with the regional values is designated Σ, the following cumulative sum based on vectors of regional observations (x_t) is optimal for detecting a change in mean from μ_G to μ_B, where these latter quantities are vectors of regional values for the good, in-control, and bad, out-of-control means, respectively:

$$S_t = \max(0, S_{t-1} + a'(x_t - \mu_G) - 0.5D)$$

where:

$$a' = \frac{(\mu_B - \mu_G)'\Sigma^{-1}}{\{(\mu_B - \mu_G)'\Sigma^{-1}(\mu_B - \mu_G)\}^{1/2}}$$

and:

$$D = \sqrt{(\mu_B - \mu_G)'\Sigma^{-1}(\mu_B - \mu_G)}.$$

Monitoring many local statistics simultaneously

ow suppose that we wish to carry out surveillance of several such local statistics simultaneously. We could either keep a Raubertas-type chart for each local statistic, or, more generally (since it is possible to account for underlying spatial autocorrelation in the regional values), keep a Healy-type chart for each region. Consider first the special case where $\Sigma = I$; the Healy and Raubertas charts will be identical. An important issue is the adjustment for multiplicity; using individual thresholds for each chart based upon *m*ARL would be too conservative, since the charts will be correlated (nearby local statistics will be similar, since they use shared regional values). On the other hand, thresholds based on ARL alone would be too liberal, unless the charts for all local statistics were identical. It is of interest to find the number of effectively independent charts (say, e); in that case each individual threshold could then be based upon e(ARL).

Let the regional variables be denoted by $\{y_i\}$ and the local statistic to be monitored by $\{z_i\}$. Rogerson (2005) suggests that a Gaussian kernel be used to define the neighborhood weights:

$$z_i = \sum_j w'_{ij}\, y_j$$

$$w'_{ij} = \frac{w_{ij}}{\sqrt{\sum_j w_{ij}^2}}$$

$$w_{ij} = \frac{1}{\sqrt{\pi}\sigma} \exp\left(\frac{-d_{ij}^2}{2\sigma^2(A/m)}\right)$$

where A is the size of the study area, and where σ is the width of the Gaussian kernel, expressed in terms of multiples of the square root of the average regional area. Then one possibility is to use the following for an estimate of e:

$$e = \frac{m}{1 + 0.81\sigma^2}.$$

This is based upon results reported in Rogerson (2001), who modified the work of Worsley (1996) on the use of Gaussian random fields to find the probability that

specified thresholds were exceeded anywhere in the study area by at least one local statistic.

Although this idea gives results that are similar to those found through Monte Carlo simulation, the adjustment is based upon the (static) correlation between regional local statistics observed at a single point in time. In practice, the cusum charts being maintained for each regional local statistic will have correlations that are not necessarily the same as this static correlation. Any adjustments to chart thresholds should ideally be based upon the probabilities of charts jointly signaling. Additional approaches to monitoring data from multiregional systems include methods designed for multivariate surveillance (Rogerson and Yamada, 2004) and monitoring regional maxima (Rogerson, 2005).

18.4. SUMMARY

The prospective surveillance of geographic patterns, based upon incoming streams of spatial data, is a field that has grown rapidly in the last decade. This growth has been motivated largely through interest in public health surveillance. There are also many potential applications in other areas, including applications to crime analysis (where there is interest in emerging areas of criminal activity), and in marketing, where the spatial pattern of customers in a competitive retailing environment could be monitored.

This chapter has only touched upon some of the major approaches and issues. The reader interested in investigating the topic further may find the edited collection of Kleinman and Lawson (2005), and the software *GeoSurveillance* (available at wings.buffalo.edu/∼rogerson) of interest.

ACKNOWLEDGMENTS

The support of Grant 1R01 ES0981–01 from the National Institutes of Health and National Cancer Institute Grant R01 CA92693–01 is gratefully acknowledged.

REFERENCES

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis,* **27**: 93–115.

Bailey, A. and Gatrell, A. (1995). *Interactive Spatial Data Analysis*. Essex: Longman (published in the U.S. by Wiley).

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society Series A*, **154**: 143–155.

Blackman, G.E. (1935). A study by statistical methods of the distribution of species in grassland associations. *Annals of Botany*, **49**: 749–777.

Breslow, N. and Clayton, D.G. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**: 9–25.

Buckeridge, D.L., Burkom, H., Campbell, M., Hogan, W.R., and Moore, A.W. (2005). *Journal of Biomedical Informatics*, **38**: 99–113.

Castro, M.C. and Singer, B.H. (2006). Controlling the false discovery rate: a new application to account for multiple and independent tests in local stastistics of spatial association. *Geographical Analysis*, **38**: 180–208.

Clark, P.J. and Evans, F.C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, **35**: 445–453.

Farrington, C.P. and Beale, A.D. 1998. The detection of outbreaks of infectious disease. In: GEOMED '97, International Workshop on Geomedical Systems. Gierl, L., Cliff, A.D., Valleron, A., Farrington, P. and Bull, M. (eds.), pp. 97–117. Stuttgart: B.G. Teubner.

Forsberg, L., Bonetti, M., Jeffery, C., Ozonoff, A. and Pagano, M. (2005). Distance-based methods for spatial and spatio-temporal surveillance. In: Kleinman, K. and Lawson, A.B. (eds), (2005). *Spatial and Syndromic Surveillance*, pp. 31–52. New York: Wiley.

Frisen, M. and Sonesson, C. (2005). Optimal surveillance. In: Kleinman, K. and Lawson, A.B. (eds), (2005). *Spatial and Syndromic Surveillance*,. pp. 31–52. New York: Wiley.

Gan, F.F. (1994). Design of optimal exponential CUSUM control charts. *Journal of Quality Technology*, **26**: 109–124.

Getis, A. and Ord, J. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, **24**: 189–206.

Hawkins, D.M. and Olwell D.H. 1998.*Cumulative Sum Charts and Charting for Quality Improvement*. New York: Springer.

Healy, J.D. (1987). A note on multivariate CUSUM procedures. *Technometrics*, **29**: 409–412.

Hunter, J.S. (1986). The exponentially weighted moving average. *Journal of Quality Technology*, **18**: 203–210.

Kleinman, K. and Lawson, A.B. (eds), (2005). *Spatial and Syndromic Surveillance*. New York: Wiley.

Kleinman, K., Lazarus, R. and Platt, R. (2004). A generalized linear mixed models approach for detecting incident clusters of disease: biological terrorism and other surveillance. *American Journal of. Epidemiology*, **156**: 217–224.

Knox, G. (1964). The detection of space-time interactions. *Applied Statistics*, **13**: 25–29.

Kulldorff, M. and Nagarwalla, N. (1994). Spatial disease clusters: detection and inference. *Statistics in Medicine*, **14**: 799–810.

Kulldorff, M. (2001). Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society Series A*, **164**: 61–72.

Lawson, A. 2001. *Statistical methods in spatial epidemiology*. New York: Wiley.

Lawson, A.B. (2005). Advanced modeling for surveillance: clustering of relative risk changes. In: Kleinman, K. and Lawson, A.B. (eds), (2005). *Spatial and Syndromic Surveillance*, pp. 31–52. New York: Wiley.

Le Strat, Y. (2005). Overview of temporal surveillance. In: Kleinman, K. and Lawson, A.B. (eds), *Spatial and Syndromic Surveillance*, pp. 13–29. New York: Wiley.

Lucas, J.M. and Saccucci, M.S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, **32**: 1–12.

Lucas, J. M. (1985). Counted data cusums. *Technometrics*, 27, 129–144.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**: 209–220.

Moran, P.A.P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, **37**: 17–23.

Moustakides, G.V. (1986). Optimal stopping-times for detecting changes in distributions. *Annals of Statistics*, **14**: 1379–1387.

Nelson, L.S. (1984). The Shewhart control chart: tests for special causes. *Journal of Quality Technology*, **16**: 237–239.

Ord, J. and Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, **27**: 286–306.

Page, E.S. (1954). Continuous inspection schemes. *Biometrika*, **41**: 100–115.

Raubertas, R.F. (1988). Spatial and temporal analysis of disease occurrence for detection of clustering. *Biometrics*, **44**: 1121–1129.

Raubertas, R.F. (1989). An analysis of disease surveillance data that uses the geographic locations of the reporting units. *Statistics in Medicine*, **8**: 267–271.

Ripley, B.D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, **13**: 255–266.

Roberts, S.W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, **1**: 239–250.

Roberts, S.W. (1966). A comparison of some control chart procedures. *Technometrics*, **8**: 411–430.

Rogerson, P. (1997). Surveillance methods for monitoring the development of spatial patterns. *Statistics in Medicine*, **16**: 2081–2093.

Rogerson, P. (2001). A statistical method for the detection of geographic clustering. *Geographical Analysis*, **33**: 215–227.

Rogerson, P. (2005). Spatial surveillance and cumulative sum methods. In: Kleinman, K. and Lawson, A. (eds), *Spatial and Syndromic Surveillance for Public Health*, pp. 95–114. New York: Wiley.

Rogerson, P. (2006). Formulas for the design of CUSUM quality control charts. *Communications in Statistics – Theory and Methods*, **35**: 373–383.

Rogerson, P. and Yamada, I. (2004). Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality Weekly Report*, **53** (Supplement): 79–85.

Rogerson, P. and Yamada, I. (2004). Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Statistics in Medicine*, **23**: 2195–2214.

Rossi, G., Lampugnani, L., and Marchi, M. (1999). An approximate CUSUM procedure for surveillance of health events. *Statistics in Medicine*, 18: 2111–2122.

Shiryaev, A.N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and its Applications*, **8**: 22–46.

Skellam, J.G. (1952). Studies in statistical ecology. I. Spatial pattern. *Biometrika*, **39**: 346–362.

Tango, T. (1995). A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine*, **7**: 649–660.

Waller, L. (2006). Detection of clustering in spatial data. *Handbook of Spatial Analysis*. London: Sage Publications.

Waller, L. and Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*. New York: Wiley.

Wolter, C. (1987). Monitoring intervals between rare events: a cumulative score procedure compared with Rina Chen's sets technique. *Methods of Information in Medicine*, **26**: 215–219.

Woodall, W.H. (2006). The use of control charts in health-care and public health.

Worsley, K.J. (1996). The geometry of random images. *Chance*, **9** (1): 27–40.

19

Case-Control Clustering for Mobile Populations

Geoffrey M. Jacquez and Jaymie R. Meliker

> The effect of [human] mobility could be a time–space lag between causes and effects that makes conventional mapping spurious.
>
> A. Shaerstrom (2003)

19.1. INTRODUCTION

Traditionally, geographic clustering techniques have concerned themselves with static spatial distributions in which human mobility is ignored. For example, within the case-control framework, place-of-residence at time of diagnosis or death is often analyzed even though there may be a substantial space time lag or latency between timing of causative exposures and disease diagnosis. The few techniques currently available for accounting for human mobility when assessing case-clustering often do not adequately account for known risk factors (e.g., smoking), covariates (e.g., age, gender, race, education, etc.) and the space–time lag between exposure and disease. This chapter is based closely on two previous papers published by our research group (Jacquez *et al*., 2005, 2006). It provides background on human mobility and its implications in disease clustering, and then offers an approach for analyzing case-control data for mobile individuals that addresses latency and incorporates covariates and other risk factors in the analysis. Called Q-statistics, this approach is used for analyzing clustering in case-control data for mobile individuals. An example analysis of bladder cancer in southeastern Michigan is presented within an inductive framework in

which the plausible explanatory hypotheses are first enumerated and then systematically evaluated. We demonstrate that clustering of residential histories of bladder cancer cases is only partially explained by smoking, age, gender, race, and education. We also identify clusters of unexplained risk (focused clusters) surrounding the business address histories of 22 industries whose reported emissions and/or business processes release known or suspected bladder cancer carcinogens. The methods developed and demonstrated in this chapter provide a systematic approach for evaluating increasingly realistic alternative hypotheses regarding the identification and explanation of clusters in case-control data.

19.1.1. A historical perspective on human mobility

Recent generations have seen an exponential increase in human mobility (Cliff and Haggett, 2003) and a global shift in the population distribution such that cities and developing countries are growing the fastest. Geographical space has collapsed, and travel times have fallen exponentially from the 1800s to the present (Davies, 1964). Improved transport and population growth have contributed to changing travel patterns, as exemplified by Figure 19.1 which illustrates the increasing size and complexity of travel networks over four male generations of the same family

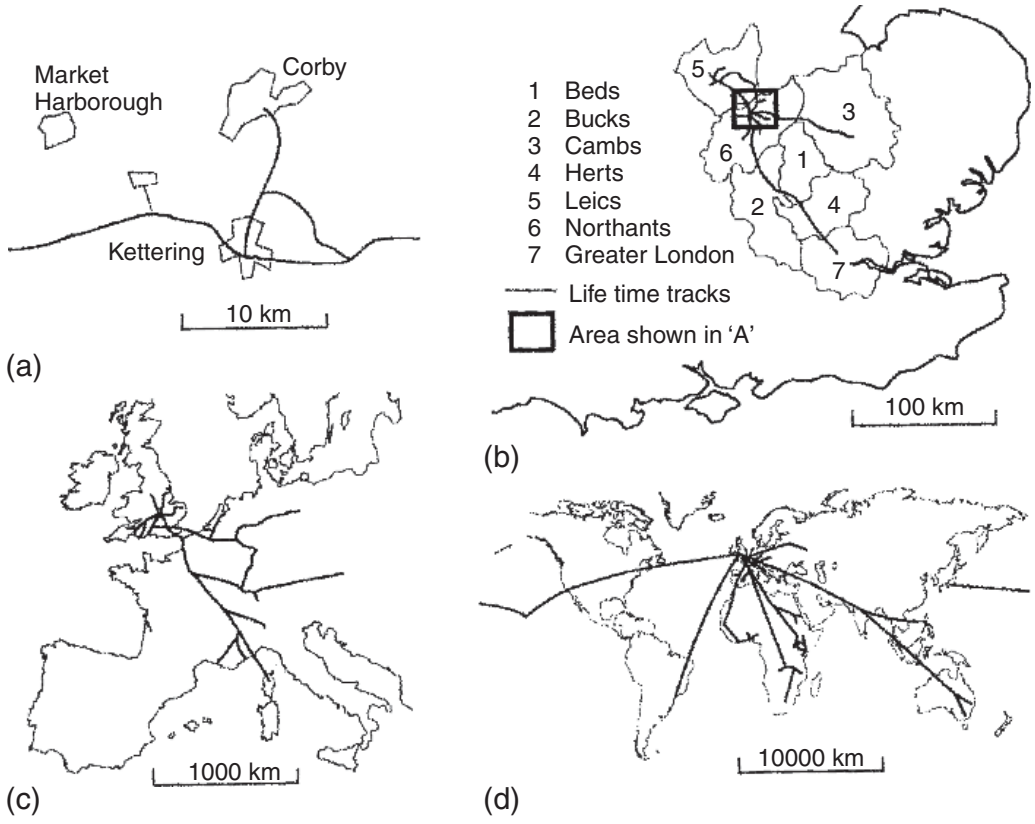


Figure 19.1 Exponential increase in lifetime distances traveled over generations of males from great grandfather (A), grandfather (B), father (C), and son (D). From Bradley (1988), with kind permission of Dr. Bradley and Springer Science and Business Media.

(Bradley, 1988). The lifetime travel-track of the great-grandfather remained within 40 km of a village in Northamptonshire, whereas the grandfather ranged throughout southern England as far as 400 km. The father traveled throughout Europe to a scale of 4000 km and the son was a global traveler, reaching a scale of 40,000 km. Although this is only one illustrative example, it demonstrates what is commonly accepted, that people traveled much greater distances at the turn of the 21st century, on average, than they did at the turn of the 20th century.

In addition to travel mobility, other aspects of human mobility include short-term daily mobility (e.g., commuting to work and running errands) and long-term mobility (e.g., housing mobility and choice of location) (Scheiner and Kaspar, 2003). U.S. population-based surveys estimate that adults spend 87% of their day indoors, 69% in their place of residence, and 6% in a vehicle in transit (Collia *et al*., 2003; Klepeis *et al*., 2001; Reuscher *et al*., 2002). Residential mobility histories compared across several countries in the early 1980s found nearly 13 moves per person over a lifetime in New Zealand, 11 in the U.S., 7 in Great Britain, 6½ in Japan, 5 in Belgium, and 4 in Ireland (Long, 1992). Individuals in their early 20s in New Zealand will, on average, have experienced as many moves as a resident of Ireland over a lifetime. Approximately 50–70% of the moves occur within localities (e.g., counties), 20–35% between localities, 10–15% between regions (e.g., states or provinces), and 0–10% between countries (Long, 1992). While the median distance moved was just 3 km in Great Britain and 10 km in the US, 17–20% of the moves were between regions or between countries, demonstrating considerable mobility for a large segment of the population. The challenge thus is to incorporate residential and other forms of human mobility into environmental health investigations.

19.1.2. Background on residential mobility in environmental health studies

In recent years residential mobility has increasingly been incorporated in exposure assessment. Exposure reconstruction often involves assessment of proximity of individual place-of-residence to environmental hazards such as super-fund sites, incinerators, and hazardous waste sites. In these instances Geographic Information Systems (GIS) have been used to reconstruct individual-level exposures to environmental contaminants (Beyea and Hatch, 1999; Nuckols *et al*., 2004; Ward *et al*., 2000). Examples include assessments of proximity of individuals to landfills (O'Leary *et al*., 2004), hazardous waste sites (Elliott *et al*., 2001; McNamee and Dolk, 2001), and farms for assessing exposures to pesticide application (Reynolds *et al*., 2005). Perhaps because of the emphasis on the individual, exposure reconstruction has concerned itself both with human mobility as well as with temporally dynamic environmental contaminants for which concentrations may change through time. Residential histories and changes through time in the concentrations of environmental contaminants have been addressed in studies of air pollutants (Bellander *et al*., 2001; Bonner *et al*., 2005; Nyberg *et al*., 2000), drinking water contaminants (Swartz *et al*., 2003), pesticides (Aschengrau *et al*., 1996; Brody *et al*., 2002) and herbicides (Stellman *et al*., 2003).

19.1.3. Unrealistic assumptions of disease clustering

Only recently has the role of human mobility and temporally varying exposures been addressed within the context of disease clustering. That risk of disease may vary from one geographic sub-population to another,

and that this risk is time-dependent, is a fact for almost all human diseases, including infectious as well as chronic diseases such as cancer. Goodchild (2000) referred to the failure to appropriately represent the time dimension as a 'static world-view'. To date, many disease clustering methods have been based on a static world-view in which individuals are considered immobile, migration between populations does not occur, and in which background disease risks under the null hypothesis are assumed to be time-invariant and uniform through geographic space. As a result, many of the applications in the published literature suffer from violations of fundamental assumptions that are inherently unrealistic (Jacquez, 2004).

19.2. CONSEQUENCES OF THE STATIC WORLD VIEW IN DISEASE CLUSTERING

When analyzing chronic diseases such as cancer, causative exposures may occur over a long time period, and the disease may be manifested only after a lengthy latency period. During this latency period individuals may move from one place of residence to another. This can make it difficult to detect clustering of cases in relation to the spatial distribution of their causative exposures. Yet the static spatial point distribution is the point of departure for many clustering approaches, including Turnbull's test (Turnbull *et al*., 1990), and tests suggested by Cuzick and Edwards (1990), Besag and Newell (1991), the Bernoulli form of the scan test (Kulldorff and Nagarwalla, 1995), Tango (1995), and a host of others. Especially for chronic diseases with long latencies, human mobility must be accounted for.

Hagerstand (1970) developed conceptual models of the space–time paths formed as individuals move throughout their days and lives. In the context of human health studies these have been called 'geospatial lifelines', and their mathematical representation, properties, and means of analysis have become important research topics. Sinha and Mark (2005) employed a Minkowski metric to quantify the dissimilarity between the geospatial lifelines of cases and controls, and suggested that their technique could be used to evaluate differences in exposure histories between the case and control populations. The Minkowski metric provides a global measure of dissimilarity between cases and controls, but does not identify where or when these dissimilarities occur. Using k-function analysis, Han *et al*. (2004) evaluated clustering of breast cancer in two New York state counties and detected significant spatial clustering at the global level. Their approach incorporated knowledge of residential locations of both cases and controls at biologically relevant ages in a woman's life, namely at birth, menarche, and at woman's first birth. The k-function was applied to the spatial pattern described by place of residence at specific time slices in the participants' lives. Sabel and colleagues (Sabel *et al*., 2000, 2003) used residential histories to analyze clustering of cases of motor neurone disease in Finland. They calculated risk surfaces using kernel functions that were weighted by duration at specific locations of residence. This approach thus used the residential history information more fully, but ignored the temporal ordering of place of residence.

Jacquez and colleagues (2005) developed global, local and focused versions of so-called Q-statistics that evaluate clustering in residential histories using case-control data. Their approach is based on a space–time representation that is consistent with Hagerstrand's model of space–time paths, and evaluates local, global, and focused clustering of the residential histories of the

cases relative to the residential histories of the controls. One of the benefits of the different versions of the Q-statistics is their ability to quantify what is happening at the local, spatial, and temporal scales that are of relevance to individuals, while also providing global statistics for evaluating aggregations of cases. But their approach did not incorporate explicit models of disease latency, nor did it account for those times in a person's life when they might be most susceptible to specific exposures.

19.3. A HISTORICAL PERSPECTIVE ON LATENCY MODELS

It seems a truism to observe that people are mobile, the environment varies through time, and that populations grow and their composition changes, thereby complicating the adjustment for covariates. We therefore need to understand the contributions to individual exposure that transpire at home, at work, and while commuting. Substantial disease latencies may need to be accounted for, and an individual's susceptibility to disease and to environmental insults may vary with age. Metabolic responses may be non-linear and synergistic, and observed impacts of current exposures may be mediated by past exposures. Enzymes involved in metabolism may be inducible, such as the example of alcohol dehydrogenase and alcohol metabolism. In addition, exposures are temporally dynamic, may be episodic or cyclic, and can occur on time scales including days, weeks, years, decades, and potentially over the entire life-course. For example, in summer, air pollutants may vary over the course of day; while concentrations of naturally-occurring metals in groundwater may be relatively static over months and even years. And certain carcinogens and biologically active compounds are of anthropogenic origin (e.g., PCBs) and were not present in the environment in prior generations.

As noted earlier, the majority of cluster methods assume a static geography and work with static spatial point patterns (instead of location histories) to represent cases and controls. The spatial coordinate employed may be the place of residence of cases at time of diagnosis, death, hospitalization, or whatever health-related event is being studied. But clustering of cases at time of diagnosis or death is often of little scientific or practical interest in terms of enhancing our understanding of health–environment relationships. Of greater import is whether there is clustering in the locations where the causative exposures occurred, but this question cannot be adequately addressed by techniques that employ a static world-view because those approaches implicitly assume the duration between exposure and the date of the health related event (e.g., diagnosis, death) to be negligible. When exploring space–time interaction – whether nearby cases tend to occur at about the same time – the Knox test (Knox, 1964) employs critical time and space distances that may be specified to reflect a latency period and the average distance individuals might move during this period. But to date and to our knowledge none of the available tests for geographic clustering take into account disease latency for location histories. Methods for addressing this need are proposed later in this chapter.

For purposes of this chapter we make a distinction between the evolution of risk through time of a known exposure (e.g., when the exposure began, ended, the mid-point, as well as changes in the exposure level through time) and the definition of a time window within which an unknown exposure *might have occurred* that plausibly could explain a known disease outcome (what we refer to

in this article as the *exposure window*). Let us now consider approaches that have been used for modeling latency.

Langholz *et al*. (1999) observed that effects of latency as described in the epidemiological literature are largely insufficient for addressing questions related to public health. They proposed latency models using bilinear and exponential decay functions, and fitted these models to case-control data within a likelihood framework. Their working definition of latency is the function describing how the relative risk associated with a *known exposure* changes through time. So, for example, in their analysis of lung cancer in a cohort of uranium miners they found that ' . . . relative risk associated with exposure increases for about 8.5 years and thereafter decreases until it reaches background levels after about 34 years'. As for most latency models of occupational studies, Langholz's metric was calculated for a known exposure – for example, the period of employment. For purposes of clustering we are interested in determining whether the residential histories of cases clustered during those times when causative exposures plausibly might have occurred, but *we do not necessarily know what those exposures might be*. We thus wish to use our admittedly inadequate knowledge of cancer latency to define exposure windows that bracket those time periods within which an environmental exposure *might* be associated with an observed cancer. This could indicate, for example, those times in a person's life when exposures (should they occur) are most likely to result in a cancer at some later date. This is an important distinction that, as noted above, must be kept in mind for the remainder of this chapter.

Exposures early in life and over an individual's life course may be important risk factors for the onset of chronic diseases such as cancer (Barker, 1992; Han *et al*., 2004; Kuh and Ben-Shlomo, 1997). But how can exposures during the life course be accounted for when modeling latency and exposure windows? Robins and Greenland (1991) showed that in cohort analyses, years of life lost (YLL) due to early exposures cannot be estimated without bias in the absence of causal models for how exposure causes death. Morfeld (2004) demonstrated this result analytically, resulting in a proposed framework for formulating such causal models (e.g., Robin's G-estimation procedure (Robins, 1997; Rothmann and Greenland, 1998) that can be used to estimate the latency between exposure and death). Of course any results from an exploratory analysis with no *a priori* hypothesis would need to be verified with another study. A model that links exposures and latency periods to the health outcomes thus appears to be required in order to evaluate alternative specifications of exposure windows, an important result that we will refer to later in this chapter.

For purposes of clustering, the putative exposure is often unknown, and we therefore must be able to handle uncertainty in exposure windows. Later in this chapter we define approaches for explicitly modeling exposure windows, and for specifying sampling distributions for exposure windows. These can then be used to evaluate the sensitivity of the cluster statistics to alternative specifications of and uncertainty in the exposure windows. But in general, the latency model employed should be specified to correspond to some *a priori* hypothesis regarding disease causation – a causal model.

19.4. AGE-DEPENDENT MODEL OF DISEASE LATENCY AND EXPOSURE WINDOWS

Detailed specification of a latency model requires a causal model of how disease results

in death. At this writing our knowledge of the causes of most cancers is incomplete and in almost all instances is insufficient to fully specify such a model. But in order to tackle this problem it first is necessary to develop an understanding of the information the construction of such a model might require. We therefore now consider how one might construct and then employ a model of disease latency within the framework of Q-statistics, using a simple and necessarily unrealistic age-dependent function as our point of departure. As more realistic models of disease causation are developed they may be radically different in form and will replace what we acknowledge is a simplistic first step. But for now and for convenience define the latency $\Delta L(A_d)$ as the duration between the age of the participant at the time of onset of the condition, $E_1(A_d)$ (age of the participant at that date when the participant has the beginnings of a cancer, yet to be diagnosed) and the age at diagnosis, A_d (Figure 19.2). Further, suppose the exposure window – the time in an individual's life course when he or she is biologically vulnerable should an exposure occur – commences at age $E_0(A_d)$ and ends at age $E_1(A_d)$. Recall the distinction made in the Introduction regarding exposure windows and an actual exposure. The exposure window is simply that time in a person's life when a causative exposure *might* have occurred and given rise to the observed cancer – that time interval from $E_0(A_d)$ and $E_1(A_d)$.

For the purposes of this chapter we will assume the age at which latency begins is the age at which the exposure window ends ($E_1(A_d)$) although this does not have to be the case and the modeling approach (below) is readily adapted to instances in which the end of the exposure window is not the same as the beginning of the latency period. We would like to model the exposure window and latency as functions of the age at diagnosis, A_d. The duration of the exposure window is therefore age dependent and we now write $\Delta E(A_d) = E_1(A_d) - E_0(A_d)$, and the duration of the latency period is $\Delta L(A_d) = A_d - E_1(A_d)$. For our purposes we wish to construct a model of $\Delta E(A_d) + \Delta L(A_d)$ so that the duration of the latency period and exposure window becomes shorter as the age at diagnosis decreases, since we wish to avoid implausible situations such as causative exposures occurring after the age at diagnosis. Notice, however, that the model can be specified in a manner that would allow maternal exposures prior to conception. We would also like the model to allow *in utero* exposures occurring after conception. To accomplish these objectives we employ a modified form of the logistic equation initially attributed to Verhulst (1838, 1845). Define the variable g at a given age of diagnosis to be:

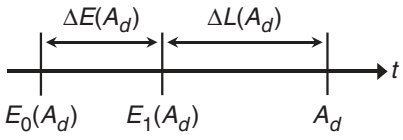


Figure 19.2 Schematic of a model of age-dependent exposure windows beginning at $E_0(A_d)$, ending at $E_1(A_d)$, and followed by latency $\Delta L(A_d)$, with latency ending at diagnosis at age A_d.

$$g(A_d) = \frac{\Delta E(A_d) + \Delta L(A_d)}{\max(\Delta E(A_d) + \Delta L(A_d))}. \quad (19.1)$$

This is the duration between the beginning of the exposure window to the age at diagnosis, scaled to the range 0 ... 1, by dividing by the maximum of that duration over all

ages considered. Now define the parameter g_0 to be:

$$g_0 = \frac{\min(\Delta E(A_d) + \Delta L(A_d))}{\max(\Delta E(A_d) + \Delta L(A_d))}. \quad (19.2)$$

This is the smallest possible value of $g(A_d)$. The model of the latency and exposure window as a function of age is then:

$$g(A_d) = \frac{1}{1 + \left(\frac{1}{g_0} - 1\right) e^{-rA_d}}. \quad (19.3)$$

Here r is a parameter describing the rate of increase of $g(A_d)$ as a function of age at diagnosis, with positive values indicating that the time period between the onset of the causative exposure and the end of the latency period increases as the age at diagnosis increases (Figure 19.3). Hence equation (19.3) is how we model $g(A_d)$ and equation (19.1) is the relationship between $g(A_d)$ and the latency and exposure windows at a given age of diagnosis.

19.5. SAMPLING DISTRIBUTIONS FOR EXPOSURE WINDOWS

With an age-dependent model of the latency and exposure windows defined we now concern ourselves with models of their uncertainty. Recall that exposure windows represent that time interval within which a causative environmental exposure plausibly could have occurred. Notice that we observe the cancer outcome (e.g., date of diagnosis) but do not know whether the cancer was in fact caused by an environmental exposure, nor what the exposure might actually be. This is in contrast to models of latency that were summarized in the Introduction, for which the exposure and its timing are known (or at least presumed known, being related for example to employment dates), as well as the date of diagnosis or death. Since in our case the exposures are not observable we require a sampling distribution for exposure windows that will allow us to assess the sensitivity of any observed case clustering to uncertainty in that exposure window.

We will accomplish this by modeling exposure windows for an individual with a given age at diagnosis as the waiting time from the beginning of the exposure window ($E_0(A_d)$) to the end of that exposure window ($E_1(A_d)$). Our approach will be to find the duration of the exposure window for individuals of a given age using the function in equation (19.3) and solving for $\Delta E(A_d)$ in equation (19.1). We then obtain individual realizations of that exposure window by sampling from a distribution of waiting times. Suppose we define events as being the beginning and end of an exposure window, and that these events are separated by a waiting time $\Delta E(A_d)$. Assume $E_0(A_d)$ and $E_1(A_d)$ are Poisson distributed and that the Poisson process has intensity λ. For a given waiting time we can estimate the intensity of the Poisson process adjusting for edge effects as:

$$\hat{\lambda} = \frac{2}{\Delta E(A_d) + 1}. \quad (19.4)$$

Or when ignoring edge effects as:

$$\hat{\lambda} = \frac{1}{\Delta E(A_d)}. \quad (19.5)$$

The cumulative distribution of $\Delta E(A_d)$ is then estimated by:

$$D(\Delta E(A_d)) = 1 - e^{-\hat{\lambda}\Delta E(A_d)} \quad (19.6)$$

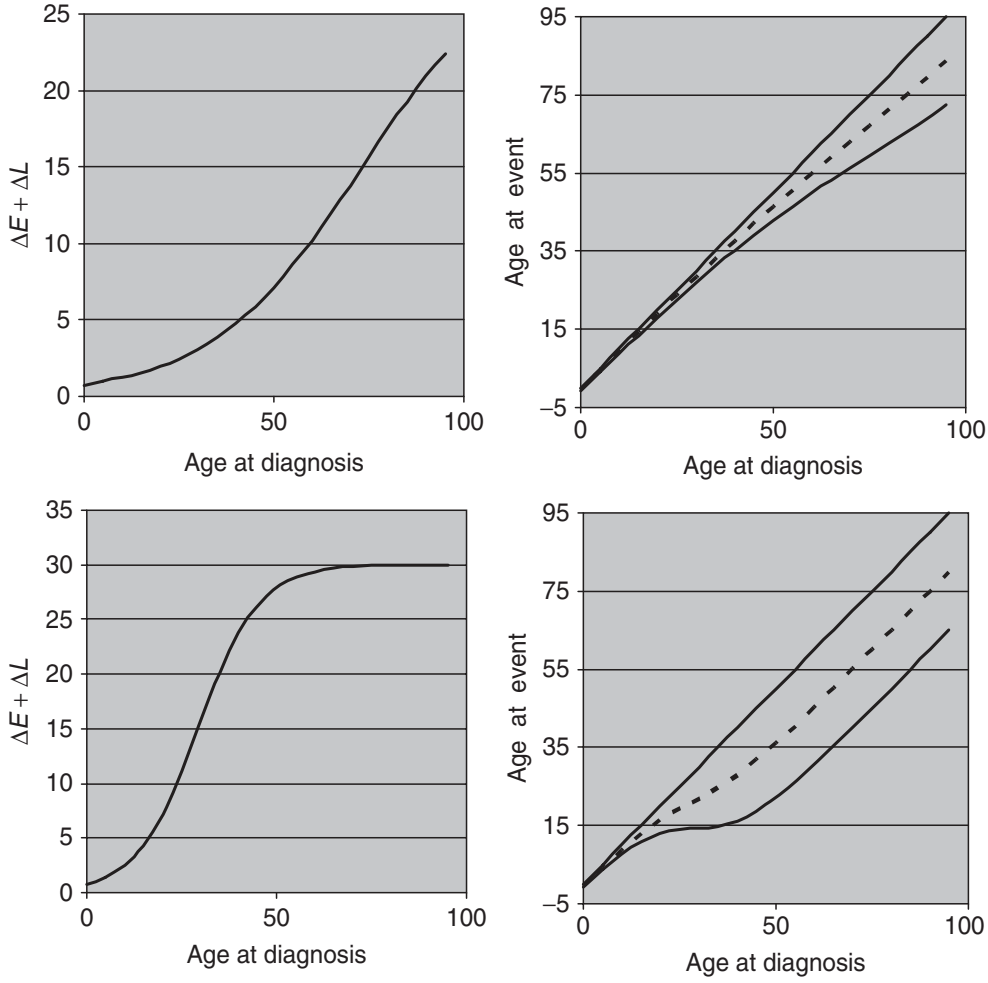


Figure 19.3 Age dependent model of exposure window and latency. The sum of the exposure window plus the latency as a function of age at diagnosis is shown in the first column. The second column shows the age at diagnosis (top solid line), the age at the end of the exposure window (dashed line) and the age at the beginning of the exposure window (bottom solid line). Top row: $r = 0.05$; bottom row $r = 0.125$. Minimum latency is 0.375 years, maximum latency is 15 years. Minimum exposure window is 0.375 years, maximum exposure window is 15 years.

And the probability of $\Delta E(A_d)$ is given by:

$$P(\Delta E(A_d)) = \hat{\lambda} e^{-\hat{\lambda}_{\Delta E(A_d)}}. \qquad (19.7)$$

Having defined exposure windows and their uncertainty we now turn to cluster statistics that account for human mobility.

19.6. THE DETECTION OF CLUSTERING IN RESIDENTIAL HISTORIES

In this section we first review Q-statistics. We then define exposure traces that are the geographic projection of exposure windows and extend the Q-statistics to provide global, local, and focused tests that account

for risk factors, covariates, and disease latency. We then describe an experimental data set for bladder cancer in southeastern Michigan, and apply some of these new methods to this dataset to illustrate the approach.

Jacquez *et al*. (2005, 2006) developed global, local, and focused tests for case-control clustering of residential histories for use with chronic diseases such as cancer and that account for covariates and other risk factors such as smoking. Readers unfamiliar with Q-statistics may wish to refer to the original works. We now briefly present these techniques and then extend them to account for exposure windows.

Define the coordinate $\mathbf{u}_{i,t} = \{x_{i,t}, y_{i,t}\}$ to indicate the geographic location of the ith case or control at time t. Residential histories can then be represented as the set of space–time locations:

$$\mathbf{R}_i = \{\mathbf{u}_{i0}, \mathbf{u}_{i1}, \ldots, \mathbf{u}_{iT}\}. \tag{19.8}$$

This defines individual i at location $\mathbf{u}_{i0}$ at the beginning of the study (time 0), and moving to location $\mathbf{u}_{i1}$ at time $t = 1$. At the end of the study individual i may be found at $\mathbf{u}_{iT}$. T is defined to be the number of unique location observations on all individuals in the study. We now define a case-control identifier, c_i, to be:

$$c_i = \begin{cases} 1 & \text{if and only if } i \text{ is a case} \\ 0 & \text{otherwise.} \end{cases} \tag{19.9}$$

Define n_a to be the number of cases and n_b to be the number of controls. The total number of individuals in the study is then $N = n_a + n_b$. Let k indicate the number of nearest neighbors to consider when evaluating nearest neighbor relationships, and define a nearest neighbor indicator to be:

$$\eta_{i,j,k,t} = \begin{cases} 1 & \text{if and only if } j \text{ is a } k \text{ nearest} \\ & \text{neighbor of } i \text{ at time } t \\ 0 & \text{otherwise} \end{cases} \tag{19.10}$$

We then can define a binary matrix of kth nearest neighbor relationships at a given time t as:

$$\boldsymbol{\eta}_{k,t} = \begin{bmatrix} 0 & \eta_{1,2,k,t} & \cdot & \cdot & \eta_{1,N,k,t} \\ \eta_{2,1,k,t} & 0 & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \eta_{N-1,N,k,t} \\ \eta_{N,1,k,t} & \cdot & \cdot & \eta_{N,N-1,k,t} & 0 \end{bmatrix}. \tag{19.11}$$

This matrix enumerates the k nearest neighbors (indicated by a 1) for each of the N individuals. The entries of this matrix are 1 (indicating that j is a k nearest neighbor of i at time t) or 0 (indicating j is not a k nearest neighbor of i at time t). It may be asymmetric about the 0 diagonal since nearest neighbor relationships are not necessarily reflexive. Since two individuals cannot occupy the same location, we assume at any time t that any individual has k unique k-nearest neighbors. The row sums thus are equal to $k(\eta_{i,\bullet,k,t} = k)$ although the column sums vary depending on the spatial distribution of case control locations at time t. The sum of all the elements in the matrix is Nk.

Alternative specifications of the proximity metric may be used – the metrics do not have to be nearest neighbor relationships in order for the Q-statistics to work. We prefer to use nearest neighbor relationships because

they are invariant under changing population densities, unlike geographic distance and adjacency measures. There is also some evidence that nearest neighbor metrics are more powerful than distance- and adjacency-based measures (Jacquez and Waller, 1997). Still, one then may be faced with the question of 'how many nearest neighbors (k) should I consider?' In certain instances one may have prior information that suggests that clusters of a certain size should be expected, and this can serve as a guide to specification of k. When prior information is lacking one may wish to explore several levels of k. In these instances Tango (2000, 2006) advocates using the minimum p-value obtained under each level of k considered as the test statistic. Jacquez *et al*. (2006) evaluated different levels of k to determine sensitivity of the results to specification of k. Each of these approaches has advantages and may be preferred in different situations.

There exists a $1 \times T + 1$ vector denoting those instants in time when the system is observed and the locations of the individuals are recorded. We can then consider the sequence of T nearest neighbor matrices defined by:

$$\boldsymbol{\eta}_k^T = \{\eta_{k,t} \; \forall t = 0 \ldots T\}. \qquad (19.12)$$

This defines the sequence of k nearest neighbor matrices for each unique temporal observation recorded in the data set, and thus quantifies how spatial proximity among the N individuals changes through time.

19.7. ADJUSTING FOR COVARIATES AND OTHER RISK FACTORS

In the absence of knowledge of covariates and other risk factors simple randomization may be used when evaluating the statistical significance of the above statistics. This is accomplished by holding the location histories for the cases and controls constant, and by then sprinkling the case-control identifiers at random over the residential histories. This corresponds to a null hypothesis where the probability of an individual being declared a case ($c_i = 1$) is proportional to the number of cases in the data set, or:

$$p(c_i = 1|H_{0,I}) = \frac{n_1}{n_0 + n_1}. \qquad (19.13)$$

Here n_1 is the number of cases and n_0 is the number of controls, and $H_{0,I}$ indicates a null hypothesis corresponding to Goovaerts and Jacquez's (2004) type I neutral model of spatial independence. This null hypothesis assumes the risk of being declared a case is the same over all of the N case and controls. When covariates and risk factors are quantified we may wish to incorporate that information into the null hypothesis. Any case-clustering that is found then will be *above and beyond* the modeled risk factors and covariates, and will thus indicate the possible presence of risk sources beyond those specified under this null hypothesis.

19.7.1. Logistic model of the probability of being a case

In order to provide a more realistic model of the risk of being a case, we must make the probability of being declared a case a function of the covariates and risk factors one wishes to incorporate under the null hypothesis. We will accomplish this task using logistic regression. Logistic models are used for binary response variables. Let $\mathbf{x}$ denote the vector of covariates and risk factors. Further, let $p = \Pr(c = 1|\mathbf{x})$ denote the response probability to be modeled, which is the probability of person i being a

case given that person's vector of covariates and risk factors. The linear logistic model may then be written as:

$$\text{logit}\,(p) = \log\,(p/1-p) = \alpha + \boldsymbol{\beta}'\mathbf{x} \tag{19.14}$$

and the equation for predicting the probability of being a case given the vector of covariates and risk factors for the ith individual is:

$$\hat{p}(c_i = 1|\mathbf{x}_i) = \frac{\mathrm{e}^{\alpha+\boldsymbol{\beta}'\mathbf{x}_i}}{1+\mathrm{e}^{\alpha+\boldsymbol{\beta}'\mathbf{x}_i}}. \tag{19.15}$$

Here the logit function is the natural log of the odds, α is the intercept parameter, and $\boldsymbol{\beta}$ is the vector of regression (slope) coefficients. One then fits the regression model to the vector of covariates and risk factors to calculate the intercept and slope parameters. These are then used to calculate, for each individual, the probability of being a case given that individual's known covariates and risk factors.

19.7.2. Randomization accounting for risk factors and covariates

We use approximate randomization to evaluate the probability of a given Q-statistic under the null hypothesis that the likelihood of being a case is a function of the covariates and risk factors specified under the logistic model in equation (9.14). This null hypothesis thus effectively accounts for the risk factors and covariates in the vector $\mathbf{x}$. To evaluate the reference distribution for a given Q-statistic we follow these steps.

Step 1. Calculate statistic (Q^*) for the observed data.

Step 2. Sprinkle the case-control identifier c_i over the residential histories of the participants in a manner consistent with the desired null hypothesis, and conditioned on the observed number of cases. Assume we have n_1 cases, N participants and that P_i is the probability of the ith participant being a case. Notice the P_i are provided by the logistic equation.

Step 2.1. Rescale the P_i as follows:
$$P_i' = P_i / \sum_{i=1}^{N} P_i.$$

Step 2.2. Map the P_i' to the interval 0 ... 1. For example, assume we have $N = 2$ participants, $n_1 = 1$ case and that $P_1 = 0.7$ and $P_2 = 0.8$. P_1' then maps to the interval [0 ... 0.7/1.5) and P_2' maps to the interval [0.7/1.5 ... 1.5/1.5).

Step 2.3. Allocate a case by drawing a uniform random number from the range [0 ... 1). Set the case identifier equal to 1 ($c_i = 1$) where i is the identifier corresponding to the study participant whose interval for P_i' contains the random number.

Step 2.4. Rescale as shown in Step 2.1 but not including the probability for the participant whose case identifier was assigned in Step 2.3.

Step 2.5. Repeat Steps 2.2–2.4 until all of the n_1 case identifiers are assigned.

Step 2.6. Set the remaining $N - n_1$ case identifiers to 0, these are the controls.

Notice steps 2.1–2.6 result in one realization of the distribution of case-control identifiers.

Step 3. Calculate Q for the realization from Step 2.

Step 4. Repeat Steps 2 and 3 a specified number of times (e.g., 999) accumulating the reference distribution of Q.

Step 5. Compare Q^* to this reference distribution to evaluate the statistical probability of observing Q^* under the null hypothesis that accounts for the known risk factors and covariates.

19.8. LOCAL AND FOCUSED CLUSTERING OF EXPOSURE TRACES

Exposure traces are defined as the residential mobility that transpires for an individual during the exposure window, $\Delta E(A_{di})$. Notice we are now subscripting the age of diagnosis with the letter i to indicate the age of diagnosis for the ith individual. Therefore, exposure traces are those portions of a case's residential history that were traversed while that individual was thought to be at risk to a cancer-causing exposure – where they were when they were in that portion of their lifespan corresponding to $\Delta E(A_{d,i})$. This concept of an *exposure trace* assumes a natural history of carcinogenesis in which the causative exposures occur, followed by a latency period which concludes when the cancer is diagnosed. This is easily modified to fit other models of the natural history of carcinogenesis, including other relevant windows such as the lag between the onset of a fully developed cancer and diagnosis. Given the residential history for case i, $\mathbf{R}_i$, denote the space–time coordinate at time of diagnosis as $\mathbf{u}_{i,t_D}$, noting that $\mathbf{u}_{i,t_D} \in \mathbf{R}_i$. We can then define that subset of the residential history $\mathbf{R}_i$ during which causative exposures might have occurred as:

$$\mathbf{R}_i^E = \{\mathbf{u}_{i,t};\ (t_{i,D} - \Delta L(A_{d,i}) > t > (t_{i,D} - \Delta L(A_{d,i}) - \Delta E(A_{d,i})). \tag{19.16}$$

Here $t_{i,D}$ is the time of diagnosis for individual i. The term $(t_{i,D} - \Delta L(A_{d,i}))$ is the time when the exposure window ended and the latency period began. The term $(t_{i,D} - \Delta L(A_{d,i}) - \Delta E(A_{d,i}))$ indicates the time prior to diagnosis when the exposure window began. This allows us to move between the age representation where things are measured relative to age at diagnosis to an absolute time representation using, for example, the Gregorian calendar. Hence equation (9.16) denotes that portion of case i's residential history in which s/he was in the exposure window. Call this the *exposure trace.* As noted earlier in the Introduction, effective specification of exposure windows, and hence of exposure traces, requires a causal model of how the exposure(s) causes cancer. The exposure trace for case $i(\mathbf{R}_i^E)$ records those places where that individual resided while s/he might have experienced causative exposures. Now define an indicator, $e_{i,t}$ as:

$$e_{i,t} = \begin{cases} 1 & \text{if and only if } t \text{ is within the} \\ & \text{exposure trace is for individual } i, \\ 0 & \text{otherwise.} \end{cases} \tag{19.17}$$

When $e_{i,t}$ is 1, let us say the exposure trace is 'active'. A local case-control test for spatial clustering of exposure traces at time t is then:

$$Q_{i,k,t}^E = c_i\ e_{i,t} \sum_{j=1}^{N} \eta_{i,j,k,t}\ c_j\ e_{j,t}. \tag{19.18}$$

This is the count, at time t, of the number of k nearest neighbors of case i's exposure trace that are also cases and whose exposure traces are also active. This statistic will be large when the active exposure traces of a group of cases cluster about case i at time t.

We can explore whether exposure traces of cases tend to cluster spatially about certain individuals through time. A statistic sensitive to this pattern is:

$$Q_{i,k}^E = \sum_{t=0}^{T} Q_{i,k,t}^E. \tag{19.19}$$

$Q^E_{i,k}$ will tend to be large when active exposure traces for the other cases tend to persistently cluster around the active exposure trace of the ith case.

We can also ask whether the exposure traces of cases cluster about specific locations (e.g., point source releases of carcinogens) that we refer to as a focus:

$$Q^E_{F,k,t} = \sum_{j=1}^{N} \eta_{F,j,k,t}\, c_j\, e_{j,t}. \qquad (19.20)$$

Here $\eta_{F,j,k,t}$ is 1 if individual j is a k nearest neighbor of the focus at time t, and 0 otherwise. The statistic $Q^E_{F,k,t}$ is the count of the number of cases whose exposure traces are k nearest neighbors of the focus at time t. Notice these statistics can also be duration weighted as described by Jacquez *et al*. (2005).

19.8.1. Statistical probability of exposure traces

In order to evaluate whether exposure traces of the cases cluster we first must derive a procedure for generating representative times of diagnosis, latency periods, and exposure windows for the controls. Once this is accomplished we will be able to determine whether the exposure traces for the cases cluster relative to those so constructed for the controls. Given the residential history of a control, steps involved to accomplish this are:

1 Set the 'age at diagnosis' for each control to be their age at their time of interview for the study (notice researchers often may subtract one year from age at time of interview, to account for time between diagnosis and interview for cases).

2 Define the exposure window and latency period for each case and control using the time of diagnosis assigned in (1) and the model of latency as a function of age defined earlier. Notice this function could also be based on the covariates for each participant, or on the times of occurrence of a putative exposure source of interest. Completion of (1) and (2) will result in dates of diagnosis, and definition of exposure windows, latency periods, and exposure traces for both cases and controls.

3 Calculate the desired test statistic for exposure traces, for the original (not randomized data), Q^* (e.g., equation (19.20) for focused clustering, equation (19.19) for local clustering, etc.).

4 Assign case-control identifiers across the residential histories employing the logistic model described earlier in order to account for known risk factors and covariates. This will result in a possible arrangement of cases and controls (a realization) that accounts for the risk factors and covariates. Hence any statistically significant clustering observed in the exposure traces may be attributable to causes other than the risk factors and covariates included in the logistic model.

5 For the realization from (4), calculate the desired test statistic for clustering of exposure traces (Q).

6 Repeat (4) and (5) a desired number of times to construct the reference distribution of the statistic under the null hypothesis (the null distribution of Q).

7 Evaluate the probability of the observed clustering of exposure traces under the null hypothesis by comparing the value of the test statistic for the observed data (Q^*) to the reference distribution for Q from (6).

19.9. EXAMPLE: BLADDER CANCER IN SOUTHEASTERN MICHIGAN

A population-based bladder cancer case-control study is underway in southeastern Michigan.

Cases diagnosed in the years 2000–2004 are being recruited from the Michigan State Cancer Registry. Controls are being frequency matched to cases by age (±5 years), race, and gender, and are being recruited using a random digit dialing procedure from an age-weighted list. At this stage of recruitment, controls are not adequately matched; therefore, age, race, and gender are included in the logistic regression model that accounts for covariates. To be eligible for inclusion in the study, participants must have lived in the eleven county study area for at least the past five years and had no prior history of cancer (with the exception of non-melanoma skin cancer). Participants are offered a modest financial incentive and research is approved by the University of Michigan IRB-Health Committee. The data analyzed here are from 219 cases and 437 controls. As part of the study, participants complete a written questionnaire describing their residential mobility. The duration of residence and exact street address were obtained, otherwise the closest cross streets were provided. Approximately 66% of cases' person-years and 63% of controls' person-years were spent in the study area. Of the residences within the study area, 88% were automatically geocoded or interactively geocoded with minor operator assistance. The unmatched addresses were manually geocoded using self-reports of cross streets with the assistance of internet mapping services (6%); if cross streets were not provided or could not be identified, residence was matched to town centroid (6%).

Address histories were collected for those industries believed to emit contaminants associated with bladder cancer. These were identified using the Toxics Release Inventory (EPA, 2000) and the Directory of Michigan Manufacturers. Standard Industrial Classification (SIC) codes were adopted, but prior to SIC coding, industrial classification titles were selected. Characteristics of 268 industries, including, but not limited to, fabric finishing, wood preserving, pulp mills, industrial organic chemical manufacturing, and paint, rubber, and leather manufacturing, were compiled into a database. Each industry was assigned a start year and end year, based on best available data. Industries were geocoded following the same matching procedure as for residences: 89% matched to the address, 5% were placed on the road using best informed guess, and as a last resort, 6% were matched to town centroid.

19.10. BLADDER CANCER: ANALYSIS

Jacquez *et al*. (2006) addressed four hypotheses regarding clusters of bladder cancer in southeastern Michigan:

A0: Bladder cancer cases in southeastern Michigan are not clustered.

A1: There is global and local space–time clustering of bladder cancer cases.

A2: The clusters may be explained entirely by known risk factors (e.g., smoking) and covariates.

This probability was then incorporated in the randomization procedure as described earlier, resulting in a null hypothesis that accounts for smoking, age, gender, education, and race. Any clustering that is observed thus is above and beyond any case clustering due to these risk factors and covariates. Increased smoking is associated with higher probability of being a case; this risk increases with age, and is elevated for whites and females. Bladder cancer typically afflicts older white males to a greater extent than the remainder of the population (Silverman *et al*., 1996).

A3: There is clustering of bladder cancer cases about industries known to emit bladder cancer carcinogens that is not explained by known risk factors and covariates.

They used the global and local Q-statistics not adjusting for covariates and risk factors to address hypotheses A0 and A1. They then used the logistic model to adjust for smoking, age, gender, education, and race in order to evaluate hypotheses A2–A3, employing the following function to evaluate the probability of being a case:

$$\hat{p}\,(c_i = 1|\mathbf{x}_i) = \frac{\mathrm{e}^{\left(\begin{array}{l}2.0359 - 0.0125^{*}Age_i - 0.9396^{*}Gender_i + 0.1900^{*}Educate_i + \\ 0.0557^{*}Race_i - 0.2438^{*}Cignum_i\end{array}\right)}}{1 + \mathrm{e}^{\left(\begin{array}{l}2.0359 - 0.0125^{*}Age_i - 0.9396^{*}Gender_i + 0.1900^{*}Educate_i + \\ 0.0557^{*}Race_i - 0.2438^{*}Cignum_i\end{array}\right)}} \tag{19.21}$$

Here females experience a higher risk because controls are in the process of being frequency matched to cases in the ongoing study, and in this dataset, a greater proportion of cases are females than controls. In this chapter, results are presented for $k = 7$ nearest neighbors. Results for additional nearest neighbors are discussed in Jacquez *et al*. (2006).

The first hypothesis ***A0: Bladder cancer cases in southeastern Michigan are not clustered*** was evaluated without correcting for the known risk factors and covariates. The Global Q statistic was 1.198437 and was significant ($p = 0.001$), and hypothesis A0 was rejected. Next, hypothesis ***A1: There is space–time clustering of bladder cancer cases in southeastern Michigan*** was evaluated using the spatial and temporally local Q-statistics of equations (19.10) and (19.12) in Jacquez *et al*. (2005). This effectively decomposed the observed global clustering into local contributions. Persistent case clusters were found in Oakland, Ingham, and Jackson counties. Hypothesis A1 was accepted and Jacquez *et al*. (2006) concluded there is persistent case clustering in these three counties. However, whether these clusters may be explained by smoking and the covariates age, gender, race, and education remained to be evaluated.

Next, the researchers evaluated hypothesis ***A2: The clusters may be explained by known risk factors and covariates***. To accomplish this they incorporated the probabilities calculated from the logistic model in equation (19.21) into the randomization procedure as described in section 19.7.2. They then recalculated the probabilities of the global Q statistic used to evaluate A0. Because the geometry of the residential histories doesn't change, the values of the statistic were unchanged. After adjustment for smoking and covariates the P value slightly increased to 0.003 from 0.001 before adjustment. Hypothesis A2 was not accepted, and the authors concluded the global case clustering of residential histories was not sufficiently explained by smoking and the covariates. Significant local clustering also remained, and was persistent through time. In all, 26 local clusters were significant after covariate adjustment. They were found in Lapeer, Ingham, Oakland, and

Jackson counties. The clusters in Lapeer and Jackson counties were comprised of 1–3 cluster centers, and are ephemeral. The clusters in northwestern Ingham county appeared in 1950, concentrated to the northwest of Lansing and persisted into 2000. Numerous clusters appeared in central and southeastern Oakland county beginning in the 1950s and persisted to the present day. The authors suggested that the grouping of these local case clusters into two areas and their persistence through time might indicate the possible action of a causal agent or an unknown covariate. They therefore explored hypothesis ***A3: There is clustering of bladder cancer cases about industries known to emit bladder cancer carcinogens that is not explained by known risk factors and covariates***. Bladder cancers have a multiplicity of possible causative exposures. Using a database of 268 industries that emitted known or suspected bladder cancer carcinogens, they analyzed case clustering of residential histories about these industries both with and without adjustment for smoking and the four covariates. The global version of the focused test was significant at the 0.015 level before covariate adjustment and remained significant ($p = 0.035$) after the covariates and smoking were accounted for. Considering the 268 business address histories one at a time, the researchers found 22 industries that were significant cluster foci, located in Oakland (19 clusters), Ingham (2), and Jackson (1) counties. Clusters in central and southeastern Oakland county appeared in the 1930s and persisted to the present day.

The prospect of environmental pollution originating from these facilities being associated with bladder cancer is intriguing; however, caution is necessary until the study is complete. Occupational histories are being collected and will be incorporated as risk factors in the logistic regression model, thus creating a neutral model that includes smoking and occupational exposures, along with key covariates. Until then, we cannot rule out occupational exposures in explaining the focused clustering around certain industries. In the interest of public health, however, it is worth exploring those facilities with the most extreme p-values to single out those that consistently are at the center of a cluster of cases. Once identified, additional epidemiological investigation may be warranted to uncover a biologically plausible exposure, and to determine whether individuals in the vicinity of the operation actually demonstrate a body burden for the suspected carcinogen.

19.11. DISCUSSION AND FUTURE DIRECTIONS

The case-control epidemiological study design provides a wealth of information at the individual level regarding exposures, risks, risk modifiers, and covariates. When designing such a study the researcher often is concerned with assessing a few putative exposures, and in determining whether there are significant differences in these exposures between the case and control populations. As such, the case-control design is not inherently spatial, nor is it particularly well suited or even capable of assessing risk factors other than those specified in the original design.

The approaches described in this chapter may prove to be a highly useful addition to the traditional aspatial case-control design because they allow researchers to identify local groups of individuals whose risk exceeds that accounted for by the known risk factors and covariates incorporated under the designed study. Efforts in developing causal models for latency and exposure timing are evolving, and the approach outlined here will allow researchers to incorporate these models into future cluster analyses that account for human mobility. In addition,

while the application presented here uses residential histories, this approach may also be used to investigate disease clustering using occupational histories, or other forms of human mobility.

The ability of local and focused tests to quantify pockets of cases whose excess risk might be attributable to specific locations or point sources is a powerful addition to the inferential toolbox. While such a tool can never of itself assess the dose–response relationship necessary to attribute risk to a specific location or point source, the ability to temporally and geographically localize the putative exposure source makes it possible to begin the assessment of dose–response relationships. Once such a putative focus has been identified, the next step may involve techniques for modeling exposure that will provide a more accurate and detailed description of the spatial and temporal variability in exposure. And once a specific point source is identified, the task of quantifying the type and quantity of releases of agents that plausibly might give rise to the observed health outcome may begin.

ACKNOWLEDGMENTS

This research was funded by grants R43CA117171, R01CA096002, and R44CA092807 from the National Cancer Institute. The views expressed in this publication are those of the researchers and do not necessarily represent that of the NCI.

REFERENCES

Aschengrau, A., Ozonoff, D., Coogan, P., Vezina, R., Heeren, T. and Zhang, Y. (1996). Cancer risk and residential proximity to cranberry cultivation in Massachusetts, *American Journal of Public Health*, **86**(9): 1289–1296.

Barker, D. (1992). *Fetal and Infant Origins of Adult Disease.* London: BMJ Publishing.

Bellander, T., Berglind, N., Gustavsson, P., Jonson, T., Nyberg, F., Pershagen, G. and Jarup, L. (2001). Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in Stockholm. *Environmental Health Perspectives*, **109**(6): 633–639.

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society Series A*, **154**: 143–155.

Beyea, J. and Hatch, M. (1999). Geographic exposure modeling: a valuable extension of geographic information systems for use in environmental epidemiology. *Environmental Health Perspectives*, **107**(suppl 1): 181–190.

Bonner, M.R., Han, D., Nie, J., Rogerson, P., Vena, J.E., Muti, P., Trevisan, M., Edge, S.B. and Fraudenheim, J.L. (2005). Breast cancer risk and exposure in early life to polycyclic aromatic hydrocarbons using total suspended particulates as a proxy measure. *Cancer Epidemiology Biomarkers and Prevention*, **14**(1): 53–60.

Bradley, D. (1988). The scope of travel medicine. In: *First Conference on International Travel Medicine*, pp. 1–9. Zurich: Springer Verlag.

Brody, J.G., Vorhees, D.J., Melly, S.J., Swedis, S.R., Drivas, P.J. and Rudel, R.A. (2002). Using GIS and historical records to reconstruct residential exposure to large-scale pesticide application. *Journal of Exposure Analysis and Environmental Epidemiology*, **12**(1): 64–80.

Cliff, A.D. and Haggett, P. (2003). On changing contexts for epidemic modeling. In: Toubiana, L., Viboud, C., Flahault, A. and Valleron, A.-J. (eds), *Geography and Health*, pp. 1–18. Paris: Inserm.

Collia, D.V., Sharp, J. and Giesbrecht, L. (2003). The 2001 National Household Travel Survey: a look into the travel patterns of older Americans. *Journal of Safety Research*, **34**(4): 461–470.

Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society Series B*, **52**: 73–104.

Davies, R. (1964). *A History of the World's Airlines.* New York: Oxford University Press.

Elliott, P., Briggs, D., Morris, S., de Hoogh, C., Hurt, C., Jensen, T.K., Maitland, I., Richardson, S.,

Wakefield, J. and Jarup, L. (2001). Risk of adverse birth outcomes in populations living near landfill sites. *British Medical Journal*, **323**(7309): 363–368.

EPA (2000). Toxics Release Inventory (TRI) Data Files, Environmental Protection Agency.

Goodchild, M. (2000). GIS and transportation: status and challenges. *GeoInformatica*, **4**: 127–139.

Goovaerts, P. and Jacquez, G.M. (2004). Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *International Journal of Health Geographics*, **3**(1): 14.

Hagerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association*, **24**: 7–21.

Han, D., Rogerson, P.A., Nie, J., Bonner, M.R., Vena, J.E., Vito, D., Muti, P., Trevisan, M., Edge, S.B. and Freudenheim, J.L. (2004). Geographic clustering of residence in early life and subsequent risk of breast cancer (United States). *Cancer Causes and Control*, **15**(9): 921–929.

Jacquez, G.M. (2004). Current practices in the spatial analysis of cancer: flies in the ointment. *International Journal of Health Geographics*, **3**(1): 22.

Jacquez, G.M., Kaufmann, A., Meliker, J., Goovaerts, P., AvRuskin, G. and Nriagu, J. (2005). Global, local and focused geographic clustering for case-control data with residential histories. *Environmental Health*, **4**(1): 4.

Jacquez, G.M., Meliker, J.R., AvRuskin, G.A., Goovaerts, P., Kaufmann, A., Wilson, M. and Nriagu, J. (2006). Case-control geographic clustering for residential histories accounting for risk factors and covariates. **5**: 32 *International Journal of Health Geographics*.

Jacquez, G.M. and Waller, L. (1997). The effect of uncertain locations on disease cluster statistics. In: Mowerer, H.T. and Congalton, R.G. (eds), *Quantifying Spatial Uncertainty in Natural Resources: Theory and Application for GIS and Remote Sensing*. Chelsea MI: Arbor Press.

Klepeis, N.E., Nelson, W.C., Ott, W.R., Robinson, J., Tsang, A.M., Switzer, P., Behar, J.V., Hern, S. and Engelmann, W. (2001). The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Analysis and Environmental Epidemiology*, **11**(3): 231–252.

Knox, G. (1964). The detection of space time interactions. *Applied Statistics*, **13**: 25–29.

Kuh, D. and Ben-Shlomo, Y. (1997). *A Life Course Approach to Chronic Disease Epidemiology: Tracing the Origins of Ill-health from Early to Later Life*. Oxford: Oxford University Press.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine*, **14**(8): 799–810.

Langholz, B., Thomas, D., Xiang, A. and Stram, D. (1999). Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado Plateau uranium miners cohort. *American Journal of Industrial Medicine*, **35**(3): 246–256.

Long, L. (1992). Changing residence: comparative perspectives on its relationship to age, sex, and marital status. *Population Studies*, **46**: 141–158.

McNamee, R. and Dolk, H. (2001). Does exposure to landfill waste harm the fetus? Perhaps, but more evidence is needed. *British Medical Journal*, **323**(7309): 351–352.

Morfeld, P. (2004). Years of Life Lost due to exposure: Causal concepts and empirical shortcomings. *Epidemiologic Perspectives and Innovation*, **1**(1): 5.

Nuckols, J.R., Ward, M.H. and Jarup, L. (2004). Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental Health Perspectives*, **112**(9): 1007–1015.

Nyberg, F., Gustavsson, P., Jarup, L., Bellander, T., Berglind, N., Jakobsson, R. and Pershagen, G. (2000). Urban air pollution and lung cancer in Stockholm. *Epidemiology*, **11**(5): 487–495.

O'Leary, E.S., Vena, J.E., Freudenheim, J.L. and Brasure, J. (2004). Pesticide exposure and risk of breast cancer: a nested case-control study of residentially stable women living on Long Island. *Environmental Research*, **94**(2): 134–144.

Reuscher, T., Schmoyer, R. and Hu, P.S. (2002). Transferability of Nationwide Personal Transportation Survey data to regional and local scales. *Transportation Research Record*, **1817**: 25–32.

Reynolds, P., Hurley, S.E., Gunier, R.B., Yerabati, S., Quach, T. and Hertz, A. (2005). Residential proximity to agricultural pesticide use and incidence of breast

cancer in California, 1988–1997. *Environmental Health Perspectives*, **113**(8): 993–1000.

Robins, J. (1997). Causal inference from complex longitudinal data. In: Berkane, M. (ed.), *Latent Variable Modeling with Applications to Causality*, pp. 69–117. New York: Springer.

Robins, J. and Greenland, S. (1991). Estimability and estimation of expected years of life lost due to a hazardous exposure. *Statistics in Medicine*, **10**(1): 79–93.

Rothmann, K. and Greenland, S. (1998). *Modern Epidemiology*. Philadelphia: Lippincott-Raven.

Sabel, C.E., Boyle, P.J., Löytönen, M., Gatrell, A.C., Jokelainen, M., Flowerdew, R. and Maasilta, P. (2003). Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death. *American Journal of Epidemiology*, **157**(10): 898–905.

Sabel, C.E., Gatrell, A.C., Löytönen, M., Maasilta, P. and Jokelainen, M. (2000). Modelling exposure opportunities: estimating relative risk for motor neurone disease in Finland. *Social Science and Medicine*, **50**(7–8): 1121–1137.

Schaerstrom, A. (2003). The potential for time geography in medical geography. In: Toubiana, L., Viboud, C., Flahault, A. and Valleron, A.-J. (eds), *Geography and Health*, pp. 195–207. Paris: Inserm.

Scheiner, J. and Kaspar, B. (2003). Lifestyles, choice of housing location and daily mobility: the lifestyle approach in the context of spatial mobility and planning. *International Social Science Journal*, **55**: 319–332.

Silverman, D., Morrison, A. and Devesa, S. (1996). Bladder cancer. In: Schottenfeld, D. and Fraumeni, J. Jr., (eds), *Cancer Epidemiology and Prevention*, pp. 1156–1179. New York: Oxford University Press.

Sinha, G. and Mark, D. (2005). Measuring similarity between geospatial lifelines in studies of environmental health. *Journal of Geographical Systems*, **7**(1): 115–136.

Stellman, J.M., Stellman, S.D., Weber, T., Tomasallo, C., Stellman, A.B. and Christian, R. (2003). A geographic information system for characterizing exposure to Agent Orange and other herbicides in Vietnam. *Environmental Health Perspectives*, **111**(3): 321–328.

Swartz, C.H., Rudel, R.A., Kachajian, J.R. and Brody, J.G. (2003). Historical reconstruction of wastewater and land use impacts to groundwater used for public drinking water: exposure assessment using chemical data and GIS. *Journal of Exposure Analysis and Environmental Epidemiology*, **13**(5): 403–416.

Tango, T. (1995). A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine*, **14**(21–22): 2323–2334.

Tango, T. (2000). A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine*, **19**(2): 191–204.

Tango, T. (in press). A test with minimized *p*-value for spatial clustering applicable to case-control point data, *Biometrics*.

Turnbull, B.W., Iwano, E.J., Burnett, W.S., Howe, H.L. and Clark, L.C. (1990). Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology*, **132** (Suppl 1): S136–143.

Verhulst, P.F. (1838). Notice sur la loi que la population pursuit dans son accroissement. *Correspondance Mathematique et Physique*, **10**: 113–121.

Verhulst, P.F. (1845). Recherches Mathematiques sur La Loi D'Accroissement de la Population (Mathematical Researches into the Law of Population Growth Increase). *Nouveaux Memoires de l'Academie Royale des Sciences et Belles-Lettres de Bruxelles*, **18**: Art. 1, 1–45.

Ward, M., Nuckols, J., Weigel, S., Maxwell, S., Cantor, K. and Miller, R. (2000). Geographic information systems. A new tool in environmental epidemiology. *Annals of Epidemiology*, **10**(7): 477.

20

Neural Networks for Spatial Data Analysis

Manfred M. Fischer

20.1. INTRODUCTION

The term 'neural network' has its origins in attempts to find mathematical representations of information processing in the study of natural neural systems (McCulloch and Pitts, 1943; Widrow and Hoff, 1960; Rosenblatt, 1962). Indeed, the term has been used very broadly to include a wide range of different model structures, many of which have been the subject of exaggerated claims to mimic neurobiological reality.[1] As rich as neural networks are, they still ignore a host of biologically relevant features. From the perspective of applications in spatial data analysis, however, neurobiological realism is not necessary. In contrast, it would impose entirely unnecessary constraints. Thus, the focus in this chapter is on neural networks as efficient nonlinear models for spatial data analysis. We can not do justice to the entire spectrum of such models. Instead, attention is limited to a particular class of neural networks that have proven to be of great practical importance, the class of *feedforward neural networks*.[2]

The attractiveness of such networks is due to two features. *First*, they provide a very flexible framework to approximate arbitrary nonlinear mappings from a set of input variables to a set of output variables where the form of the mapping is governed by a number of adjustable parameters, called weights. *Second*, they are devices for nonparametric statistical inference. No particular structure or parametric form is assumed *a priori*. This is particularly useful in the case of problems where solutions require knowledge that is difficult to specify *a priori*, but for which there are sufficient observations.

The objective of this chapter is to provide an entry point and appropriate background, for those spatial analysts wishing to engage in the field of neural networks, required to fully realize its potential. The chapter is organized as follows. In section 20.2 we begin by introducing the functional form of feedforward neural network models, including the specific parameterization of the nonlinear transfer functions. Section 20.3 proceeds to discuss the problem of determining the network parameters within a framework that involves the solution of a nonlinear optimization problem. Because there is no hope of finding an analytical solution to this optimization problem, section 20.4 reviews some of the most important iterative search procedures that utilize gradient information for solving the problem. This requires the evaluation of derivatives of the objective function – known as *error function* in the machine learning literature – with respect to the network parameters, and section 20.5 shows how these can be obtained computationally efficient using the technique of *error backpropagation*.

The section that follows addresses the issue of network complexity and briefly discusses some techniques (in particular *regularization* and *early stopping*) to determine the number of hidden units. This problem is shown to essentially consist of optimizing the complexity of the network model (complexity in terms of free parameters) in order to achieve the best *generalization performance*. Section 20.7 then moves attention to the issue of how to appropriately test the generalization performance of a neural network. Some conclusions and an outlook for the future are given in the final section.

The bibliography that is included intends to provide useful pointers to the literature rather than a complete record of the whole field of neural networks. The readers should recognize that there are several wide ranging text books with introductory character, of which Hertz *et al*. (1991), Ripley (1996) and Bishop (2006) appear to be most suitable for a spatial analysis audience. Readers interested in spatial interaction or flow data analysis are referred to a paper by Fischer and Reismann (2002b) to find a useful methodology for neural spatial interaction modelling.

20.2. FEEDFORWARD NEURAL NETWORKS

Feedforward neural networks consist of nodes (also known as processing units or simply units) that are organized in layers. Figure 20.1 shows a schematic diagram of a typical feedforward neural network containing a single intermediate layer of processing units separating input from output units. Intermediate layers of this sort are often called *hidden* layers to distinguish them from the input and output layers. In this network there are N input nodes representing input variables $x_1, \ldots, x_N$; H hidden units representing hidden variables $z_1, \ldots, z_H$; and K output nodes representing output variables $y_1, \ldots, y_K$. Weight parameters are represented by links between the nodes. The bias parameters are denoted by links coming from additional input and hidden variables x_0 and z_0. Observe the feedforward structure where the inputs are connected only to units in the hidden layer, and the outputs of this layer are connected only to units in the output layer.

The term *architecture* or topology of a network refers to the topological arrangement of the nodes. We call the network architecture shown in Figure 20.1 a single hidden layer network or a two layer rather than a three layer network because it is the number of layers of adaptive weights that is important for determining the network properties. This architecture is most widely used in practice.[3]

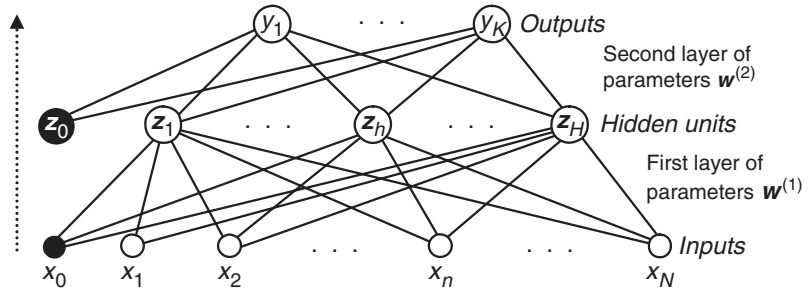


Figure 20.1 Network diagram for the single hidden layer neural network corresponding to equation (20.6). The input, hidden and output variables are represented by nodes, and the weight parameters by links between the nodes, where the bias parameters are denoted by links coming from additional input and hidden variables x_0 and z_0. The arrow denotes the direction of information flow through the network during forward propagation

Kurková (1992) has shown that one hidden layer is sufficient to approximate any continuous function uniformly on a compact input domain. But note that it may be more parsimonious to use fewer hidden units connected in two or more hidden layers.

Any network diagram can be converted into its corresponding mapping function, provided that the diagram is feedforward as in Figure 20.1 so that it does not contain closed directed cycles.[4] This guarantees that the network output y_k $(k = 1, \ldots, K)$ can be described by a series of functional transformations as follows. First, we form a linear combination[5] of the N input variables $x_1, \ldots, x_N$ to get the input, say net_h, that hidden unit h receives:

$$net_h = \sum_{n=1}^{N} w_{hn}^{(1)} x_n + w_{h0}^{(1)} \tag{20.1}$$

for $h = 1, \ldots, H$. The superscript (1) indicates that the corresponding parameters are in the first layer of the network. The parameters $w_{hn}^{(1)}$ represent connection weights going from input $n\,(n = 1, \ldots, N)$ to hidden unit $h\,(h = 1, \ldots, H)$, and $w_{h0}^{(1)}$ biases.[6] These quantities are known as activations in the field of neural networks. Each of them is then transformed using a differentiable continuous nonlinear or *activation* (transfer) function[7] φ to give the output:

$$z_h = \varphi\,(net_h) \tag{20.2}$$

for $h = 1, \ldots, H$. These quantities are again linearly combined to generate the input, called net_k, that output unit $k\,(k = 1, \ldots, K)$ receives:

$$net_k = \sum_{h=1}^{H} w_{kh}^{(2)} z_h + w_{k0}^{(2)}. \tag{20.3}$$

The parameters $w_{kh}^{(2)}$ represent the connection weights from hidden unit $h\,(h = 1, \ldots, H)$ to output unit $k\,(k = 1, \ldots, K)$, and the $w_{k0}^{(2)}$ are bias parameters. Finally, the net_k are transformed to produce a set of network outputs $y_k (k = 1, \ldots, K)$:

$$y_k = \psi_k(net_k) \tag{20.4}$$

where ψ_k denotes a real valued activation function of output unit k.

Information processing in such networks is, thus, straightforward. The input units just provide a 'fan-out' and distribute the input to the hidden units. These units sum their inputs, add a constant (the bias) and take a fixed transfer function φ_h of the result. The output units are of the same form, but with output activation function ψ_k. Network output y_k can then be expressed in terms of an output function $g_k(\boldsymbol{x}, \boldsymbol{w})$ as:

$$y_k = g_k(\boldsymbol{x}, \boldsymbol{w}) = \psi_k\left(\sum_{h=1}^{H} w_{kh}^{(2)} \varphi_h\left(\sum_{n=0}^{N} w_{hn}^{(1)} x_n + w_{h0}^{(1)}\right) + w_{k0}^{(2)}\right) \tag{20.5}$$

where $\boldsymbol{x} = (x_1, \ldots, x_N)$ and $\boldsymbol{w}$ represents a vector of all the weights and bias terms. Note that the bias terms $w_{h0}^{(1)}$ $(h = 1, \ldots, H)$ and $w_{k0}^{(2)}$ $(k = 1, \ldots, K)$ in equation (20.5) can be absorbed[8] into the set of weight parameters by defining additional input and hidden unit variables, x_0 and z_0, whose values are clamped at one so that $x_0 = 1$ and $z_0 = 1$. Then the network function (20.5) becomes

$$y_k = g_k(\boldsymbol{x}, \boldsymbol{w}) = \psi_k\left(\sum_{h=1}^{H} w_{kh}^{(2)} \varphi_h\left(\sum_{n=0}^{N} w_{hn}^{(1)} x_n\right)\right). \tag{20.6}$$

Neural networks of type (20.6) are rather general. They can be seen as a flexible way to parameterize a fairly general nonlinear function from some N-dimensional input space $\boldsymbol{X}$ to some K-dimensional output space $\boldsymbol{Y}$.

Several authors including Cybenko (1989), Funahashi (1989), Hornik *et al.* (1989) and many others have shown that such neural network models, with more or less general types of activation functions φ and ψ, have universal approximation capabilities. They can uniformly approximate any continuous function f on a compact input domain to arbitrary accuracy, provided the network has a sufficiently large number of hidden units. This approximation result, however, is non-constructive and provides no guide to how many hidden units might be needed for a practical problem at hand.

This result holds for a wide range of hidden and output layer activation functions. The functions can be any non-linearity as long as they are continuous and differentiable. The hidden unit activation functions $\varphi_h(.)$ are typically sigmoid, and almost always taken to be logistic sigmoid[9] so that:

$$\varphi_h(net_k) = \frac{1}{1 + \exp(-net_h)} \tag{20.7}$$

whose outputs lie in the range (0, 1), while the choice of the activation function $\psi_k(.)$ of the output units is generally determined by the nature of data and the assumed distribution of the target variables. Section 20.3 will show that different activation functions should be chosen for different types of problems. For standard regression problems the identity function appears to be an appropriate choice so that $y_k = net_k$. For multiple binary classification each output unit activation should be transformed using a logistic sigmoid function, while the standard multi-class classification problem in which each input is assigned to one of K mutually exclusive classes gives rise

to the softmax activation function (Bridle, 1994):

$$y_k = \psi_k(net_k) = \frac{\exp(net_k)}{\sum_{c=1}^{K} \exp(net_c)} \qquad (20.8)$$

where $0 \leq y_k \leq 1$ and $\sum_{k=1}^{K} y_k = 1$.

A neural network with a single logistic output unit can be seen as a nonlinear extension of logistic regression. With many logistic units, it corresponds to linked logistic regressions of each class versus the others. If the transfer functions of the output units in a network are taken to be linear, we have a standard linear model augmented by nonlinear terms. Given the popularity of linear models in spatial analysis, this form is particularly appealing, as it suggests that neural network models can be viewed as extensions of – rather than as alternatives to – the familiar models. The hidden unit activations can then be viewed as latent variables whose inclusion enriches the linear model.

20.3. NETWORK TRAINING

So far, we have considered neural networks as a general class of parametric nonlinear functions from a vector $\boldsymbol{x}$ of input variables $x_1, \ldots, x_N$ to a vector $\boldsymbol{y}$ of output variables $y_1, \ldots, y_K$. The process of determining the network parameters is called network training or network learning. The problem of determining the network parameters can be viewed from different perspectives. We view it as an unconstrained nonlinear function optimization problem,[10] the solution of which requires the minimization of some (continuous and differentiable) error function.

This error function, say E, is defined in term of deviations of the network outputs $\boldsymbol{y} = (y_1, \ldots, y_K)$ from corresponding desired (target) outputs $\boldsymbol{t} = (t_1, \ldots, t_K)$, and expressed as a function of the weight vector $\boldsymbol{w}$ representing the free parameters (connection weights and bias terms) of the network. The goal of training is then to minimize the error function so that:

$$\min_{\boldsymbol{w} \in \boldsymbol{W}} E(\boldsymbol{w}) \qquad (20.9)$$

where $\boldsymbol{W}$ is a weight space appropriate to the network architecture. The smallest value of $E(\boldsymbol{w})$ will occur at a point such that the gradient of the error function vanishes $\nabla E(\boldsymbol{w}) = 0$, where $\nabla E(\boldsymbol{w})$ denotes the gradient (the vector containing the partial derivatives) of $E(\boldsymbol{w})$ with respect to $\boldsymbol{w}$. A single hidden layer network of the kind shown in Figure 20.1, with H hidden units, generally has many points at which the gradient vanishes. The point $\boldsymbol{w}^*$ is called a *global* minimum for $E(\boldsymbol{w})$ if $E(\boldsymbol{w}^*) \leq E(\boldsymbol{w})$ for all $\boldsymbol{w} \in \boldsymbol{W}$. Other minima are called *local* minima, and each corresponds to a different set of parameters. For a successful application of neural networks, however, it may not be necessary to find the global minimum, and in general it will not be known whether the minimum found is the global one or not. But it may be necessary to compare several minima in order to find a sufficiently good solution of the problem under scrutiny.

Training is performed using a training set $S_p = \{(\boldsymbol{x}^p, \boldsymbol{t}^p) : p = 1, \ldots, P\}$, consisting of P ordered pairs of vectors. $\boldsymbol{x}^p$ denotes an N-dimensional input vector and $\boldsymbol{t}^p$ the associated K-dimensional desired output (target) vector. The choice of a suitable error function depends on the problem to be performed. We follow Bishop (1995: chapter 6) to provide a maximum likelihood motivation for the choice, and start by considering

regression problems. If we assume that the K target variables are independent conditional on $\boldsymbol{x}$ and $\boldsymbol{w}$ with shared noise precision α, then the conditional distribution of the target values is given by a Gaussian:

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}) = N(\boldsymbol{t}|\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{w}), \alpha^{-1}\boldsymbol{I}) \quad (20.10)$$

where α is the precision (inverse variance) of the Gaussian noise. For the conditional distribution given by equation (20.10), it is sufficient to take the output unit transfer function ψ to be the identity. Given that $\boldsymbol{t} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_P)$, are independent, identically distributed observations, we can construct the corresponding likelihood function:

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \alpha) = \prod_{p=1}^{P} p(\boldsymbol{t}_p|\boldsymbol{x}^p, \boldsymbol{w}, \alpha). \quad (20.11)$$

Maximizing the likelihood function is equivalent to minimizing the *sum-of-squares function* given by:

$$E(\boldsymbol{w}) = \tfrac{1}{2}\sum_{p=1}^{P}\sum_{k=1}^{K}\left\|g_k(\boldsymbol{x}^p, \boldsymbol{w}) - t_k^p\right\|^2. \quad (20.12)$$

The value of $\boldsymbol{w}$ found by solving equation (20.9) will be denoted $\boldsymbol{w}_{ML}$ because it corresponds to the maximum likelihood estimation. Having formed $\boldsymbol{w}_{ML}$, the noise precision is then provided by:

$$\frac{1}{\alpha_{ML}} = \tfrac{1}{PK}\sum_{p=1}^{P}\left\|\boldsymbol{g}(\boldsymbol{x}^p, \boldsymbol{w}_{ML}) - \boldsymbol{t}^p\right\|^2. \quad (20.13)$$

The assumption of independence can be dropped at the expense of a slightly more complex optimization problem. Note that in practice the nonlinearity of the network function $\boldsymbol{g}(\boldsymbol{x}^p, \boldsymbol{w})$ causes the error $E(\boldsymbol{w})$ to be convex, and so in practice local maxima of the likelihood may be found, which correspond to local minima of the error function.

There is a natural pairing of the error function given by the negative log likelihood and the output unit transfer function. In the regression case we can view the network as having a transfer function ψ that is the identity, so that $y_k = net_k$. The corresponding sum-of-squares error function then has the characteristic:

$$\frac{\partial E}{\partial net_k} = (y_k - t_k). \quad (20.14)$$

This property will be used when discussing the technique of error backpropagation in section 20.5.

Now let us consider the case of binary classification where we have a single target variable t such that $t = 1$ denotes class C_1 and $t = 0$ class C_2. We consider a network with a single output whose transfer function is a logistic sigmoid so that $0 \leq g(\boldsymbol{x}, \boldsymbol{w}) \leq 1$, and we can interpret $g(\boldsymbol{x}, \boldsymbol{w})$ as the conditional probability $p(C_1, \boldsymbol{x})$, with $p(C_2, \boldsymbol{x})$ given by $1 - g(\boldsymbol{x}, \boldsymbol{w})$. The conditional probability of targets given inputs is then a Bernoulli distribution of the form:

$$p(t\,|\,\boldsymbol{x}, \boldsymbol{w}) = g(\boldsymbol{x}, \boldsymbol{w})^t\{1 - g(\boldsymbol{x}, \boldsymbol{w})\}^{1-t}. \quad (20.15)$$

If we have a training set of independent observations, then the error function, given

by the negative log likelihood, is the cross-entropy error function of the form:

$$E(\boldsymbol{w}) = -\sum_{p=1}^{p} \{t^p \ln y^p + (1-t^p)\ln(1-y^p)\} \tag{20.16}$$

where y^p denotes $g(\boldsymbol{x}^p, \boldsymbol{w})$. Note there is no analogue of the noise precision α because the target values are assumed to be correctly labelled.

For classification problems, the targets represent labels defining class membership or – more generally – estimates of the probabilities of class membership. If we have K separate binary classifications to perform, then a neural network with K logistic sigmoid output units is an appropriate choice. In this case a binary class label $t_k^p \in \{0, 1\}$ is associated with each output k. If we assume that the class labels are independent, given the input vector $\boldsymbol{x}^p$, then the conditional distribution is:

$$p(\boldsymbol{t} \,|\, \boldsymbol{x}, \boldsymbol{w}) = \prod_{k=1}^{K} g_k(\boldsymbol{x}, \boldsymbol{w})^{t_k} [1-g_k(\boldsymbol{x}, \boldsymbol{w})]^{1-t_k}. \tag{20.17}$$

Taking the negative logarithm of the corresponding likelihood function then yields the multiple-class cross-entropy error function of the form:

$$E(\boldsymbol{w}) = -\sum_{p=1}^{P}\sum_{k=1}^{K} \left\{t_k^p \ln y_k^p + (1 - t_k^p)\ln(1 - y_k^p)\right\} \tag{20.18}$$

where $y_k^p = g_k(\boldsymbol{x}^p, \boldsymbol{w})$. It is important to note that the derivative of this error function with respect to the activation for a particular output unit k takes the simple form (20.14) as in the regression case.

If we have a standard multiple-class classification problem to solve, where each input is assigned to one of K mutually exclusive classes, then we can use a neural network with K output units each of which has a softmax output activation function. The binary target variables $t_k \in \{0, 1\}$ have a 1-of-K coding scheme indicating the correct class, and the network outputs are interpreted as $g_k(\boldsymbol{x}^p, \boldsymbol{w}) = p(t_k^p = 1 \,|\, \boldsymbol{x})$ leading to the error function, called the multiple-class cross-entropy error function (see Fischer and Staufer, 1999):

$$E(\boldsymbol{w}) = -\sum_{p=1}^{P}\sum_{k=1}^{K} t_k^p \ln\left(\frac{g_k(\boldsymbol{x}^p, \boldsymbol{w})}{t_k^p}\right) \tag{20.19}$$

which is non-negative, and equals zero when $g_k(\boldsymbol{x}^p, \boldsymbol{w}) = t_k^p$ for all k and p. Once again, the derivative of this error function with respect to the activation for a particular output unit k takes the familiar form equation (20.14). It is worth noting that in the case of $K = 2$ we can use a network with a single logistic sigmoid output, alternatively to a network with two softmax output activations.

In summary, there is natural pairing of the choice of the output unit transfer function and the choice of the error function, according to the type of the problem that has to be solved. For regression we take linear outputs and a sum-of-squares error, for (multiple independent) binary classifications we use logistic sigmoid outputs with the corresponding cross-entropy error function, and for multi-class classification softmax outputs and the multi-class cross-entropy error function. For classification problems involving two classes, we can use a single logistic sigmoid output, or alternatively we

can take a network with two softmax outputs (Bishop, 2006: 236).

20.4. PARAMETER OPTIMIZATION

There are many ways to solve the minimization problem (20.9). Closed-form optimization via the calculus of scalar fields rarely admits a direct solution. A relatively new set of interesting techniques that use optimality conditions from calculus are based on evolutionary computation (Goldberg, 1989; Fogel, 1995). But gradient procedures which use the first partial derivatives $\nabla E(\boldsymbol{w})$, so-called first order strategies, are most widely used. Gradient search for solutions gleans its information about derivatives from a sequence of function values. The recursion scheme is based on the formula:[11]

$$\boldsymbol{w}(\tau + 1) = \boldsymbol{w}(\tau) + \eta(\tau)\boldsymbol{d}(\tau) \qquad (20.20)$$

where τ denotes the iteration step. Different procedures differ from each other with regard to the choice of step length $\eta(\tau)$ and search direction $\boldsymbol{d}(\tau)$, the former being a scalar called *learning* rate and the latter a vector of unit length.

The simplest approach to using gradient information is to assume $\eta(\tau)$ being constant and to choose the parameter update in equation (20.20) to comprise a small step in the direction of the negative gradient so that:

$$\boldsymbol{d}(\tau) = -\nabla E(\boldsymbol{w}(\tau)). \qquad (20.21)$$

After each such update, the gradient is re-evaluated for the new parameter vector $\boldsymbol{w}(\tau + 1)$. Note that the error function is defined with respect to a training set S_P to be processed to evaluate ∇E. One complete presentation of the entire training set during the training process is called an *epoch*. The training process is maintained on an epoch-by-epoch basis until the connection weights and bias terms of the network stabilize and the average error over the entire training set converges to some minimum. It is good practice to randomize the order of presentation of training examples from one epoch to the next. This randomization tends to make the search in the parameter space stochastic over the training cycles, thus avoiding the possibility of limit cycles in the evolution of the weight vectors.

Gradient descent optimization may proceed in one of two ways: pattern mode and batch mode. In the *pattern mode* weight updating is performed after the presentation of each training example. Note that the error functions based on maximum likelihood for a set of independent observations comprise a sum of terms, one for each data point. Thus:

$$E(\boldsymbol{w}) = \sum_{p=1}^{P} E_p(\boldsymbol{w}) \qquad (20.22)$$

where E_p is called the *local error* while E the global error, and pattern mode gradient descent makes an update to the parameter vector based on one training example at a time so that:

$$\boldsymbol{w}(\tau + 1) = \boldsymbol{w}(\tau) - \eta\nabla E_p(\boldsymbol{w}(\tau)). \qquad (20.23)$$

Rumelhart *et al.* (1986) have shown that pattern based gradient descent minimizes equation (20.22), if the learning parameter η is sufficiently small. The smaller η, the smaller will be the changes to the weights in the network from one iteration to the next and the smoother will be the trajectory in the parameter space. This improvement, however, is attained at the cost of a slower rate of training. If we make the learning rate parameter η too large so as to speed up the

rate of training, the resulting large changes in the parameter weights assume such a form that the network may become unstable.

In the *batch mode* of training, parameter updating is performed after the presentation of all the training examples that constitute an epoch. From an online operational point of view, the pattern mode of training is preferred over the batch mode, because it requires less local storage for each weight connection. Moreover, given that the training patterns are presented to the network in a random manner, the use of pattern-by-pattern updating of parameters makes the search in parameter space stochastic in nature[12] which in turn makes it less likely to be trapped in a local minimum. On the other hand, the use of batch mode of training provides a more accurate estimation of the gradient vector ∇E. Finally, the relative effectiveness of the two training modes depends on the problem to be solved (Haykin, 1994: 152 pp).

For batch optimization there are more efficient procedures, such as conjugate gradients and quasi-Newton methods, that are much more robust and much faster than gradient descent (Nocedal and Wright, 1999). Unlike steepest gradient, these algorithms have the characteristic that the error function always decreases at each iteration unless the parameter vector has arrived at a local or global minimum. Conjugate gradient methods achieve this by incorporating an intricate relationship between the direction and gradient vectors. The initial direction vector $\boldsymbol{d}(0)$ is set equal to the negative gradient vector at the initial step $\tau = 0$. Each successive direction vector is then computed as a linear combination of the current gradient vector and the previous direction vector. Thus:

$$\boldsymbol{d}(\tau + 1) = -\nabla E(\boldsymbol{w}(\tau + 1)) + \beta(\tau)\boldsymbol{d}(\tau) \tag{20.24}$$

where $\beta(\tau)$ is a time varying parameter. There are various rules for determining $\beta(\tau)$ in terms of the gradient vectors at time τ and $\tau + 1$ leading to the Fletcher–Reeves and Polak–Ribière variants of conjugate gradient algorithms (see Press *et al*., 1992). The computation of the learning rate parameter $\eta(\tau)$ in the update formula (20.20) involves a line search, the purpose of which is to find a particular value of η for which the error function $E(\boldsymbol{w}(\tau)+\eta\,\boldsymbol{d}(\tau))$ is minimized, given fixed values of $\boldsymbol{w}(\tau)$ and $\boldsymbol{d}(\tau)$.

The application of Newton's method to the training of neural networks is hindered by the requirement of having to calculate the Hessian matrix and its inverse, which can be computationally expensive. The problem is further complicated by the fact that the Hessian matrix $\boldsymbol{H}$ would have to be non-singular for its inverse to be computed. Quasi-Newton methods avoid this problem by building up an approximation to the inverse Hessian over a number of iteration steps. The most commonly variants are the Davidson–Fletcher–Powell and the Broyden–Fletcher–Goldfarb–Shanno procedures (see Press *et al*., 1992).

Quasi-Newton procedures are today the most efficient and sophisticated (batch) optimization algorithms. But they require the evaluation and storage in memory of a dense matrix $\boldsymbol{H}(\tau)$ at each iteration step τ. For larger problems (more than 1,000 weights) the storage of the approximate Hessian can be too demanding. In contrast, the conjugate gradient procedures require much less storage, but an exact determination of the learning rate $\eta(\tau)$ and the parameters $\beta(\tau)$ in each iteration τ, and, thus, approximately twice as many gradient evaluations as the quasi-Newton methods.

When the surface modelled by the error function in its parameter space is extremely rugged and has many local minima, then a local search from a random starting point

tends to converge to a local minimum close to the initial point and to a solution worse than the global minimum. In order to seek out good local minima, a good training procedure must thus include both a gradient based optimization algorithm and a technique like random start that enables sampling of the space of minima. Alternatively, stochastic global search procedures might be used. Examples of such procedures include Alopex (see Fischer *et al*., 2003, for an application in the context of spatial interaction data analysis), genetic algorithms (see Fischer and Leung, 1998, for another application in the same context), and simulated annealing. These procedures guarantee convergence to a global solution with high probability, but at the expense of slower convergence.

Finally, it is worth noting that the question whether neural networks can have real-time learning capabilities is still challenging and open. Real-time learning is highly required by time-critical applications, such as for navigation and tracking systems in a GIS-T context, where the data observations are arriving in a continuous stream, and predictions have to be made before all the data seen. Even for offline applications, speed is still a need, and real-time learning algorithms that reduce training time are of considerable value.

20.5. ERROR BACKPROPAGATION

One of the greatest breakthroughs in neural network modelling has been the introduction of the technique of error backpropagation[13] in that it provides a computationally efficient technique to calculate the gradient vector of an error function for a feedforward neural network with respect to the parameters. This technique – sometimes simply termed backprop – uses a local message passing scheme in which information is sent alternately forwards and backwards through the network. Its modern form stems from Rumelhart *et al*. (1986), illustrated for gradient descent optimization applied to the sum-of-squares error function. It is important to recognize, however, that error backpropagation can also be applied to error functions other than just sum-of-squares and to a wide variety of optimization schemes for weight adjustment other than gradient descent, in pattern or batch mode.

We describe the backpropagation algorithm for a general network of type (20.6) that has a single hidden layer, arbitrary differentiable activation functions with a corresponding local error function $E_p(\boldsymbol{w})$. For each pattern p in the training data set, we shall assume that we have supplied the corresponding input vector $\boldsymbol{x}^p$ to the network and calculated the activations of all the hidden and output units in the network by applying equations (20.1)–(20.4). Recall that each hidden unit h has input net_h^p and output $z_h^p = \varphi_h(net_h^p)$, and each output unit k has input net_k^p and output $y_k^p = \psi_k(net_k^p)$. This process is called forward propagation because it can be seen as a forward flow of information (signals) provided by $\boldsymbol{x}^p$ through the network. For the rest of this section we consider one example and drop the superscript p in order to keep the notation uncluttered.

We evaluate the gradient E_p with respect to a hidden-to-output parameter $w_{kh}^{(2)}$ first, by noting that E_p depends on the weight $w_{kh}^{(2)}$ only via the summed input, net_k, to the output unit k. Thus, we can apply the chain rule for partial derivatives to get:

$$\frac{\partial E_p}{\partial w_{kh}^{(2)}} = \frac{\partial E_p}{\partial net_k} \frac{\partial net_k}{\partial w_{kh}^{(2)}} \tag{20.25}$$

where

$$\frac{\partial net_k}{\partial w_{kh}^{(2)}} = \frac{\partial}{\partial w_{kh}^{(2)}} \sum_{h=0}^{H} w_{kh}^{(2)} z_h = z_h. \quad (20.26)$$

If we define:

$$\delta_k := \frac{\partial E_p}{\partial net_k} = \psi'(net_k) \frac{\partial E_p}{\partial y_k} \quad (20.27)$$

where the δs are often referred to as *errors*, and substitute equations (20.26) and (20.27) into equation (20.25), we obtain:

$$\frac{\partial E_p}{\partial w_{kh}^{(2)}} = \delta_k z_h. \quad (20.28)$$

This equation tells us that the required partial derivative with respect to $w_{kh}^{(2)}$ is obtained simply by the multiplication of two expressions: the value of δ for unit k at the output end of the connection concerned and the value of z at the input end h of the connection. Thus, in order to evaluate the partial derivatives with respect to the second layer parameters we need only to compute the value of δ_k for each output unit $k = 1, \ldots, K$ in the network, and then apply equation (20.28).

For linear outputs associated with the sum-of-squares error function, for logistic sigmoid outputs associated with the cross-entropy error function and for softmax outputs associated with the multiple-class cross-entropy error function, the δs are given by:

$$\delta_k = y_k - t_k \quad (20.29)$$

while for logistic sigmoid outputs associated with the sum-of-squares error function the δs are found as:

$$\delta_k = y_k(1 - y_k)(y_k - t_k). \quad (20.30)$$

For the input-to-hidden connections we must differentiate the chosen error function with respect to the parameters $w_{hn}^{(1)}$, which are more deeply embedded in the error function. Using again the chain rule for partial derivatives, we get:

$$\frac{\partial E_p}{\partial w_{hn}^{(1)}} = \delta_h \frac{\partial net_h}{\partial w_{hn}^{(1)}} = \delta_h x_n \quad (20.31)$$

with:

$$\delta_h := \varphi'_h(net_h) \sum_{k=1}^{K} \delta_k w_{kh}^{(2)} \quad (20.32)$$

where the use of the prime signifies differentiation with respect to the argument. In the case of logistic hidden units we get the following backpropagation formula:

$$\begin{aligned} \delta_h &= \varphi'_h(net_h) \sum_{k=1}^{K} \delta_k w_{kh}^{(2)} \\ &= \varphi(net_h)(1 - \varphi(net_h)) \sum_{k=1}^{K} \delta_k w_{kh}^{(2)} \\ &= z_h(1 - z_h) \sum_{k=1}^{K} \delta_k w_{kh}^{(2)}. \end{aligned} \quad (20.33)$$

Since the formula for δ_h contains only terms in a later layer, it is clear that it can be calculated from output to input on the network. Thus, the basic idea behind the technique of error backpropagation is to use a

forward pass through the network to calculate the z_h and y_k values by propagating the input vector, followed by a backward pass to calculate δ_k and δ_h, and hence the partial derivatives of the error function. Note that for the presentation of each training example the input pattern is fixed throughout the message passing scheme, encompassing the forward pass followed by the backward pass.

The backpropagation technique can be summarized in the following four steps:

Step 1 Apply an input vector $\boldsymbol{x}^p$ to the network and forward propagate through the network, using equations (20.1)–(20.4), to generate the hidden and output unit activations based on current weight settings.

Step 2 Evaluate the δ_k for all the output units ($k = 1, \ldots, K$) using equation (20.29) or equation (20.30), depending on the problem type to be studied.

Step 3 Backpropagate the deltas, using equation (20.33), to get δ_h for each hidden unit $h (h = 1, \ldots, H)$ in the network.

Step 4 Use equations (20.28) and (20.31) to evaluate the required derivatives.

For batch procedures the gradient of the global error can be obtained by repeating *Step 1* to *Step 4* for each pattern p in the training set, and then summing over all patterns.

20.6. NETWORK COMPLEXITY

So far we have considered neural networks of type (20.6) with *a priori* given numbers of input, hidden and output units. While the number of input and output units in a neural network is basically problem dependent, the number H of hidden units is a free parameter that can be adjusted to provide the best testing performance on independent data, called testing set. But the testing error is not a simple function of H due to the presence of local minima in the error function. The issue of finding a parsimonious model for a real world problem is critical for all models but particularly important for neural networks because the problem of overfitting is more likely to occur.

A neural network model that is too simple (i.e., small H), or too inflexible, will have a large bias and smooth out some of the underlying structure in the data (corresponding to high bias), while one that has too much flexibility in relation to the particular data set will overfit the data and have a large variance. In either case, the performance of the network on new data (i.e., generalization performance) will be poor. This highlights the need to optimize the complexity in the model selection process in order to achieve the best generalization (Bishop, 1995: 332; Fischer, 2000). There are some ways to control the complexity of a neural network, complexity in terms of the number of hidden units or, more precisely, in terms of the independently adjusted parameters. Practice in spatial data analysis generally adopts a trial and error approach that trains a sequence of neural networks with an increasing number of hidden units and then selects that one which gives the predictive performance on a testing set.[14]

There are, however, other more principled ways to control the complexity of a neural network model in order to avoid overfitting.[15] One approach is that of *regularization*, which involves adding a regularization term $R(\boldsymbol{w})$ to the error function in order to control overfitting, so that the total error function to be minimized takes the form:

$$\tilde{E}(\boldsymbol{w}) = E(\boldsymbol{w}) + \mu R(\boldsymbol{w}) \qquad (20.34)$$

where μ is a positive real number, the so-called regularization parameter, that controls the relative importance of the data dependent error $E(\boldsymbol{w})$ and the regularization term $R(\boldsymbol{w})$,

sometimes also called complexity term. This term embodies the *a priori* knowledge about the solution, and therefore depends on the nature of the particular problem to be solved. Note that $\tilde{E}_p(\boldsymbol{w})$ is called the *regularized error function*.

One of the simplest forms of regularizer is defined as the squared norm of the parameter vector $\boldsymbol{w}$ in the network, as given by:

$$R(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 . \tag{20.35}$$

This regularizer[16] is known as a weight decay function that penalizes large weights. Hinton (1987) has found empirically that a regularizer of this form can lead to significant improvements in network generalization.

Sometimes, a more general regularizer is used, for which the regularized error takes the form:

$$E(\boldsymbol{w}) + \mu \|\boldsymbol{w}\|^m \tag{20.36}$$

where $m = 2$ corresponds to the quadratic regularizer (20.35). The case $m = 1$ is known as the 'lasso' in the statistics literature (Tibshirani, 1996b). It has the property that – if μ is sufficiently large – some of the parameter weights are driven to zero in sequential learning algorithms, leading to a sparse model. As μ is increased, so an increasing number of parameters are driven to zero.

One of the limitations of this regularizer is inconsistency with certain scaling characteristics of network mappings. If one trains a network using original data and one network using data for which the input and/or target variables are linearly transformed, then consistency requires obtaining equivalent networks which differ only by a linear transformation of the weights. Any regularizer should possess this characteristic, otherwise one solution is arbitrary favoured over an equivalent solution. In particular, the weights should be scale invariant (Bishop, 2006: 257–258). A regularized error function that satisfies this property is given by:

$$E(\boldsymbol{w}) + \mu_1 \left\|\boldsymbol{w}_{q1}\right\|^m + \mu_2 \left\|\boldsymbol{w}_{q2}\right\|^m \tag{20.37}$$

where $\boldsymbol{w}_{q1}$ denotes the set of the weights in the first layer, that is $w_{11}^{(1)}, \ldots, w_{h1}^{(1)}, \ldots, w_{HN}^{(1)}$, and $\boldsymbol{w}_{q2}$ those in the second layer, that is $w_{11}^{(2)}, \ldots, w_{kh}^{(2)}, \ldots, w_{KH}^{(2)}$. Under linear transformations of the weights, the regularizer will remain unchanged, provided that the parameters μ_1 and μ_2 are suitably rescaled.

The more sophisticated control of complexity that regularization offers over adjusting the number of hidden units by trial and error is evident. Regularization allows complex neural network models to be trained on data sets of limited size without severe overfitting, by limiting the effective network complexity. The problem of determining the appropriate number of hidden units is, thus, shifted to one of determining a suitable value for the regularization parameter(s) during the training process.

The principal alternative to regularization as a way to optimize the model complexity for a given training data set is the procedure of *early stopping*. As we have seen in the previous sections, training of a nonlinear network model corresponds to an iterative reduction of the error function defined with respect to a given training data set. For many of the optimization procedures used for network training (such as conjugate gradient optimization) the error is a nondecreasing function of the iteration steps τ. But the error measured with respect to independent data, called the *validation data set*, often shows a decrease first, followed by an

increase as the network starts to overfit, as illustrated in Fischer and Gopal (1994) for a spatial interaction data analysis problem. Thus, training can be stopped at the point of smallest error with respect to the validation data, in order to get a network that shows good generalization performance. But, if the validation set is small, it will give a relatively noisy estimate of generalization performance, and it may be necessary to keep aside another data set, the test set, on which the performance of the network model is finally evaluated.

This approach of stopping training before a minimum of the training error has been reached is another way of eliminating the network complexity. It contrasts with regularization because the determination of the number of hidden units does not require convergence of the training process. The training process is used here to perform a directed search of the weight space for a neural network model that does not overfit the data and, thus, shows superior generalization performance. Various theoretical and empirical results have provided strong evidence for the efficiency of early stopping (see, e.g., Weigend *et al*., 1991; Baldi and Chauvin, 1991; Finnoff, 1991). Although many questions remain, a picture is starting to emerge as to the mechanisms responsible for the effectiveness of this procedure. In particular, it has been shown that stopped training has the same sort of regularization effect (i.e., reducing model variance at the cost of bias) that penalty terms provide.

20.7. GENERALIZATION PERFORMANCE

The ability of a neural network to predict correctly new observations that differ from those used for training is known as *generalization* (see, e.g., Moody, 1992). To assess the generalization performance of a neural network model is of crucial importance. The performance on the training set is not a good indicator due to the problem of overfitting. As often in statistics, there is a trade-off between accuracy on the training data and generalization. This is a well-studied dilemma (see, e.g., Bishop, 1995: chapter 9).

The simplest way to assess the generalization performance is the use of a test set. Here, of course, it is assumed that the test data are drawn from the same population used to generate the training data. If the test set is too small, an accurate assessment cannot be obtained. Test set validation becomes practical only if the data sets are very large or new data can be generated cheaply. As the training and test sets are independent samples, an unbiased estimate of the prediction risk is obtained. But the estimate can be highly variable across different data splittings.

One way to overcome this problem is by *cross-validation*. Cross-validation is a sample re-use method for assessing generalization performance. It makes maximally efficient use of the available data. The idea is to divide the available data set into – generally equally sized – D parts, and then to use one part to test the performance of the neural network model trained on the remaining $(D-1)$ parts. The resulting estimator is again unbiased, and we can average the D such estimates. Leave-one-out cross-validation is a special case, in which each observation is tested on the remaining $(P-1)$ observations. This version evidently requires a large number of computations. Choosing $D = P$ should give the most accurate assessment, as the 'true' size of the training set is most closely mimicked, but also involves the most computation. In addition, cross-validation estimates of performance for large

D might be expected to be rather variable. Taking a smaller D can give a larger bias, but smaller variance and mean-square error. This is an argument in favour of smaller D (Ripley, 1996: 70–71). Bootstrap estimates of bias can be used for bias correction (Efron, 1982).

With small samples of data – precisely when structural uncertainty is greatest – cross-validation may not be feasible, because there are too few data values with which to carry out the estimation, validation and testing activities in a stable way. *Bootstrapping* the neural network modelling process – creating bootstrap copies of the available data to generate copies of training, validation and test sets – may be used instead as a general framework for evaluating generalization performance. The idea underlying the bootstrap is appealingly simple. For an introduction see, for example, Efron (1982), Efron and Tibshirani (1993) or Hastie *et al.* (2001).

Suppose, we are interested in a single hidden layer neural network together with a sum-of-squares error function to solve a regression problem. The standard procedure for estimating and evaluating the neural network is to split the available data set, say $S_P = \{z^p = (\boldsymbol{x}^p, \boldsymbol{t}^p) : p = 1, \ldots, P\}$, into three parts: a training set $S_{P1} = \{z^{p1} = (\boldsymbol{x}^{p1}, \boldsymbol{t}^{p1}) : p1 = 1, \ldots, P1\}$, a validation set $S_{P2} = \{z^{p2} = (\boldsymbol{x}^{p2}, \boldsymbol{t}^{p2}) : p2 = 1, \ldots, P2\}$ and a test set $S_{P3} = \{z^{p3} = (\boldsymbol{x}^{p3}, \boldsymbol{t}^{p3}) : p3 = 1, \ldots, P3\}$, with $P1 + P2 + P3 = P$. The training set serves for parameter estimation such as by means of gradient descent on the sum-of-squares error function. The validation set is used, for example, to determine the stopping point before overfitting occurs, and the test set to evaluate the generalization performance of the model, using some measure of error between a prediction and an observed value, such as the familiar root-mean-square error ρ of the form:

$$\hat{\rho}(\hat{\boldsymbol{w}}) = \frac{\sum_{p3=1}^{P3} \left\| \boldsymbol{g}(\boldsymbol{x}^{p3}, \hat{\boldsymbol{w}}) - \boldsymbol{t}^{p3} \right\|^2}{\sum_{p3=1}^{P3} \left\| \boldsymbol{t}^{p3} - \bar{\boldsymbol{t}} \right\|^2} \quad (20.38)$$

which is a function of $\hat{\boldsymbol{w}}$, obtained by solving the minimization problem (20.9). $\bar{\boldsymbol{t}}$ is defined to be the average test set target vector. Care has to be taken in interpreting the results obtained as accurate estimates of the generalization performance.

Randomness enters into this standard approach to neural network modelling in two ways: in the splitting of the data samples, and in choices about the parameter initialization. This leaves one question wide open. What is the variation of test performance as one varies training, validation, and test sets? This is an important question, since there is not just one 'best' split of the data or obvious choice for the initial weights. Thus, it is useful to vary both the data partitions and parameter initializations to find out more about the distributions of generalization errors. One way is to use a computer intensive bootstrapping approach to evaluate the performance, reliability, and robustness of the neural network model, an approach that combines the purity of splitting the data into three disjoint data sets with the power of a resampling procedure. Implementing this approach involves the following steps (see Fischer and Reismann, 2002a, b, for an application in the context of spatial interaction modelling).

Step 1: Generation of bootstrap training, validation and test sets

Using the sample S_P, we first build a test set by choosing $P3$ patterns randomly,[17] with replacement. The patterns used in this specific test set are then removed from the pool S_P. From the remainder,

we then randomly set aside $P2$ patterns for the bootstrap validation set. They are picked randomly without replacement and removed from the pool. The remaining patterns constitute the training set. This process is repeated B times (typically $20 < B < 200$) to generate $b = 1, \ldots, B$ training data sets of size $P1$, $S_{P1}^{*b} = \{{}^{*b}\boldsymbol{z}^{p1} : p1 = 1, \ldots, P1\}$, called bootstrap training sets; $b = 1, \ldots, B$ validation data sets of size $P2$, $S_{P2}^{*b} = \{{}^{*b}\boldsymbol{z}^{p2} : p2 = 1, \ldots, P2\}$, called bootstrap validation sets; and $b = 1, \ldots, B$ test data sets of size $P3$, $S_{P3}^{*b} = \{{}^{*b}\boldsymbol{z}^{p3} : p3 = 1, \ldots, P3\}$, called bootstrap test sets.

Step 2: Computation of the bootstrap parameter estimates

Each bootstrap training set S_{P1}^{*b} is used to compute a new parameter vector by minimizing:

$$\arg\min\{E({}^{*b}\boldsymbol{w}) : {}^{*b}\boldsymbol{w} \in \boldsymbol{W}, \boldsymbol{W} \subseteq R^{Q}\} \tag{20.39}$$

where Q is the number of parameters, and $E({}^{*b}\boldsymbol{w})$ the (global) sum-of-squares error for the bth bootstrap training sample. This is given by:

$$E({}^{*b}\boldsymbol{w}) = \frac{1}{2}\sum_{p1=1}^{P1}\left\|\boldsymbol{g}({}^{*b}\boldsymbol{x}^{p1}, {}^{*b}\boldsymbol{w}) - {}^{*b}\boldsymbol{t}^{p1}\right\|^2 \tag{20.40}$$

where the sum runs over the bth bootstrap training set, and $b = 1, \ldots, B$. The corresponding bootstrap validation set is used to determine the stopping point before overfitting occurs and/or to set additional parameters or hyperparameters. This yields B bootstrap parameter estimates ${}^{*b}\hat{\boldsymbol{w}}$(b=1,. . ., B).

Step 3: Estimation of the bootstrap statistic of interest

From S_{P3}^{*b} calculate ${}^{*b}\hat{\rho}({}^{*b}\hat{\boldsymbol{w}})$, the bootstrap analogue of $\hat{\rho}(\hat{\boldsymbol{w}})$ given by equation (20.38), in the same manner as $\hat{\rho}(\hat{\boldsymbol{w}})$ but with resample S_{P3}^{*b} replacing S_{P3} and ${}^{*b}\hat{\boldsymbol{w}}$ replacing $\hat{\boldsymbol{w}}$. This yields a sequence of bootstrap statistics, ${}^{*1}\hat{\rho}, \ldots, {}^{*B}\hat{\rho}$.

Step 4: Estimation of the standard deviation

The statistical accuracy of the performance statistic can then be evaluated by looking at the variability of the statistic between the different bootstrap test sets. Estimate the standard deviation, $\hat{\sigma}$, of ${}^{*}\hat{\rho}$ as approximated by bootstrap:

$$\hat{\sigma}_{P3}^{B} = \left\{\frac{1}{B-1}\sum_{b=1}^{B}\left[{}^{*b}\hat{\rho}({}^{*b}\boldsymbol{w}) - {}^{*}\hat{\rho}(.)\right]^2\right\}^{1/2} \tag{20.41}$$

where

$${}^{*}\hat{\rho}(.) = \frac{1}{B}\sum_{b=1}^{B}{}^{*b}\hat{\rho}({}^{*b}\boldsymbol{w}). \tag{20.42}$$

The true standard error of $\hat{\rho}$ is a function of the unknown density function F of ρ, that is $\sigma(F)$. With the bootstrapping approach described above one obtains $\hat{F}_{P3}^{*}$ which is supposed to describe closely the empirical probability distribution $\hat{F}_{P3}$, in other words $\hat{\sigma}_{P3}^{B} \approx \sigma(\hat{F}_{P3})$. Asymptotically, this means that as $P3$ tends to infinity, the estimate $\hat{\sigma}_{P3}^{B}$ tends to $\sigma(F)$. For finite sample sizes, however, generally there will be deviations.

Step 5: Bias estimation

The bootstrap scheme can be used to estimate not only the variability of the performance statistic $\hat{\rho}$, but also its bias (Zapranis and Refenes, 1998). Bias can be thought of as a function of the unknown probability density function F of ρ that is $\beta = \beta(F)$. The bootstrap estimate of bias is simply:

$$\hat{\beta}_B = \beta(\hat{F}_{P3}) = E^{*}[\rho(\hat{F}_{P3}^{*}) - \rho(\hat{F}_{P3})] \tag{20.43}$$

where E^{*} indicates expectation with respect to bootstrap sampling and $\hat{F}_{P3}^{*}$

the bootstrap empirical distribution. The bootstrap estimate of bias is:

$$\hat{\beta} = \frac{1}{B} \sum_{b=1}^{B} \left[{}^{*b}\hat{\rho}\,({}^{*b}\boldsymbol{w}) - \hat{\rho}\,(\boldsymbol{w}) \right]. \tag{20.44}$$

The bias is removed by subtracting $\hat{\beta}_B$ from the estimated $\hat{\rho}$.

20.8. SUMMARY AND OUTLOOK

In one sense neural networks are nonlinear models having a methodology of their own. From a spatial analysis point of view neural networks can generally be used anywhere one would ordinarily use a linear or nonlinear specification, with estimation proceeding via appropriate techniques. The now rather well-developed theory of estimation of misspecified models applies immediately to provide appropriate interpretations and inferential procedures.

Neural networks have essentially a broader utility that has yet to be fully appreciated by spatial analysts, but which has the potential to significantly enhance scientific understanding of spatial phenomena and spatial processes subject to neural network modelling. In particular, the estimates obtained from neural network learning may serve as a basis for formal statistical inference, making possible statistical tests of specific hypotheses of interest. Because of the ability of neural networks to extract complex nonlinear effects, the alternatives against which such tests can have power may extend usefully beyond those with the reach of more conventional methods, such as linear models for regression and classification.

Although we have covered a fair amount of ground in this chapter we have only scratched the surface of the modelling possibilities offered by neural networks. To mention some additional models treated in the field of neural networks, we note that competitive learning networks have been much studied with applications, for example, to the travelling salesman problem and remote sensing classification problems, and that radial basis function networks in which the activation for a hidden unit is determined by the distance between the input vector and a prototype vector are also standard objects of investigation. Leung (1997), for example, illustrates the use of radial basis function networks for rule learning. We, moreover, did not consider neural networks for unsupervised feature discovery which in statistical terms correspond to cluster analysis and/or latent structure analysis.

For neural networks to find a place in spatial data analysis they need to overcome their current limitations, mainly due to the relative absence of established procedures for model identification, comparable to those for spatial econometric modelling techniques. In particular, providing tests specifically designed to test the adequacy of neural models is a research issue on its own right. Despite significant improvements in our understanding of the fundamentals of neural network modelling, there are many open problems and directions for future research. From a spatial analytic perspective an important avenue for further investigation is the incorporation of spatial dependency in the network representation that received less attention in the past than it deserves. Another is the application of Bayesian inference techniques to neural networks. A Bayesian approach would provide an alternative framework for dealing with the issues of network complexity and would avoid many of the problems discussed in this chapter. In particular, error bars and confidence intervals can easily be assigned to the predictors generated by neural networks, without the need of bootstrapping.

NOTES

1 Neural networks can model cortical local learning and signal processing, but they are not the brain, neither are many special purpose systems to which they contribute (Weng and Hwang, 2006).

2 Feedforward neural networks are sometimes also called multilayer perceptrons even though the term perceptron is usually used to refer to a network with linear threshold gates rather than with continuous nonlinearities. Radial basis function networks, recurrent networks rooted in statistical physics, self-organizing systems and ART (Adaptive Resonance Theory) models are other important classes. For a fuzzy ARTMAP multispectral classification see Gopal and Fischer (1997).

3 A generalization of this network architecture is to allow skip-layer connections from input to output, each of which is associated with a corresponding adaptive parameter. But note that a network with sigmoidal hidden units can always mimic skip-layer connections for bounded input values by using sufficiently small single hidden layer weights. Skip-layer connections, however, can be easier to implement and interpret in practice.

4 Networks with closed directed cycles are called *recurrent* networks. There are three types of such networks: *first*, networks in which the input layer is fed back into the input layer itself; *second*, networks in which the hidden layer is fed back into the input layer, and *third*, networks in which the output layer is fed back into the input layer. These feedback networks are useful when input variables represent time series.

5 Note, we could alternatively use product rather than summation hidden units to supplement the inputs to a neural network with higher-order combinations of the inputs to increase the capacity of the network in an information capacity sense. These networks are called *product unit* rather than summation unit networks (see Fischer and Reismann, 2002b).

6 This term should not be confused with the term bias in a statistical sense.

7 The inverse of this function is called link function in the statistical literature. Note that radial basis function networks may be viewed as single hidden layer networks that use radial basis function nodes in the hidden layer. This class of neural networks asks for a two stage approach for training. In the first stage the parameters of the basis functions are determined, while in the second stage the basis functions are kept fixed and the second layer weights are found (see Bishop, 1995: 170 pp.).

8 This is the same idea as incorporating the constant term in the design matrix of a regression by inserting a column of ones.

9 In some cases there may be some practical advantage to use a *tanh* function instead. But note that this leads to results equivalent to the logistic function.

10 This viewpoint directs attention to the literature on numerical optimization theory, with particular reference to optimization techniques that use higher-order information such as conjugate gradient procedures and Newton's method. The methods use the gradient vector (first-order partial derivatives) and/or the Hessian matrix (second-order partial derivatives) of the error function to perform optimization, but in different ways. A survey of first- and second-order optimization techniques applied to network training can be found in Cichocki and Unbehauen (1993).

11 When using an iterative optimization algorithm, some choice has to be made of when to stop the training process. There are various criteria that may be used. For example, training may be stopped when the error function or the relative change in the error function falls below some prespecified value.

12 The particular form of $\eta(\tau)$ most commonly used is described by $\eta(\tau) = c/\tau$ where c is a small constant. Such a choice is sufficient to guarantee convergence of the stochastic approximation algorithm (Ljung, 1977).

13 The term *backpropagation* is used in the literature to mean very different things. Sometimes, the feedforward neural network architecture is called a backpropagation network. The term is also used to describe the training of a feedforward neural network using gradient descent optimization applied to a sum-of-squares error function.

14 Note that limited data sets make the determination of H more difficult if there is not enough data available to hold out a sufficiently large independent test sample.

15 A neural network is said to be overfitted to the data if it obtains an excellent fit to the training data, but gives a poor representation of the unknown function which the neural network is approximating.

16 In conventional curve fitting, the use of this regularizer is termed *ridge regression*.

17 Note that a reliable pseudo-random number generator is essential for the valid application of the bootstrap approach.

REFERENCES

Anders, U. and Korn, O. (1999). Model selection in neural networks. *Neural Networks*, **12**(2): 309–323.

Baldi, P. and Chauvin, Y. (1991). Temporal evolution of generalization during learning in linear networks. *Neural Computation*, **3**(4): 589–603.

Bäck, T., Fogel, D.B. and Michaelewicz, Z. (eds) (1997). *Handbook of Evolutionary Computation*. New York and Oxford: Oxford University Press.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**(6): 2350–2383.

Bridle, J.S. (1994). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In: Fogelman Soulié, F. and Hérault, J. (eds), *Neurocomputing. Algorithms, Architectures and Applications*, pp. 227–236. Berlin, Heidelberg and New York: Springer.

Carpenter, G.A. (1989). Neural network models for pattern recognition and associative memory. *Neural Networks*, **2**(4): 243–257.

Carpenter, G.A., Grossberg, S. and Reynolds, J.H. (1991). ARTMAP supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, **4**(5): 565–588.

Cichocki, A. and Unbehauen, R. (1993). *Neural Networks for Optimization and Signal Processing*. Chichester: Wiley.

Corne, S., Murray, T., Openshaw, S., See, L. and Turton, I. (1999). Using computational intelligence techniques to model subglacial water systems. *Journal of Geographical Systems*, **1**(1): 37–60.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control Signals and Systems*, **2**: 303–314.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Finnoff, W. (1991). Complexity measures for classes of neural networks with variable weight bounds. *Proceedings of the International Geoscience and Remote Sensing Symposium* (IGARSS'94, Volume 4), pp. 1880–1882. Piscataway, NJ: IEEE Press.

Finnoff, W., Hergert, F. and Zimmerman, H.G. (1993). Improving model selection by nonconvergent methods. *Neural Networks*, **6**(6): 771–783.

Fischer, M.M. (1998). Computational neural networks – A new paradigm for spatial analysis. *Environment and Planning A*, **30**(10): 1873–1892.

Fischer, M.M. (2000). Methodological challenges in neural spatial interaction modelling: the issue of model selection. In: Reggiani, A. (ed.), *Spatial Economic Science*: *New Frontiers in Theory and Methodology*, pp. 89–101. Berlin, Heidelberg and New York: Springer.

Fischer, M.M. (2002). Learning in neural spatial interaction models: A statistical perspective, *Journal of Geographical Systems,* **4**(3): 287–299.

Fischer, M.M. (2005). Spatial analysis. In Longley, P., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds), *Geographical Information Systems. Principles, Techniques, Management and Applications*. Second Edition, Abridged, (CD-ROM). Hoboken, New Jersey: Wiley.

Fischer, M.M. (2006a). Neural networks. A general framework for non-linear function approximation. *Transactions in GIS*, **10**(4): 521–533.

Fischer, M.M. (2006b). *Spatial Analysis and Geocomputation. Selected Essays*. Berlin, Heidelberg and New York: Springer.

Fischer, M.M. and Getis, A. (eds) (1997). *Recent Developments in Spatial Analysis. Spatial Statistics, Behavioural Modelling, and Computational Intelligence*. Berlin, Heidelberg and New York: Springer.

Fischer, M.M. and Gopal, S. (1994). Artificial neural networks: A new approach to modelling interregional telecommunication flows. *Journal of Regional Science,* **34**(4): 503–527.

Fischer, M.M. and Leung, Y. (1998). A genetic-algorithm based evolutionary computational neural network for modelling spatial interaction data. *The Annals of Regional Science,* **32**(3): 437–458.

Fischer, M.M. and Leung, Y. (eds) (2001). *GeoComputational Modelling: Techniques and Applications*. Berlin, Heidelberg and New York: Springer.

Fischer, M.M. and Reismann, M. (2002a). Evaluating neural spatial interaction modelling by bootstrapping. *Networks and Spatial Economics,* **2**(3): 255–268.

Fischer, M.M. and Reismann, M. (2002b). A methodology for neural spatial interaction modeling. *Geographical Analysis,* **34**(2): 207–228.

Fischer, M.M. and Staufer, P. (1999). Optimization in an error backpropagation neural network environment

with a performance test on a spectral pattern classification problem. *Geographical Analysis*, **31**(2): 89–108.

Fischer, M.M., Hlaváckova-Schindler, K. and Reismann, M. (1999). A global search procedure for parameter estimation in neural spatial interaction modelling. *Papers in Regional Science,* **78**(2): 119–134.

Fischer, M.M., Reismann, M. and Hlaváckova-Schindler, K. (2003). Neural network modelling of constrained spatial interaction flows: Design, estimation and performance issues. *Journal of Regional Science,* **43**(1): 35–61.

Fischer, M.M., Gopal, S., Staufer, P. and Steinnocher, K. (1997). Evaluation of neural pattern classifiers for a remote sensing application. *Geographical Systems,* **4**(2): 195–226.

Fogel, D.B. (1995). *Evolutionary Computation*: *Toward a New Philosophy of Machine Intelligence*. Piscataway, NJ: IEEE Press.

Fogel, D.B. and Robinson, C.J. (eds) (1996). *Computational Intelligence*. Piscataway: IEEE Press and Wiley-Interscience.

Foody, G.M., and Boyd, D.S. (1999). Fuzzy mapping of tropical land cover along an environmental gradient from remotely sensed data with an artificial neural network. *Journal of Geographical Systems*, **1**(1): 23–35.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, **2**(3): 183–192.

Gahegan, M. (2000). On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, **32**(1): 113–133.

Gahegan, M., German, G. and West, G. (1999). Improving neural network performance on the classification of complex geographic datasets. *Journal of Geographical Systems*, **1**(1): 3–22.

Goldberg, D.E. (1989). *Genetic Algorithms*. Reading, MA: Addison-Wesley.

Gopal, S. and Fischer, M.M. (1996). Learning in single hidden-layer feedforward network models. *Geographical Analysis* **28**(1): 38–55.

Gopal, S. and Fischer, M.M. (1997). Fuzzy ARTMAP – a neural classifier for multispectral image classification. In: Fischer, M.M. and Getis, A. (eds), *Recent Developments in Spatial Analysis*, pp. 306–335. Berlin, Heidelberg and New York: Springer.

Grossberg, S. (1988). Nonlinear neural networks. Principles, mechanisms and architectures. *Neural Networks*, **1**(1): 17–61.

Hassoun, M.H. (1995). *Fundamentals of Artificial Neural Networks*. Cambridge, MA and London, England: MIT Press.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Berlin, Heidelberg and New York: Springer.

Haykin, S. (1994). *Neural Networks. A Comprehensive Foundation*. New York: Macmillan College Publishing Company.

Hertz, J., Krogh, A. and Palmers, R.G. (1991). *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Wesley.

Hinton, G.E. (1987). Learning translation invariant recognition in massively parallel networks. In: Bakker, J.W. de, Nijman, A.J. and Treleaven, P.C. (eds), *Proceedings PARLE Conference on Parallel Architectures and Languages Europe*, pp. 1–13. Berlin, Heidelberg and New York: Springer.

Hlaváckova-Schindler, K. and Fischer, M.M. (2000). An incremental algorithm for parallel training of the size and the weights in a feedforward neural network. *Neural Processing Letters*, **11**(2): 131–138.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**(5): 359–368.

Huang, G.-B. and Siew, C.-K. (2006). Real-time learning capability of neural networks. *IEEE Transactions on Neural Networks*, **17**(4): 863–878.

Janson, D.J. and Frenzel, J.F. (1993). Training product unit neural networks with genetic algorithms. *IEEE Expert*, **8**(5): 26–33.

Kohonen, T. (1988). *Self-Organization and Associative Memory*. Berlin, Heidelberg and New York: Springer.

Kuan, C.-M. and White, H. (1991). Artificial neural networks: An econometric perspective. *Econometric Reviews*, **13**(1): 1–91.

Kůrková, V. (1992). Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, **5**(3): 501–506.

Leung, K.-S., Ji, H.-B. and Leung, Y. (1997). Adaptive weighted outer-product learning associative memory. *IEEE Transactions on Systems, Man, and Cybernetics – Part B*, **27**(3): 533–543.

Leung, Y. (1997). Feedforward neural network models for spatial data classification and rule learning. In: Fischer, M.M. and Getis, A. (eds), *Recent Developments in Spatial Analysis*, pp. 289–305. Berlin, Heidelberg and New York: Springer.

Leung, Y. (2001). Neural and evolutionary computation methods for spatial classification and knowledge acquisition. In: Fischer, M.M. and Leung, Y. (eds), *GeoComputational Modelling. Techniques and Applications*, pp. 71–108. Berlin, Heidelberg and New York: Springer.

Leung, Y., Chen, K.-Z. and Gao, X.-B. (2003). A high-performance feedback neural network for solving convex nonlinear programming problems. *IEEE Transactions on Neural Networks* **14**(6): 1469–1477.

Leung, Y., Dong, T.-X. and Xu, Z.-B. (1998). The optimal encodings for biased association in linear associative memory. *Neural Networks* **11**(5): 877–884.

Leung, Y., Gao, X.-B. and Chen, K.-Z. (2004). A dual neural network for solving entropy-maximising models. *Environment and Planning A*, **36**(5): 897–919.

Leung, Y., Chen, K.-Z., Jiao, Y.-C., Gao, X.-B. and Leung, K.S. (2001). A new gradient-based neural network for solving linear and quadratic programming problems. *IEEE Transactions on Neural Networks*, **12**(5): 1074–1083.

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, **AC-22**: 551–575.

McCulloch, W.S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**: 115–133.

Mineter, M.J. and Dowers, S. (1999). Parallel processing for geographical applications: A layered approach. *Journal of Geographical Systems*, **1**(1): 61–74.

Moody, J.E. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In: Moody, J.E., Hanson, S.J. and Lippman, R.P. (eds), *Advances in Neural Information Processing Systems 4*, pp. 683–690. San Mateo, CA: Morgan Kaufmann.

Murata, N. Yoshizawa, S. and Amari, S. (1993). Learning curves, model selection and complexity of neural networks. In: Hanson, S.J., Cowan, J.D. and Giles, C.L. (eds), *Advances in Neural Information Processing Systems 5*, pp. 607–614. San Mateo, CA: Morgan Kaufmann.

Nocedal, J. and Wright S.J. (1999). *Numerical Optimization*. Berlin, Heidelberg and New York: Springer.

Openshaw, S. (1993). Modelling spatial interaction using a neural net. In: Fischer, M.M. and Nijkamp, P. (eds), *GIS, Spatial Modelling, and Policy*, pp. 147–164. Berlin, Heidelberg and New York: Springer.

Openshaw, S. (1994). Neuroclassification of spatial data. In: Hewitson, B.C. and Crane, R.G. (eds), *Neural Nets: Applications in Geography*. pp. 53–70. Boston: Kluwer Academic Publishers.

Openshaw, S. and Abrahart, R.J. (eds) (2000). *GeoComputation*. London and New York: Taylor & Francis.

Openshaw, S. and Openshaw, C. (1997). *Artificial Intelligence in Geography*. Chichester: Wiley.

Plutowski, M., Sakata, S. and White, H. (1994). Cross-validation estimates IMSE. In: Cowan, J.D., Tesauro, G. and Alspector, J. (eds), *Advances in Neural Information Processing Systems 6*, pp. 391–398. San Francisco: Morgan Kaufmann.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, **78**(9): 91–106.

Press, W.H., Teukolky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in C. The Art of Scientific Computing*. 2nd edn. Cambridge: Cambridge University Press.

Ripley, B.D. (1994). Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society B*, **56**(3): 409–456.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Washington DC: Spartan Books.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986). Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pp. 318–362. Cambridge, MA: MIT Press.

Schwefel, H.-P. (1994). *Evolution and Optimum Seeking*. New York: Wiley.

Specht, D.F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks*, **2**(6): 568–576.

Stepniewski, W. and Keane, J. (1997). Pruning backpropagation neural networks using modern stochastic optimization techniques. *Neural Computing and Applications*, **5**(2): 76–98.

Tibshirani, R. (1996a). A comparison of some error estimates for neural network models. *Neural Computation*, **8**(1): 152–163.

Tibshirani, R. (1996b). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**: 267–288.

Wedge, D., Ingram, D., McLean, D., Mingham, C. and Bandar, Z. (2006). On global-local artificial neural networks for function approximation. *IEEE Transactions on Neural Networks*, **17**(4): 942–952.

Weigend, A.S., Rumelhart, D.E. and Huberman, B.A. (1991). Generalization by weight elimination with application to forecasting. In: Lippman, R., Moody, J. and Touretzky, D. (eds), *Advances in Neural Information Processing Systems* 3, pp. 875–882. San Mateo, CA: Morgan Kaufmann.

Weng, J. and Hwang, W.-S. (2006). From neural networks to the brain: Autonomous mental development. *IEEE Computational Intelligence Magazine*, **1**(3): 15–31.

White, H. (1989). Learning in artificial neural networks: a statistical perspective. *Neural Computation*, **1**(4): 425–464.

White, H. (1992). *Artificial Neural Networks. Approximation and Learning Theory*. Oxford, UK and Cambridge, USA: Blackwell.

White, H. and Racine, J. (2001). Statistical inference, the bootstrap, and neural-network modeling with applications to foreign exchange rates. *IEEE Transactions on Neural Networks*, **12**(4): 657–673.

Widrow, B. and Hoff, M.E. Jr. (1960). Adaptive switching circuits. *IRE Western Electric Show and Convention Record*, **Part 4**: 96–104.

Wilkinson, G.G. (1997). Neurocomputing for earth observation – recent developments and future challenges. In: Fischer, M.M. and Getis, A. (eds), *Recent Developments in Spatial Analysis*, pp. 289–305. Berlin, Heidelberg and New York: Springer.

Wilkinson, G.G. (2001). Spatial pattern recognition in remote sensing by neural networks. In: Fischer, M.M. and Leung, Y. (eds), *GeoComputational Modelling. Techniques and Applications*, pp. 145–164. Berlin, Heidelberg and New York: Springer.

Wilkinson, G.G., Fierens, F. and Kanellopoulos, I. (1995). Integration of neural and statistical approaches in spatial data classification. *Geographical Systems*, **2**(1): 1–20.

Yao, X. (1996). A review of evolutionary artifical neural networks. *International Journal of Intelligent Systems*, **8**(4): 539–567.

Yao, X. (2001). Evolving computational neural networks through evolutionary computation. In: Fischer, M.M. and Leung, Y. (eds), *GeoComputational Modelling. Techniques and Applications*, pp. 35–70. Berlin, Heidelberg and New York: Springer.

Yao, X., Fischer, M.M. and Brown, G. (2001). Neural network ensembles and their application to traffic flow prediction in telecommunication networks. In: *Proceedings of the 2001 IEEE-INNS-ENNS International Joint Conference on Neural Networks*, pp. 693–698. Piscataway, NJ: IEEE Press.

Zapranis, A. and Refenes, A.-P. (1998). *Principles of Neural Model Identification, Selection and Adequacy. With Applications to Financial Econometrics*. London: Springer.

Zhang, G., Patuwo, B.E. and Hu, M.Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, **14**(1): 35–62.

21

Geocomputation

Harvey J. Miller

21.1. INTRODUCTION

Geocomputation concerns the application of high-performance computers to explore and analyze digital representations of the Earth and related phenomena. Geocomputation is predicated on the belief that computational techniques are useful for explaining and predicting geographic phenomena. This is due to three reasons: (1) astonishing increases in computational power allowing new forms of modeling, analysis, and simulation; (2) greater capabilities for collecting and storing geographic data, allowing unprecedented detailed representations of geographic phenomena, and; (3) a postulate that computation is meaningful for understanding reality beyond the computational process itself, and perhaps better than traditional, analytical approaches.

The next section of this chapter reviews the conceptual foundation for geocomputation. This includes discussions of the broader field of computational science and the complexity of natural processes. It also reviews the motivations for geocomputation, its relationship to spatial analysis and geographic information systems (GIS), and elements of a theory of geocomputation. The next major section reviews selected techniques to illustrate the application of geocomputation principles in spatial analysis. The final section concludes this chapter by briefly discussing the future of GC.

21.2. COMPUTATIONAL SCIENCE AND COMPLEXITY

21.2.1. Computational science

In contrast with *computer science* or the study of computers and computation, *computational science* (CS) uses computers to study other scientific problems.

CS involves the development and application of computational techniques using high performance computers to explore massive databases and to simulate complex and intricate processes. This complements the use of traditional scientific techniques such as experimentation, analytical modeling, and statistical techniques. These techniques are limited since they can only explore a small portion of the vast information spaces implied by some phenomena and often require harsh assumptions that are not met in reality, or by the data acquired through measuring reality (Openshaw, 2000).

One motivation behind CS is that computers have become incredibly powerful and will continue to do so for the foreseeable future. These potent platforms can provide new, revolutionary tools for scientific investigation. Another motivation is the collapse in costs of data collection and storage.

Moore's Law

The now famous Moore's Law of Integrated Circuits best describes the incredible growth in computing power. In 1965, Gordon Moore (one of the inventors of the integrated circuit and then Chair of Intel, Inc.) noted that the surface area of each transistor being etched on an integrated circuit was being reduced by about 50% every 12 months. In 1975, he revised this to 18 months. This is known as *Moore's Law*: the processing capacity of the integrated chip doubles every six months. Figure 21.1 provides evidence of Moore's Law (Kurtzweil, 1999).

Moore's Law implies *exponential* growth in computational power. We have orders of magnitude more computing power available to us than to researchers as recently as 10 years ago, and many orders of magnitude more than the time when many of our analytical, statistical and experimental techniques were developed. Perhaps we should re-think

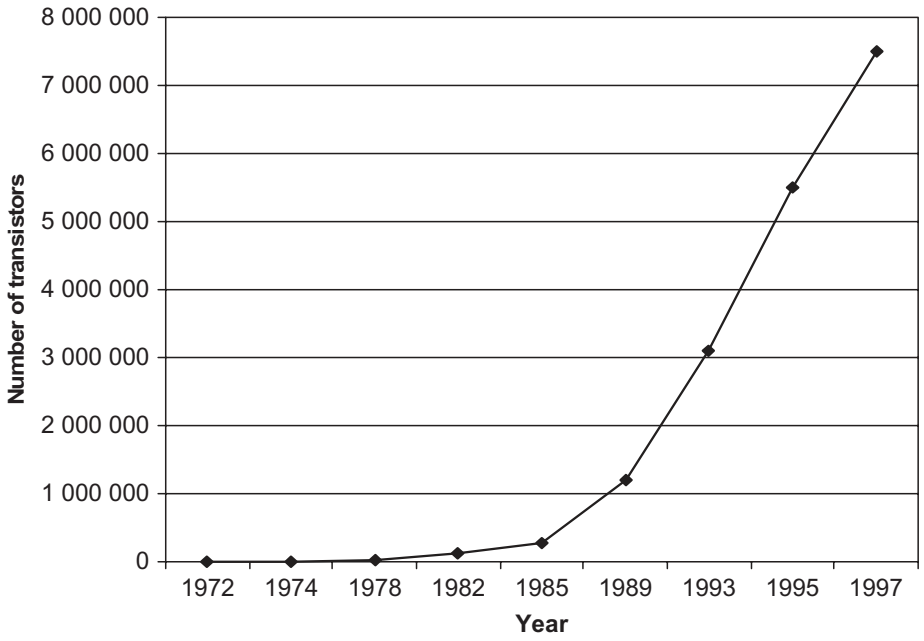


Figure 21.1 Evidence of Moore's Law – the number of transistors on Intel IC chips (based on Kurtzweil, 1999).

our tools and methods given this growth in computing power (Openshaw, 2000).

Data collection and storage

Paralleling the astonishing increases in computational power is an equally stunning collapse in the cost of collecting and storing digital data. The computerization of many government and business transactions as well as the increasing capabilities for direct digital data capture through devices such as bar code scanners and environmental sensors has greatly reduced the cost of data collection. At the same time, database, and data warehousing techniques have become more powerful and affordable (Chen *et al*., 1996). The hardware costs of storing data are also a minute fraction of the costs a decade or two ago. These trends are shifting science from a data-poor to a data-rich environment.

21.2.2. Nature and complexity

Just because we are drowning in CPU cycles and data does not mean we should apply them to understanding non-computer related phenomena. Computers may not be appropriate tools for gaining new scientific understanding about reality apart from the computational process itself. Perhaps computers should just be used to manage our data and documents, run our personal digital assistants and cell phones, and coordinate transportation and logistics. In other words, just because computers are great for solving engineering problems, does not mean that they are useful to discover knowledge about the Earth. However, computational science is also predicated on a belief that computation can mimic natural processes. Nature may behave much like a computer; in fact, *perhaps the universe is a computer* (Kelley, 2002).

This may sound far-fetched. But until recently, science has worked under a similar but equally pervasive metaphor, namely, *universe as machine*. This assumed that the universe behaved much like an engine, with continuous and well-behaved processes with effects that are proportional to causes. Most importantly, this implied that *the whole is equal to sum of the parts*, and we can understand the whole by studying its parts independently. The tools for this exploration were algebra and calculus: these are tools that examine quantities (magnitudes) in continuous mathematical space (Flake, 1998).

There is a fundamental reason why computation may be a better description of nature than mechanics: frugality. Natural processes have a remarkable ability to extract much from a minimal investment of resources: consider, for example, the surface area generated by the leaves of a tree or in the interface between the lungs and the circulatory system. Similarly, a great deal of biological complexity results from a code that consists of only four symbols, namely, DNA. Similarly, computing tries to obtain the most with the least investment of computational resources and, similar to biological growth, simple computational rules can result in complex behavior that is not predictable. The property of *complexity from simplicity* in both nature and computation means that the *whole is greater than the sum of the parts*: phenomena cannot be understood entirely by independent analysis of their components. This implies a middle path to scientific knowledge: instead of looking at the individual or aggregate, we see how the aggregate emerges from the interactions of individuals (Flake, 1998).

There are some powerful mechanisms in both nature and computation that facilitate complexity from simplicity. These principles are *parallelism*, *iteration*, and *adaptation* (Flake, 1998). Complex systems are often highly parallel collections of relatively

simple units, for example, consider an ant colony (ants), a brain (neurons), an ecosystem (animals), or a city (people). Parallel systems are more efficient and robust than sequential systems since they can specialize, explore a wider range of solutions simultaneously, and survive the failure of many components. Iteration over time allows feedback from the environment to determine the success or failure of different units and their strategies. Iteration also supports the closely related concept of *recursion* or self-reference. Finally, adaptation is a consequence of parallelism and iteration within an environment with scarce resources and therefore competition. Many of the techniques used in computational science incorporate some or all of these principles.

21.3. GEOCOMPUTATION

21.3.1. Motivation

Similar to CS, a factor motivating geocomputation is the increasing ability to capture, store, and process digital geographic data. In particular, it is increasingly possible to capture geo-data at high levels of spatial and temporal resolution as well as manipulate very detailed representations of geography using geographic information systems (GIS) and related technologies. Geo-spatial data capture technologies include intelligent transportation systems, hyperspectral, and laser-based remote sensing systems, environmental monitoring devices, and location-aware technologies (LATs) that can report their geo-location densely with respect to time. GIS allow analysis of geographic relationships and morphology at levels of detailed hardly imaginable even a short time in the past (Miller and Wentz, 2003). It is possible that there are surprising and useful patterns in these data and representations that are not being discovered by the analytic and statistical methods in traditional spatial analysis.

Another motivating force for geocomputation is the increasing recognition of the complexity of the spatio-temporal systems of concern in geography and the Earth sciences (Fischer and Leung, 2001). For example, the dynamic evolution of an urban system emerges from the individual agents of change, their interactions and the co-evolution of the context in which these interactions occur. This suggests not only that these systems are more complicated than previously supposed, but also that we cannot engineer their growth; rather, we can only influence or shape their evolution. We have seen this time and time again when a relatively modest change in infrastructure or policy (e.g., a new highway interchange, a change in zoning regulations) leads to wildly disproportionate outcomes (e.g., traffic congestion, urban sprawl). This is not defeatist; rather, it suggests humility and the need for sophisticated, nuanced approaches to understanding and directing these systems to efficient, equitable and sustainable outcomes.

In addition to increasing recognition of the complexity of geographic phenomena, it is also likely that *intrinsic* complexity of these systems is increasing. As the world continues to become more crowded, mobile and connected, small local actions can have large-scale outcomes. Saturated road networks mean that an accident in one location results in traffic jams across town. Airline networks distribute diseases across continents and around the world within hours. The Internet spreads innovative ideas and wild rumors throughout the globe nearly at the speed of light. Interconnected financial networks mean that decisions made in a conference room can have huge economic consequences for large

regions thousands of miles away. Population pressure and consequent desertification in Central Asia creates air quality problems in North America. A crowded planet with high consumption, mobile and connected lifestyles creates complex spatio-temporal dynamics at all geographic scales.

Many in geography and other Earth sciences have known for quite some time that the world is more sophisticated, and becoming even more sophisticated, than the mechanical metaphor bequeathed to us by 18th century science. The problem is that we did not have the tools or data for dealing with basic and applied scientific problems from a complex system perspective. The geocomputational revolution is shattering this barrier.

21.3.2. Distinctive features of geocomputation

An increasing recognition of the complexity of geographic systems and the collapse in the costs of capturing and storing geographic data does not necessarily justify a separate field of study. It is possible that standard CS techniques could be applied to geographic phenomena with the same success with which they are applied to other scientific questions. What is it about GC that makes it distinct from CS?

A major distinction is the emphasis on the geospatial framework that conditions the phenomenon under investigation (Openshaw, 2000). A *geo-space* is a set of locations with defined shortest path relationships between all pairs (Beguin and Thisse, 1979). These locations often correspond to the Earth's surface, although this is not necessarily the case. Shortest path relationships are usually based on physical distance, but other relationships such as time, direction and connectivity, or some combination, are possible (see Miller and Wentz, 2003). Two major properties conditioned by a geospatial framework are *spatial dependency* and *spatial heterogeneity*; these refer to the tendency of attributes at locations that are proximal with respect to shortest paths to be related and the tendency of processes to vary by location in geospace (respectively). Rather than confounding, these properties are rich sources of information about spatial processes (Fotheringham, 2000). In addition, many geographic entities cannot be represented as simple points in an information space without significant loss. Geographic entities have morphological properties such as size and shape that can have non-trivial effects on their evolution and interactions (see Miller and Wentz, 2003). GC is the development and application of computational techniques that are sensitive to spatial relationships and spatial morphology inherent in geo-spatial phenomena.

21.3.3. Relationship to spatial analysis and geographic information systems

At its core, the argument for GC being unique with respect to CS is identical to the arguments for the uniqueness of spatial analysis with respect to statistics, and GIS being a unique subset of information systems. What, if anything, makes GC unique with respect to spatial analysis and GIS?

Fotheringham (2000) distinguishes between computer-based spatial analysis and GC. Computer-based spatial analysis uses the computer merely as a convenient tool (i.e., nothing more than, say, a very fast slide rule or abacus). GC refers to the case where the computer *drives* the form of spatial analysis instead of just being a convenient vehicle. In other words, in GC *the computer plays a pivotal role*.

GC is more concerned with finding numerical approximations rather than precise

analytical solutions. While this may sound like a drawback (and it is), this is a necessary trade-off. Traditional modeling methods rely on simplistic representations of space and behavior in order to facilitate precise analytical solutions. GC determines numerical approximations of solutions for systems with more complex representations of space and behavior. The argument is that it is better to have an approximate solution to a richly represented system than an exact solution to a sterile representation. Numerical approximations are necessary consequences of richer, more accurate representations of geographic phenomena.

Much of the digital geographic data available to researchers no longer meets many of the assumptions of inferential statistics, including the more relaxed assumptions of spatial analysis. Geographic data is increasingly no longer carefully structured and limited samples from a much larger population. Rather, digital geographic data are often monitored entire populations (in the statistical sense) collected using ill-structured and 'noisy' methods. Computational techniques that do not require strict assumptions are better suited for these rich but sloppy data (Atkinson and Martin, 2000).

GIS provides a source of data and a toolkit environment for GC. GC is distinct since it emphasizes dynamic processes over static form and user interaction over passive receipt of information. GC is about matching technology with environment, process with data model, geometry with application, analysis with local context, and, philosophy of science with practice (Longley, 1998). We can also make a distinction between GIS and GC that is similar to Fotheringham's (2000) distinction between computer-based spatial analysis and GC. In many respects, the computer is nothing more than a convenient vehicle for GIS. For example, the overlay operation pre-dates much of the development of computer-based GIS (see McHarg, 1969). We have also been constructing, storing and using maps for 5000 years. In other words, we can 'do' GIS even without computers, although it would be very slow and tedious. GC is about what we could not do before the development of powerful computers.

In sum, GC uses the traditional techniques of spatial analysis (statistics, mathematical modeling) and GIS as parts of a more flexible and expansive tool kit. GC is concerned with the use of computational techniques and technologies within a scientific framework. This involves GIS as the data and information manager, computational methods as the tools, and high performance computing as the driver (Fischer and Abrahart, 2000; Fischer and Leung, 2001; Openshaw, 2000).

21.3.4. A theory of geocomputation?

Couclelis (1998a, b) provides a more skeptical view of GC. She argues that GC is in fact a loosely connected 'grab-bag' of techniques rather than a focused scientific endeavor. She challenges the GC community to develop a rigorous computational theory of spatiotemporal processes that justifies the prefix 'geo.'

Couclelis points out that computational science is based on the *theory of computation*, a highly developed and rigorous theory of what can (and cannot) be computed and how things that can be computed should be. This involves questions such as determining which processes in the world can be described in the precise manner required by computation, and what is the appropriate language for describing specific processes. These are much deeper questions than what available computational technique is best for a particular data set or problem. (For an excellent introduction

to the theory of computation, see Sipser (1997).)

Noting the formal equivalence between the theory of computing and the theory of algebra, Couclelis (1998b) also develops a rough typology that demarcates the types of techniques that can legitimately be called geocomputation. In Table 21.1, the upper left quadrant contains techniques that are definitely geocomputational. The lower right quadrant contains techniques that are definitely not geocomputational. The upper right and lower left quadrants contain borderline cases.

Couclelis also distinguishes between hard GC and soft GC, paralleling a similar distinction between hard and soft artificial intelligence. *Hard GC* involves efforts to understand and represent complex geographical processes using computational techniques. *Soft GC* refers to the development of geographic problem representations and solutions using spatially oriented computational techniques.

Couclelis' (1998a, b) challenge has yet to be met by the GC community. Answering these fundamental questions can indicate the deep connections between different geocomputational tools as well as between different geographic phenomena; this will undoubtedly advance the field as well as provide it with a more rigorous foundation.

21.4. GEOCOMPUTATIONAL TECHNIQUES

This section reviews selected geocomputational techniques, specifically, *fractals, dynamical systems and chaotic behavior, cellular automata, agent-based modeling*, and *artificial neural networks.* This is not an exhaustive list. Other techniques such as geographic knowledge discovery, visualization, local spatial statistics, and optimization techniques such as genetic algorithms could also be included since these are data and computation-hungry techniques where the computer plays a pivotal role. However, some of these techniques are better treated independently, as indeed they are elsewhere in this volume. The survey below intends to provide illustrative examples of geocomputation, as well as demonstrate the pervasive theme of complexity from simplicity that underlies most geocomputational techniques.

21.4.1. Fractals

Platonic objects such as points, lines, polygons and solids are simple, smooth and ideal, and typically only result from deliberate design. In contrast, naturally occurring objects such as coastlines, clouds, trees and the human circulatory system are highly self-similar, hierarchical and irregular. Each part

Table 21.1 A general classification of geocomputational and non-geocomputational techniques (Couclelis, 1998b)

		Operators	
		Spatial	Nonspatial
Operands	Spatial	Cellular automata Shape grammars Fractals	Map classification Neural networks Multimedia imaging
	Nonspatial	Cartographic labeling	Traditional modeling Global statistics

appears to be a scaled-down version of the entire object, these self-similar features form a hierarchy, and the boundaries of these objects are highly irregular (Mandlebroit, 1983). Also, many human-made objects such as transportation networks and cities that grow in a bottom-up, organic manner also appear irregular, hierarchical and self-similar (Batty and Longley, 1994).

A fractal is a class of complex geometric shapes that have *fractional dimension*, a concept first introduced by the mathematician Felix Hausdorff in 1918. Benoit Mandelbrot coined the term 'fractal' from the Latin word fractus ('fragmented,' or 'broken') since the shapes are irregular rather than smooth (Batty and Longley, 1994; Encyclopedia Britannica, 2000). For example, points, lines and polygons have dimensions of zero, one and two, respectively. In contrast, a fractal curve has a dimension between one and two, and has a highly irregular and complex shape, while a fractal region has a dimension between two and three.

Fractals were discovered well over a century ago, but were considered to be 'pathological' and 'monsters' (Mandlebroit, 1983). The rise of the digital computer has facilitated the analysis and appreciation of these monsters since they seem to be based on the computational principles of iteration and recursion. Indeed, it is possible that many natural objects and processes exhibit fractal properties since iteration and recursion are efficient ways to grow objects. Many fractals can be generated through recursive functions that are very compact and require little information to encode their algorithm. Fractals are also very good at maximizing functionality with minimal resource inputs. Fractals such as the Koch snowflake and Peano curve can cram an incredible amount of length (in fact, infinite) into a finite area (Flake, 1998).

Since many natural and geographic phenomena display fractal properties, they are becoming important in spatial analysis and geocomputation (Goodchild and Mark, 1987; Longley, 2000). Fractals can also serve as the basis for spatial sampling strategies and other forms of spatial analysis (e.g., Appleby, 1996; De Cola, 1991; Lam and Liu, 1996). Because of their natural look, fractals are also becoming popular in computer graphics, particularly for rending natural landscapes such as mountains or other complex terrain (Clarke, 1993; Illingworth and Pyle, 1997). Fractals also provide principles for spatial data structures that map two- and three-dimensional data to the one-dimensional data structures in computers; examples include space-filling curves such as the Peano curve and the Hilbert curve (Goodchild and Mark, 1987).

Measuring the fractal dimension

As the properties discussed above suggest, fractals are more complex than Platonic objects. The fractal dimension is a measure of this complexity: Mandlebroit (1983) noted that the often paradoxical properties of fractals (such as enclosing a finite space with in infinite boundary, or packing an infinite length into a finite space) are a result of their 'dimensional discordance'.

The complexity of a fractal object relates to scale-dependency when measuring its size. The first person to notice this was Lewis Richardson, although this recognition may go back to the ancient Greeks (Batty and Longley, 1994). In 1967, Mandelbroit published a paper based on Richardson's insight entitled 'How long is the coast of Britain?' (Mandlebroit, 1967). The apparent length of a coastline seems to increase whenever the resolution of the measurement unit is increased: higher resolutions mean that smaller and smaller features become relevant, increasing the measured length. At the extreme, using an infinitely precise measure, the coast will appear to be infinite

in length. Therefore, we must conclude that the length of this naturally occurring object is meaningless, independent of the scale of measurement.

We can estimate the fractal dimension of an entity by comparing the growth in its apparent length or size with the change in the scale of the measurement. Essentially, this is an attempt to estimate the following power law (Peitgen *et al*., 2004):

$$y \propto x^d$$

where y is the size of the object, x is the measurement scale, and d is an empirical parameter related to the dimension of the object. In practice, estimating this relationship is complex: there are several definitions and measures of the fractal dimension, not all of which agree (see Lam and De Cola, 1993; Moon, 1992; Peitgen *et al*., 2004). Common fractal dimensions include the *similarity*, *capacity*, and *Hausdorff–Besicovich* dimensions (Batty and Longley, 1994; Goodchild and Mark, 1987; Williams, 1997). Methods for calculating these dimensions include *box-counting, compass*, *area-perimeter*, and *variogram* methods (Burrough, 1993; Peitgen *et al*., 2004).

Measuring the fractal dimension of geographic phenomena allows determination of its *scale-invariance* (self-similarity at different scales) as well as other fractal properties such as space-filling and irregularity. The increasing availability of digital geographic data as well as GIS tools for handling these data can support these analyses, allowing more detailed examination of the relationships between spatial process and geographic form (Batty and Longley, 1994; Longley, 2000). Applications include spatial population distributions (Appleby, 1996), transportation network morphology (Benguigui and Daoud, 1991), urban morphology (Batty and Longley, 1987, 1994; De Keersmaecker *et al*., 2003; Shen, 2002) land cover patterns (De Cola, 1989), landscape analysis (Burrough, 1993; Clarke and Schweizer, 1991) and riparian networks (Phillips, 1993a). Wentz (2000) uses a fractal dimension measure as a component of a general, trivariate shape measure.

Simulating fractal growth

In addition to fractal analysis of geographic patterns, it is also possible to simulate fractal growth using rule sets and iterated systems. Simulating fractal growth from finite systems such as rule sets and iterated systems captures a key property of fractal growth in the real world: the ability to generate highly complex entities using very simple processes. Physical, biological and human systems evolve from some baseline apparently without encoding complex information such as systems of simultaneous equations, constrained optimization problems, or partial differential equations to govern their growth. Rather, real world phenomena may emerge through simple growth mechanisms applied recursively. Many methods for simulating fractal growth use the powerful technique of recursion to generate complex structures with minuscule base information.

Two well-known recursive methods for generating fractals are *iterated functional systems* (IFS) and *L-systems*. The IFS algorithm starts with a seed object and maps a point on that object back onto itself through some randomly chosen affine transformation. This recursive process iterates and the object approaches a fractal object consisting of the union of smaller copies of the seed object (see Batty and Longley, 1994; Barnsley, 1988; Flake, 1998). L-systems simulate biological growth through a rule-based system that generates progressively complex strings through recursively applying the production rules to the axioms and the strings generated through these applications. This results in structures

with fractal properties that can be visualized using systems such as turtle graphics (Flake, 1998; Peitgen *et al*., 2004).

Other methods that simulate fractal growth include tessellation methods such as cellular automata (White and Engelen, 1993; also see below) and diffusion-limited aggregation (Batty, 1991; Fotheringham *et al*., 1989); these methods have been applied to simulating urban dynamics. Brownian motion methods have been applied to simulate natural objects with fractal properties such as riparian networks, geological time series and terrain (Goodchild and Klinkenberg 1993).

Fractal analysis and fractal simulation appear to be powerful methods that can reveal or mimic the structure and processes underlying many natural and human systems. The critical question remains whether explicit linkages can be identified between fractal processes and the natural and behavioral mechanisms identified from the domain sciences. It is important to note that some fractal algorithms are heuristics that imply unrealistic growth processes. To this end, correspondence between fractal processes and central place theory (Arlinghaus, 1985; Arlinghaus and Arlinghaus, 1989) and von Thünian theories of urban structure (Cavailhès *et al*., 2004) have been established.

21.4.2. Dynamical systems and chaotic behavior

A dynamical system is a system that experiences some change or motion. Many (if not most) natural and human made systems are dynamic. The traditional way to study dynamical systems is through *differential equations* and *difference equations*. Differential equations are continuous-time equations where one or more of the variables are rates of change expressed as derivatives. Difference equations capture discrete-time dynamics, with rates of change expressed in terms of differences in the values of variables at different points in time.

For many years it was assumed dynamical systems exhibited one of three types of behavior with respect to time (Flake, 1998): (1) fixed point (static); (2) periodic (orbit), and; (3) quasi-periodic (orbits that never quite repeat themselves). However, it was also known that certain types of dynamical systems exhibited behaviors that were intractable analytically. In particular, *non-linear dynamical systems* were known to be notoriously difficult. Since the rise of the digital computer, it has become easier to study non-linear dynamical systems using numerical simulation. This has led to the discoveries that these systems are not just intractable: they show very complex behavior now referred to as *chaos*.

Chaos is not randomness: completely deterministic systems can exhibit chaotic behavior. Yet this behavior is seemingly random with respect to prediction: forecasts about these systems over the long-run are poor, even though the mechanisms of the system are known. In particular, chaotic systems are highly sensitive to initial conditions: small differences in the starting points can lead to huge differences in their trajectories later in time. Chaotic behavior seems to be inherent in many types of nonlinear dynamical systems, even those with very simple structures: population dynamics, predator–prey dynamics, weather, and the stock market are all examples of real world processes that can be difficult to predict even if we know the underlying mechanics (Flake, 1998).

Chaotic behavior and strange attractors

Two well-known non-linear systems that generate chaotic behavior are the *Lorenz attractor* and generalized *Lotka–Volterra systems.* The Lorenz attractor consists of

three linked differential equations that model convection flow in weather systems. Generalized Lotka–Volterra systems model predator–prey relationships through n linked differential equations, where n is the number of species. This system displays of wide range of dynamic behavior under different parameterizations, including chaotic behavior (Flake, 1998).

The Lorenz attractor and generalized Lotka–Volterra systems capture many of the characteristics of chaotic dynamical systems. Both are non-linear and incorporate feedback: for example, in Lotka–Volterra systems the number of predators affects the number of prey through culling the latter, while in turn the number of prey affects the predators that can be supported. Both systems are very simple, but generate very complex behavior – behavior that often cannot be distinguished from randomness. However, the trajectories of these systems contain order, at least in a global sense. Finally, these systems are hypersensitive to initial conditions, with the consequence that while short-term behavior can be predicted, long-term predictions are meaningless (Williams, 1997).

An *attractor* is the bounded region within the phase space towards which dynamic systems evolve: examples include the fixed point, period, and quasi-periodic behaviors mentioned above. Chaotic systems are characterized by *strange attractors*. The system evolves within a finite space, but with an infinite period: visiting every location within the region but never the same location twice. Consequently, the calculated dimension of chaotic trajectories will often be fractional: contained within a finite area, but space-filling. These trajectories are often infinitely self-similar: increasing the resolution of the calculations and subsequent trajectory plots will reveal the same structure repeatedly. Thus, there is a deep linkage between fractals and chaos: both exhibit the computational principle of complexity from simplicity (Flake, 1998; Peitgen *et al*., 2004; Williams, 1997).

Spatial chaos

The non-linear dynamical systems we have discussed thus far exhibit temporal chaos, that is, chaotic behavior in the dynamic evolution of aggregate system parameters. A reasonable question is whether temporal chaos can lead to *spatial chaos* or complex spatial patterns that exhibit a high degree of sensitivity to conditions at particular locations. Theoretically, it turns out that spatial chaos can emerge from temporal chaos under very broad conditions: unless the system is perfectly isotropic with respect to space, spatial chaos will emerge and increase over time (Phillips, 1993b, 1999a). Given the broad conditions under which spatial chaos can emerge, it is not surprising that spatial chaos has been detected in physical and human geographic models and data. These include physical systems such as geomorphologic, hydrological, and ecological systems (see Phillips, 1999a, b), retail dynamics (Wilson, 2006), economic systems (White, 1990), urban systems (Wong and Fotheringham, 1990), and spatial choice processes (Nijkamp and Reggiani, 1990).

There are three general approaches to detecting spatial chaos (Phillips, 1993b). One method is to test for sensitivity to initial conditions by analyzing the Lyapunov exponents: these describe the average rate of convergence or divergence of two neighboring trajectories in phase space (Williams, 1997). A second method is numerical simulation: simulate and plot the behavior of the system in phase space, and analyze the plot using graphical techniques. A third approach is to examine an empirical temporal or spatial series for signatures of chaos, with the latter series derived by generating a spatial gradient by choosing some transect across space (see Phillips, 1993b).

Although techniques exist for detecting spatial chaos, this is nevertheless challenging since chaos often co-exists with stochastic uncertainty in real-world systems. Therefore, it can be difficult to extract the chaotic signature, particularly with systems that have a large number of variables and/or with datasets that are small and imperfectly measured. Consequently, a major challenge is to develop detection techniques that work given these real-world conditions (Phillips, 1993b; Williams, 1997).

21.4.3. Cellular automata

Cellular automata (CA) are discrete spatio-temporal dynamic systems based on local rules. Using relatively simple rule sets, CA can generate very complex spatio-temporal dynamics, including chaotic behavior (Flake, 1998). The potential for CA in spatial analysis has been recognized for quite some time. In a pioneering paper, Waldo Tobler describes the theoretical foundation for a cellular geography and defines five general classes of models, with CA being one of these classes (Tobler, 1979). Couclelis (1985, 1988) followed this with discussions of the potential of CA for capturing micro–macro spatial dynamics and the emergence of complex geographic systems from simple behaviors.

CA are becoming very popular in geographic research for a number of reasons. One is that they are inherently and explicitly spatio-temporal. A second reason is that they are computationally efficient and can be applied to problems with very high spatial resolution. There is also a natural link to GIS. GIS provides a platform for managing the spatial data required for CA. In return, CA allow GIS to go beyond static, geometric representations to include non-localized spatial processes such as spatial organization, configuration, pattern, dynamics, transformation, and change (White and Engelen, 1997).

CA components

A cellular automaton consists of the following components (Batty, 2000). The basic element of a CA is the *cell*. A cell is a memory element that stores different states. In the simplest case, each cell can have the binary states 1 or 0. In more complex simulation the cells can have several different states. The cells are arranged in a regular, discrete spatial configuration, usually a *lattice*. However, the grid configuration is not required; see O'Sullivan (2001) and Shi and Pang (2000). The state of each cell for the next time step is based on the states of the cells in its *neighborhood*. Common definitions of neighborhoods include *von Neumann* (a neighborhood with radius = 1 following the rook's case), *Moore* (an enlargement of the von Neumann neighborhood to contain diagonal cells), and *extended Moore* (a Moore neighborhood with radius = 2). It is also possible to relax the assumption of neighborhood to allow non-local effects (see Takeyama and Couclelis, 1997).

Transition rules determine the state of a cell at time $t + \Delta t$ based on the pattern of cell states within its neighborhood at time t. The set of transition rules are finite and constant across all cells. The number of possible transition rules can be enormous. If Ω is the number of possible states and h is the size of the neighborhood, then the number of possible cell patterns is $p = \Omega^h$. Given these patterns, there are $R = \Omega^p$ different transition rules for the cell. For example, in a Moore neighborhood with binary states, there are $2^9 = 512$ possible cell patterns and $2^{512} = 1.340780793$ E 154 possible rule sets; a number that is larger than the number of elementary particles

in the universe (Batty, 2000). However, although the number of possible transition rules is enormous, in practice the rules are typically very simple and refer to aggregate properties rather than detailed patterns within neighborhoods.

Empirical studies by Wolfram and others show that even the simple linear automata above behave in ways reminiscent of complex biological systems. For example, the fate of any initial configuration of a cellular automaton is to (1) die out; (2) become stable or cycle with fixed period; (3) grow indefinitely; (4) grow and contract in a complex manner (Wolfram, 1984). The important implication of these properties is that models of complex systems need not be complex themselves: simple rules can generate the complex behavior we associated with biological entities, ecosystems, economic systems, and cities (Couclelis, 1988).

Global from local

A very important property of cellular automata is the emergence of global patterns from local rules. The transition rules that drive the system are purely local: each cell's future state is based on neighboring cells only. Yet, higher-level global patterns and structure emerge from these purely local rules. The system self-organizes at the global level: there are no overarching rules yet global patterns emerge. Applications of CA span a wide range of emergent geographic phenomena: these include urban dynamics and land-use (Clarke and Gaydos, 1998; Clarke *et al*., 1997; White and Engelen, 2000; Xie, 1996; also see Benenson and Torrens, 2004, chapter 4), wildfire propagation (Clarke *et al*., 1994), traffic simulation (Esser and Schreckenberg, 1997) as well as physical geographic phenomena such as forest succession, land cover, and species composition (see Parker *et al*., 2003).

Despite the increasing popularity of CA in geocomputational modeling, standard CA have some restrictions that can cause some concerns when modeling geographical processes. One limitation is the assumption of *time–space stationarity*: a cell's future state is completely characterized by the states in its neighborhood according to static transition rules. Cells have no inherent characteristics that can affect its transitions. Therefore, a given configuration of cells in the neighborhood of that cell will result in the same transition regardless of that cell's location in space and time (White and Engelen, 1997). Phipps and Langlois (1997) note that time–space stationarity is particularly problematic when modeling geographic processes such as land use dynamics. The location of a parcel of land with respect to the rest of the system can affect its land use over time. Similarly, the conditions that affect land use at one time can change; for example, zoning laws may change as vacant parcels are filled and pressure builds to relax zoning restrictions.

Another problem is the assumption of *unconstrained transitions*: the number of cells in each state is determined endogenously by the application of transition rules to the current configuration with no recognition of potential exogenous constraints. Li and Yeh (2000) address the unconstrained transition problem by incorporating environmental constraints into their CA-based urban dynamics model.

A third weakness of using CA to model geographic processes is that a deterministic rule set is unrealistic. Other unobserved factors (such as individual choice) can influence state transitions. Phipps and Langois (1997) develop a stochastic framework for CA-based modeling of land-use dynamics. Also see de Almeida *et al*. (2003), Batty (2000), and Wu (2002) for discussions and examples of stochastic CA.

Scale is also a critical issue in CA modeling of geographic phenomena. Scale issues

are inherent in the choice of cell size as well as the neighborhood definition. Ménard and Marceua (2005) perform a sensitivity analysis of scale and the resulting spatial patterns and dynamics in a CA model of land-cover change. They discover substantial, non-linear relationships between these scale issues and the simulation results.

21.4.4. Agent-based modeling

An *agent* is some independent unit that tries to fulfill a set of goals in a complex, dynamic environment. These goals can be 'end goals' or ultimate states that the agents try to achieve, or they can be some type of reinforcement or reward that the agent attempts to maximize. The environment can be very general, and often includes other agents. An agent is *autonomous* if its actions are independent, i.e., it makes decisions based on its sensory inputs and goals. An agent is *adaptive* if its behavior can improve over time through some learning process (Maes 1995). Agents interact by exchanging physical or virtual (informational) resources. These interactions are typically very simple: they can be described by a small set of rules. From the pattern and intensity of these interactions emerge complex behavior that is not completely predictable or controllable: it materializes from the interactions of these rules (Flake, 1998).

The agent perspective is very general: many systems can be viewed as collections of autonomous, adaptive, and interacting agents. In *agent-based modeling* (ABM), we are concerned with simulated agents (software representations) as opposed to embodied agents (such as humans). *Multi-agent systems* (MAS) are ABMs that contain a distribution of simulated and interacting agents (Boman and Holm, 2004). ABM and MAS are bottom-up, individual-based approaches to simulating physical and human phenomena. The objective is to simulate the dynamics of complex systems through the behaviors and interactions of its individual agents. Agents can represent people, households, animals, firms, organizations, regions, countries, and so on, depending on the scale of the analysis and the elemental units hypothesized for that scale.

Similar to CA, ABM in many respects exemplifies the geocomputational approach. ABM is motivated by the view that many geographic phenomena are emergent: simple processes generate complex structure and patterns. In addition, the increasing availability of high-resolution data and GIS tools for handling these data facilitate ABM in geographic research (Benenson and Torrens, 2004).

Generative geographic science

ABM is a critical tool in a distinct, generative approach to science that focuses on the following question: How could the decentralized local interactions of autonomous agents generate a given pattern? The analyst attempts to answer this question by situating an initial population of autonomous agents in a relevant spatial environment and allowing them to interact according to simple rules, thereby generating the macroscopic regularity from the bottom up. If the analyst can reproduce the macroscopic pattern, than the microspecification is a candidate explanation (Epstein, 1999).

ABM is well-suited as a central tool in generative science due to the following realistic characteristics (Epstein, 1999): (1) *heterogeneity* – agents represent individual entities with unique characteristics that can change over time, as opposed to the static, aggregate representative agents in traditional social and other sciences; (2) *autonomy* – there is no central control over individuals in agent-based models, except for feedback

between macrostructures to microstructures (such as newborn agents learning social norms or shared culture), and therefore no need to postulate an abstract central authority or governing equations to facilitate the modeling; (3) *explicit space* – agent behavior and events occur in an explicit space, whether it is real (e.g., geographic) or abstract (e.g., social networks); (4) *local interactions* – agents interact only with other agents and environmental factors within some bounded regional of space and time; and (5) *bounded rationality* – agents have limited information, often based on their local neighborhoods, and limited abilities to process this information.

Similar to CA, explicit representations of space and local neighborhoods for interaction make ABM a natural tool for analyzing geographic phenomena. Unlike CA, the neighborhood over which an agent interacts can be more fluid and flexible. In addition, an agent can be mobile: in addition to changing its state, it can change its location in space over time. In some respects, these distinctions are arbitrary and historical: CA and ABM can be seen as special cases of a more general geographic automata system containing spatially fixed (CA) and non-fixed (ABM) automata (Benenson and Torrens, 2004). Similarly, Boman and Holm (2004) argue that time geography (see Hägerstrand, 1970) can serve as a unifying principle for ABM and the older tradition of *microsimulation* by providing a more explicit representation of real-world spatial and temporal constraints on agent behaviors and interactions.

ABM have been applied in diverse domains such as economics (see Epstein, 1999; Tesfatsion and Judd, 2006), environmental management (Gimblett *et al*., 2002; Hare and Deadman, 2004), land-use/land-cover change (Parker *et al*., 2003), urban dynamics (Benenson and Torrens, 2004), societies and culture (Epstein and Axell 1996), transportation (Balmer *et al*., 2004), and human movement at micoscales (Batty *et al*., 2003).

Automata-based modeling such as ABM and CA does have some substantial weaknesses and challenges (Epstein, 1999). First, automata-based modeling lacks standards for model comparisons and replication of results (Axtell *et al*., 1996). Unlike analytical modeling, subtle design differences (such as asynchronous versus synchronous agent updating) can make huge differences in the results, and there are no standards for reporting these decisions. Second, solution concepts are weak: a simulation run is only one possible path of a (typically) stochastic process, not a general solution. Consequently, there is need for careful experimental design to fully explore or sample from the information space implied by the model. However, this leads to a third challenge. While the parameter space given a postulated rule set is usually small, the space defined by combinations of possible agent rules can be enormous and difficult to explore fully (recall the earlier discussion regarding the number of potential CA rules). However, this can be mitigated to some degree by theory: similar to any good modeling, theoretical correctness should help distinguish between plausible and implausible rules.

21.4.5. Artificial neural networks

Continuing development and deployment of technologies for capturing geographic information such as remotely sensed imagery, vehicle-based GPS receivers, flow gauges, and automated weather reporting stations are generating huge but error-prone datasets. To exploit this noisy data requires techniques that are robust (fault-tolerant) and scalable (process large databases in reasonable, perhaps even real, time). *Artificial neural networks* (ANNs) are an important

class of computational techniques that can exploit noisy data, as well as solve difficult optimization problems.

ANNs are an analog to biological neural networks such as the brain. Biological neurons adjust their firing frequencies over time to other neurons in response to the firing frequencies from their input neurons. Some of these neurons are connected to external sensors (such as eyes). Through a learning process, the biological neural networks adjust firing frequencies until an appropriate response is achieved (e.g., ideas, behavior). An ANN replicates (on a very limited scale) the behavior and connectivity among biological neurons in a brain. ANNs adapt their structure based on subtle regularities in the input data. They are robust with respect to error and can find patterns in noisy data in a short amount of time. ANNs offer these advantages over 'brittle' statistical methods that require strict, well-behaved and known error distributions (Fischer and Abrahart, 2000).

ANN application modes

ANNs are very flexible and can be applied in many different modes, including *pattern classification, clustering*, *function approximation*, *forecasting*, and *optimization* (Fischer and Abrahart, 2000).

Pattern classification involves assigning input patterns into one of several prespecified categories. *Supervised classification* is one of the central problems in remote sensing: each pixel must be classified into one of several known land cover classifications based on its spectral signature and perhaps other spectral signatures in its neighborhood. However, traditional methods for supervised classification in remote sensing are failing relative to the vast amount of information available in emerging remote sensing technologies that have high spatial and spectral resolution. Also, integrating ancillary information into remote sensing to aid in classification also increases the complexity of the problem (Fischer and Abrahart, 2000). ANNs have considerable promise as pattern classifiers that can effectively handle the vast and noisy information in remotely sensed imagery and imagery combined with ancillary data (see Foody, 1995; Gong *et al*., 1996; Hepner *et al*., 1990).

In contrast to pattern classification, we often have the case where we do not have any pre-specified categories for the data. Instead, we wish to find natural groupings or clusters of the data based on inherent similarities and dissimilarities. *Cluster analysis* refers to attempts to classify a set of objects into classes or clusters such that objects within a cluster are similar while objects between clusters are dissimilar. Unsupervised ANNs such as Kohonen Maps are a type of neural clustering where weighted connectivity after training reflects proximity in the information space of the input data (see Flexer, 1999). ANNs have been used to cluster river flow data into different event types (Fischer and Abrahart, 2000).

ANNs can also be viewed as a type of universal function approximation technique. Assume a large stream of paired inputs and outputs generated from some unknown noisy function. We can view ANNs as an attempt to approximate the unknown function with an approximate function determined by the pattern of weights in the ANN (Fischer and Abrahart, 2000). Applications of ANNs as function approximations include spatial interaction (Fischer and Gopal, 1994; Gopal and Fischer, 1996; Mozolin *et al*., 2000; Nijkamp *et al*., 1996) and spatial interpolation (Rizzo and Dougherty, 1994).

The problem of function approximation is very similar to the problem of forecasting events over space and time. Formally, the problem is: given a set of n samples of

a time series, predict the value(s) of $n + 1$, $n + 2$, ... (Fischer and Abrahart, 2000). See Hill *et al*. (1996) for more detail on ANNs in time series forecasting. Applications in physical geography include rainfall-runoff responses (Smith and Eli, 1995; Fischer and Abrahart, 2000). Gopal and Scuderi (1995) use ANNs to predict sunspot cycles and solar climate conditions. ANNs are also being used to predict traffic conditions and flow within transportation networks; see Dougherty *et al*. (1994).

ANNs have also proven effective at solving complex optimization problems. This requires transforming a given optimization problem to a neural network representation. Applications to classic optimization problems include the traveling salesman problem, scheduling, and the knapsack problem (see Peterson and Söderberg, 1993).

21.5. CONCLUSION: THE FUTURE OF GEOCOMPUTATION

The future of geocomputation can be summarized in the following sentence: more data, more power, and greater access to both. Data collection and storage costs continue to fall, and computational power continues to increase and will likely increase through the mid-part of the 21st century, perhaps beyond, and the nature of computer interfaces is changing.

Currently emerging are new computing environments that have great potential for geocomputation. Parallel processing has promise in geocomputation and GIS since most procedures can be decomposed into parallel tasks or data streams. However, this decomposition is not trivial due to the overhead involved (see Healy *et al*., 1998; Mineter and Dowers, 1999; Turton, 2000). Grid computing environments software (called *middleware*) can distribute processes across networked computers, exploiting unused resources in these clients. Grid environments can rival the performance of a high-performance mainframe at a fraction of its cost (Armstrong *et al*., 2005).

Over the longer run, computational power should continue to increase at its exponential rate for several more decades. Moore's Law was developed specifically to describe electronic computing based on the integrated circuit. However, Kurtzweil (1999) notes that an exponential increase in computing capabilities has been occurring for over a century. This includes the mechanical computer (1900–1930), electromagnetic computers (1930–mid 1940s), vacuum tube computers (mid-1940s to 1956), transistor computers (1956–1968), and the current paradigm of integrated circuit based computing (1968–2030?). Thus, Moore's Law of integrated circuits is only a special case of a more general trend that may continue through the 21st century. Limits to integrated circuit engineering techniques, as well as the laws of physics, could, however, mean an end to this growth sometime within the next few decades. But even with a conservative estimate of reaching the limits in the year 2030, we are still looking at over twenty more years of continued exponential growth in computing. It is also possible that another computing paradigm may emerge that may shatter these limits. For example, quantum computing would not only shatter these limits, but would also require an entire new theory of *simultaneous* computation.

The nature of the interface between computation, data collection, and information access is also changing. We are currently in an era of ubiquitous or *pervasive* computing characterized by the connection of things in the world through computational devices that are small, lightweight, embedded in other things (such as automobiles,

cell phones, and home appliances) and often Internet-enabled. The continuation of this trend is the *nanoclients* that are extremely small and specialized. Nanoclients include wearable computers, smart dust, and wireless geo-sensor networks. These extremely 'thin' clients combined with very 'fat' high-performance servers can revolutionize geocomputation. Not only do nanoclients allow for ambient geographic data collection, but the environment itself can become a type of computer. Space becomes a metaphor for itself, landscapes or maps become models of themselves, and geographic objects become context-aware and know their own positions and relationships to other geographic objects (Clarke, 2003).

The continuing increase in computing power, capabilities for collecting and storing geo-spatial data, and the merging of computation with the geographic environment will require entirely new modes of thinking about computation in general and geocomputation in particular. While there will always be limits to computing (at least as we now understand it) the phenomena and problems that can be analyzed and understood through geocomputational methods are limited as much by our creativity and imagination.

REFERENCES

Appleby, S. (1996). Multifractal characterization of the distribution pattern of the human population. *Geographical Analysis*, **28**: 147–160.

Arlinghaus, S.L. (1985). Fractals take a central place. *Geografiska Annaler*, **67B**: 83–88.

Arlinghaus, S.L. and Arlinghaus, W.C. (1989). The fractal theory of central place geometry: A Diophantine analysis of fractal generators for arbitrary Loschian numbers. *Geographical Analysis*, **21**: 103–121.

Armstrong, M.P., Cowles, M.K. and Wang, S. (2005). Using a computational grid for geographic information analysis: A reconnaissance. *Professional Geographer*, **57**: 365–375.

Atkinson, P. and Martin, D. (2000). Introduction. In: Atkinson, P. and Martin, D. (eds), *GIS and Geocomputation*, pp.1–7. London: Taylor and Francis.

Axtell, R., Axelrod, R., Epstein, J.M. and Cohen, M.D. (1996). Aligning simulation models: A case study and results. *Computational and Mathematical Organization Theory*, **1**: 123–141.

Balmer, M., Nagel, K. and Raney, B. (2004). Large-scale multi-agent simulations for transportation applications. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, **8**: 205–221.

Barnsley, M. (1988). *Fractals Everewhere*. London: Academic Press.

Batty, M. (1991). Cities as fractals: Simulating growth and form. In: Crilly, T., Earnshaw, R.A. and Jones, H. (eds), *Fractals and Chaos*, pp. 41–69. Berlin: Springer-Verlag.

Batty, M. (2000). Geocomputation using cellular automata. In: Openshaw, S. and Abrahart, R.J. (eds), *GeoCompuatation*, pp. 95–126. London: Taylor and Francis.

Batty, M., Desyllas, J. and Duxbury, E. (2003). The discrete dynamics of small-scale spatial events: Agent-based models of mobility in carnivals and street parades. *International Journal of Geographical Information Science*, **17**: 673–697.

Batty, M. and Longley, P. (1987). Fractal dimensions of urban shape. *Area*, **19**: 215–221.

Batty, M. and Longley, P. (1994). *Fractal Cities*. London: Academic Press.

Beguin, H. and Thisse, J.-F. (1979). An axiomatic approach to geographical space. *Geographical Analysis*, **11**: 325–341.

Benenson, I. and Torrens, P. (2004). *Geosimulation: Automata-based Modeling of Urban Phenomena*. Chichester, UK: John Wiley.

Benguigui, L. and Daoud, M. (1991). Is the suburban railway a fractal? *Geographical Analysis* **23**: 362–368.

Boman, M. and Holm, E. (2004). Multi-agent systems, time geography and microsimulations. In: Olsson, M.-O. and Sjöstedt, G. (eds), *Systems Approaches and their Applications*, pp. 95–118. Dordrecht: Kluwer Academic.

Burrough, P.A. (1993). Fractals and geostatistical methods in landscape studies. In: Lam, N.S.-N. and De Cola, L. (eds), *Fractals in Geography*, pp. 187–121. Englewood Cliffs, NJ: Prentice-Hall.

Cavailhès, J., Frankhauser, P., Peeters, D. and Thomas, I. (2004). Where Alonso meets Sierpinski: An urban economic model of a fractal metropolitan area. *Environment and Planning A*, **36**: 1471–1498.

Chen, M.S., Han, J. and Yu, P.S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, **8**: 866–883.

Clarke, K.C. (1993). One thousand Mount Everests? In: Lam, N.-S. and De Cola, L. (eds). *Fractals in Geography*, pp. 265–281. Englewood Cliffs, NJ: Prentice-Hall.

Clarke, K.C. (2003). Geocomputation's future at the extremes: High performance computing and nanoclients. *Parallel Computing*, **29**: 1281–1295.

Clarke, K.C., Brass, J.A. and Riggan, P.J. (1994). A cellular automaton model of wildfire propagation and extinction. *Photogrammetric Engineering and Remote Sensing*, **60**: 1355–1367.

Clarke, K.C. and Gaydos, L.J. (1998). Loose-coupling a cellular automaton model and GIS: Long-term urban growth prediction for San Francisco and Washington/Baltimore. *International Journal of Geographical Information Science,* **12**: 699–714.

Clarke, K.C., Hoppen, S. and Gaydos, L. (1997). A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environment and Planning B: Planning and Design,* **24**: 247–261.

Clarke, K.C. and Schweizer, D.M. (1991). Measuring the fractal dimension of natural surfaces using a robust fractal estimator. *Cartography and Geographic Information Systems*, **18**: 37–47.

Couclelis, H. (1985). Cellular worlds: A framework for modeling micro-macro dynamics. *Environment and Planning A*, **17**: 585–596.

Couclelis, H. (1988). Of mice and men: What rodent populations can teach us about complex spatial dynamics. *Environment and Planning A*, **20**: 99–109.

Couclelis, H. (1998a). Geocomputation in context. In: Longley, P.A., Brooks, S.M. McDonnell, R. and Macmillan, B. (eds), *Geocomputation: A Primer*, pp.17–29. New York: Wiley.

Couclelis, H. (1998b). Geocomputation and space. *Environment and Planning B: Planning and Design*, **25**: 41–47.

De Cola, L. (1989). Fractal analysis of a classified Landsat scene. *Photogrammetric Engineering and Remote Sensing*, **55**: 601–610.

De Cola, L. (1991). Fractal analysis of multiscale spatial autocorrelation among point data. *Environment and Planning A*, **23**: 545–556.

de Almeida, C.M., Batty, M., Monteiro, A.M.V., Câmara, G., Soares-Filho, B.S., Cerqueira, G.C. and Pennachin, C.L. (2003). Stochastic cellular automata modeling of urban land use dynamics. *Computers, Environment and Urban Systems*, **27**: 481–509.

De Keersmaecker, M.-L., Frankhauser, P. and Thomas, I. (2003). Using fractal dimensions for characterizing intra-urban diversity: The example of Brussels. *Geographical Analysis*, **35**: 310–328.

Dougherty, M.S., Kirby, H.R. and Boyle, R.D. (1994). Using neural networks to recognise, predict and model traffic. In: Bielle, M., Ambrosino, G. and Boero, M. (eds), *Artificial Intelligence Applications to Traffic Engineering*, pp. 233–250. Utrecht, The Netherlands: VSP.

Epstein, J.M. (1999). Agent-based computational models and generative social science. *Complexity*, **4**(5): 41–60.

Epstein, J.M. and Axtell, R. (1996). *Growing Artificial Societies: Social Science from the Bottom Up.* Cambridge, MA: MIT Press.

Esser, I. and Schreckenberg, M. (1997). Microscopic simulation of urban traffic based on cellular automata. *International Journal of Modern Physics C,* **8**: 1025–1036.

Fischer, M.M. and Abrahart, R.J. (2000). Neurocomputing: Tools for geographers. In: Openshaw, S. and Abrahart, R.J. (eds), *GeoComputation*, pp. 187–127. London: Taylor and Francis.

Fischer, M.M. and Gopal, S. (1994). Artificial neural networks: A new approach to modeling interregional telecommunication flows. *Journal of Regional Science*, **34**: 503–527.

Fischer, M.M. and Leung, Y. (2001). Geocomputational modeling – techniques and applications: Prologue. In: Fischer, M.M. and Leung, Y. (eds), *GeoCompuational Modeling: Techniques and Applications*, pp. 1–12. Berlin: Springer.

Flake, G.W. (1998). *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems and Adaptation*. Cambridge, MA: MIT Press.

Flexer, A. (1999). On the use of self-organizing maps for clustering and visualization. In: Żytkow, J.M. and Rauch, J. (eds), *Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Artificial Intelligence 1704, 80–88.

Foody, G.M. (1995). Land cover classification by an artificial neural network with ancillary information. *International Journal of Geographical Information Systems*, **9**: 527–542.

Fotheringham, A.S. (2000). GeoComputation analysis and modern spatial data. In: Openshaw, S. and Abrahart, R.J. (eds), *GeoComputation*, pp. 33–48. London: Taylor and Francis.

Fotheringham, A.S., Batty, M. and Longley, P. (1989). Diffusion-limited aggregation and the fractal nature of urban growth. *Papers of the Regional Science Association,* **67**: 55–69.

Gimblett, H.R., Richards, M.T. and Itami, R.M. (2002). Simulating wildland recreation use and conflicting spatial interactions using rule-driven intelligent agents. In: Gimblett, H.R. (ed.), *Integrating Geographic Information Systems and Agent-based Modeling Techniques for Simulating Social and Ecological Processes*, pp. 211–243. Oxford, UK: Oxford University Press.

Gong, P., Pu, R. and Chen, J. (1996). Mapping ecological land systems and classification uncertainties from digital elevation and forest-cover data using neural networks. *Photogrammetric Engineering and Remote Sensing*, **62**: 1249–1260.

Goodchild, M. and Klinkenberg, B. (1993). Statistics of channel networks on fractional Brownian surfaces. In: Lam, N.S.-N. and De Cola, L. (eds), *Fractals in Geography*, pp. 122–141. Englewood Cliffs, NJ: Prentice-Hall.

Goodchild, M. and Mark, D. (1987). The fractal nature of geographic phenomena. *Annals of the Association of American Geographers*, **77**: 265–278.

Gopal, S. and Fischer, M.M. (1996). Learning in single hidden-layer feedforward network models: Backpropagation in a spatial interaction modeling context. *Geographical Analysis*, **28**: 38–55.

Gopal, S. and Scuderi, L. (1995). Application of artificial neural networks in climatology: A case study of sunspot prediction and solar climate trends. *Geographical Analysis*, **27**: 42–59.

Hägerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association*, **24**: 7–21.

Hare, M. and Deadman, P.J. (2004). Further towards a taxonomy of agent-based simulation models in environmental management. *Mathematics and Computers in Simulation*, **64**: 25–40.

Healey, R., Dowers, S., Gittings, B. and Mineter, M. (eds) (1998). *Parallel Processing Algorithms for GIS*. London: Taylor and Francis.

Hepner, G.F., Logan, T., Ritter, N. and Bryant, N. (1990). Artificial neural network classification using a minimal training set: comparison to conventional supervised classification. *Photo- grammetric Engineering and Remote Sensing*, **56**: 469–473.

Hill, T., O'Conner, M. and Remus, W. (1996). Neural network models for time series forecasts. *Management Science*, **42**: 1082–1092.

Illingworth, V. and Pyle, I. (1997). *Dictionary of Computing*, New York: Oxford University Press.

Kelley, K. (2002). God is the machine. *Wired*, 10.12. Available at www.wired.com.

Kurtzweil, R. (1999). *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. New York: Penguin

Lam, N.-S. and De Cola, L. (1993). Fractal measurement. In: Lam, N.S.-N. and De Cola, L. (eds), Fractals in Geography, pp. 23–55. Englewood Cliffs, NJ: Prentice-Hall.

Lam, S.-N. and Liu, K. (1996). Use of space-filling curves in generating a national rural sampling frame for HIV/AIDS research. *Professional Geographer*, **48**: 321–332.

Li, X. and Yeh, A.G.-O. (2000). Modelling sustainable urban development by the integration of constrained cellular automata and GIS. *International Journal of Geographical Information Science,* **14**: 131–152.

Longley, P (1998). Foundations. In: Longley, P.A., Brooks, S.M., McDonnell, R. and MacMillan, B. (eds), *Geocomputation: A Primer*, pp. 3–15. New York: John Wiley.

Longley, P. (2000). Fractal analysis of digital spatial data. In: Openshaw, S. and Abrahart, R.J. (eds), *GeoComputation*, pp. 293–312. London: Taylor and Francis.

Maes, P. (1995). Modeling adaptive autonomous agents. In: Langton, C. (ed.), *Artificial Life: An Overview*, pp. 135–162. Cambridge, MA: MIT Press.

Mandlebroit, B.B. (1967). How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science*, **155**: 636–638.

Mandlebroit, B.B. (1983). *The Fractal Geometry of Nature,* New York: W.H. Freeman.

McHarg, I.L. (1969). *Design with Nature*, 1st edn. Garden City, NY: Natural History Press.

Ménard, A. and Marceau, D.J. (2005). Exploration of spatial scale sensitivity in geographic cellular automata. *Environment and Planning B: Planning and Design*, **32**: 693–714.

Miller, H.J. and Wentz, E.A. (2003). Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers*, **93**: 574–594.

Mineter, M.J. and Dowers, S. (1999). Parallel processing for geographic applications: A layered approach. *Journal of Geographical Systems*, **1**: 61–74.

Moon, F.C. (1992). *Chaotic and Fractal Dynamics*: *An Introduction for Applied Scientists and Engineers*. New York: John Wiley.

Mozolin, M., Thill, J.-C. and Usery, E.L. (2000). Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. *Transportation Research B*, **34**: 53–73.

Nijkamp, P. and Reggiani, A. (1990). Logit models and chaotic behaviour: A new perspective. *Environment and Planning A*, **22**: 1455–1467.

Nijkamp, P., Reggiani, A. and Tritapepe, T. (1996). Modelling inter-urban transport flows in Italy: A comparison between neural network analysis and logit analysis. *Transportation Research C*, **4C**: 323–338.

Openshaw, S. (2000). Geocomputation. In: Openshaw, S. and Abrahart, R.J. (eds), *GeoComputation*, pp. 1–31, 293–312. London: Taylor and Francis.

O'Sullivan, D. (2001). Exploring spatial process dynamics using irregular cellular automaton models. *Geographical Analysis*, **33**: 1–18.

Parker, D.C., Manson, S.M., Janssen, M.A., Hoffmann, M.J. and Deadman, P. (2003). Multi-agent systems for the simulation of land-use and land-cover change: A review. *Annals of the Association of American Geographers,* **93**: 314–337.

Peitgen, H.-O., Jürgen, H. and Saupe, D. (2004). *Chaos and Fractals: New Frontiers of Science*, 2nd edn. New York: Springer.

Peterson, C. and Söderberg, B. (1993). Artificial neural networks, in Reeves, C. R. (ed.) *Modern Heuristic Techniques for Combinatorial Problems*, New York: John Wiley, 197–242.

Phillips, J.D.. (1993a). Interpreting the fractal dimension of rivers. In: Lam, N.S.-N. and De Cola, L. (eds), *Fractals in Geography*, pp. 142–157. Englewood Cliffs, NJ: Prentice-Hall.

Phillips, J.D. (1993b). Spatial-domain chaos in landscapes. *Geographical Analysis,* **25**: 101–117.

Phillips, J.D. (1999a). *Earth Surface Systems: Complexity, Order and Scale*. Oxford, UK: Blackwell.

Phillips, J.D. (1999b). Spatial analysis in physical geography and the challenge of deterministic uncertainty. *Geographical Analysis*, **31**: 359–372.

Phipps, M. and Langlois, A. (1997). Spatial dynamics, cellular automata and parallel processing computers. *Environment and Planning B: Planning and Design*, **24**: 193–204.

Rizzo, D.M. and Dougherty, D.E. (2004). Characterization of aquifer properties using artificial neural networks: Neural kriging. *Water Resources Research*, **30**: 483–498.

Shen, G. (2002). Fractal dimension and fractal growth of urbanized areas. *International Journal of Geographical Information Science*, **16**: 419–437.

Shi, W. and Pang, M.Y.C. (2000). Development of Voronoi-based cellular automata: An integrated dynamic model for geographical information systems. *International Journal of Geographical Information Science*, **14**: 455–474.

Sipser, M. (1997). *Introduction to the Theory of Computation*. Boston, MA: PWS Publishing.

Smith, J. and Eli, R.N. (1995). Neural-network models of rainfall-runoff processes. *Journal of Water Resources Planning and Management*, **121**: 499–509.

Takeyama, M. and Couclelis, H. (1997). Map dynamics: Integrated cellular automata and GIS through geo-algebra. *International Journal of Geographical Science,* **11**: 73–91.

Tesfatsion, L. and Judd, K.L. (2006). *Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics*, Amsterdam: North-Holland.

Tobler, W. (1979). Cellular geography. In: Gale, S. and Olsson, G. (eds), *Philosophy in Geography*, pp. 379–386. Dordrecht: D. Reidel.

Turton, I. (2000). Parallel processing in geography. In: Openshaw, S. and Abrahart, R.J. (eds), *GeoComputation*. pp. 49–66. London: Taylor and Francis.

Wentz, E.A. (2000). A shape definition for geographic applications based on edge, elongation and perforation. *Geographical Analysis*, **32**: 95–112.

White, R.W. (1990). Transient chaotic behaviour in a hierarchical economic system. *Environment and Planning A*, **22**: 1309–1321.

White, R. and Engelen, G. (1993). Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns. *Environment and Planning A*, **25**: 1175–1199.

White, R. and Engelen, G. (1997). Cellular automata as the basis of integrated dynamic regional modeling. *Environment and Planning B: Planning and Design*, **24**: 235–246.

White, R. and Engelen, G. (2000). High-resolution integrated modeling of the spatial dynamics of urban and regional systems. *Computers, Environment and Urban Systems,* **24**: 383–400.

Williams, G.P. (1997). *Chaos Theory Tamed*. Washington, DC: Joseph Henry Press.

Wilson, A.G. (2006). Ecological and urban systems models: Some explorations of similarities in the context of complexity theory. *Environment and Planning A*, **28**: 633–646.

Wolfram, S. (1984). Universality and complexity in cellular automata. *Physica D,* **10**: 1–35.

Wong, D.W.S. and Fotheringham, A.S. (1990). Urban systems as examples of bounded chaos: exploring the relationship between fractal dimension, rank-size and rural-to-urban migration. *Geografiska Annaler*, **72B**: 89–99.

Wu, F. (2002). Calibration of stochastic cellular automata: The application to rural–urban land conversions. *International Journal of Geographical Information Science*, **16**: 795–818.

Xie, Y. (1996). A generalized model for cellular urban dynamics. *Geographical Analysis,* **28**: 350–373.

22

Applied Retail Location Models Using Spatial Interaction Tools

Morton E. O'Kelly

22.1. RETAIL LOCATIONAL ANALYSIS[1]

22.1.1. Spatial retail location

The demand by consumers for retail goods and services is a function of the attributes of the commodity, household income, and other factors such as home ownership status. For example, a home improvement store is likely to target a market with a housing stock that has lots of possibilities for repair, upgrades, and remodels. Both home-owning and renting populations might yield adequate density of demand, but the *effective* demand for goods and services by homeowners is much more likely to be attractive to this particular service. An electrical supplier would have a demand for services also, but for that business it could be some combination of over-the-counter sales (light fixtures) and more substantial electrical equipment sold to contractors and builders. A business with a traditional central market place location (in an older mixed use inner city neighborhood for example) might conceivably want to branch out its locations to catch the growth in the suburbs and even the outlying communities in the hinterland of that main market. In fact there are so many different ways to imagine the dynamics of retail site location that there is a real need for a general purpose simulation tool that might enable the estimation of the merit of various growth proposals (Baker, 2000; Munroe, 2001). In all these cases, it is important to have an accurate

estimate of the spatial distribution of effective demand as arising out of a combination of preferences, and disposable income.

Central place theory has long held that there is a hierarchy of goods, from frequently demanded inexpensive items to high-end expensive goods. There is both a higher spatial frequency of demand for (and provision of) the so-called lower-order goods, and a corresponding scarcity on the landscape of higher-order goods. Thus for every Mercedes or Lexus dealership in the city there might be numerous Ford and Toyota dealerships. The higher the order of goods provided, one assumes that there is a wider market scope required to provide sufficient demand to cover the operating costs of the business (the so-called threshold). Similarly, the higher-order goods, because of their relative scarcity on the landscape, require longer trip lengths; the break even calculation for the retailer is whether the spatial extent of the market required to cover costs is matched by a corresponding willingness of consumers to travel to the center for the goods (see the classic study by Berry (1967)).

Inexpensive 'low order goods' are sometimes sold in combinations with higher priced items from superstores that do not necessarily have a small range: they can in fact be attractive over a large distance, provided the assortment and price point allows the large agglomerated retailer to undercut the smaller more widely dispersed providers of retail services. This formula is used by Wal-Mart or other 'big box' retailers; they have a large assortment of goods, and price points that are competitive, and locations that in themselves act as a magnet for spatial interaction (Munroe, 2001). Customers travel to stores and therefore the spatial interaction of the purchasers must be recognized as an important behavioral factor. The retail and trade area service location problem requires knowledge of what customers want, where they are located, and that they have income that covers the price and market segment of the goods. In common with many levels of retail operation, many of the most successful chains study a massive amount of geo-demographic profile data that enables a rich portrait of customers and consumer behavior to inform the merchandize and market planning of their operations.

22.1.2. Consumer demand and behavior[3]

Measuring the total income and pool of expenditure is accomplished by combining a count of households by geo-demographic cluster (e.g., Claritas PRIZM, MapInfo PSYTE, ESRI Tapestry, AGS Mosaic etc.), the index value for each group (m), the penetration rate and some index of average per household expenditure.[4] To calculate the potential pool of expenditure for the zone i and commodity 'c' a formula such as the following might be used:

$$O_{ic} = \sum_{\text{over all groups}} N_{im}\, y_{mc}$$

where O_{ic} is the demand in zone i for commodity c, N_{im} is the number of group m households in zone i, and y_{mc} is an expenditure rate per household cluster m on commodity c.

From this aggregate, demand shares allocated *to a particular store* have two components: on the one hand the share is smaller for the more distant competitive stores (holding other factors constant), but additionally it is felt that the demand for a store *increases* with the accessibility of that origin zone to any shopping destination. Zonal accessibility, and hence aggregate demand is a function of where the stores are located, and so unlike conventional

location models, we should not treat demand as an exogenous factor (O'Kelly, 1999). While it would be naive to say 'build it and they will come', it is certainly reasonable to think that the provision of retail services can induce demand for that service that would otherwise be allocated to other discretionary uses. Some insight and market-based intelligence is needed to capture the correct demand parameters and sensitivity to locational access. The basic accessibility of each zone can be predefined, and the demand in the immediate area of a new potential store opening can increase, as a result of improved accessibility. One practical estimation approach that can be effective is to have a variety of alterative sources of judgment (like the so-called 'Delphi' method, and a variant of the judgmental methods advocated for several years by Seldin (1995)) with perhaps one figure coming from an estimate of *per capita* expenditure and saturation, one coming from *pro forma* estimates of expected sales per square foot and yet another estimate coming from an experienced local commercial real estate professional. The best model is likely to use some aspect of these data as controls on the judgmental estimate. In other words, no analyst will simply apply a sales per square foot figure to an arbitrary new built store and say that the expected sales are a product of the coefficient and the store size. Much more likely is an analysis that takes the current sales situation of the competition into account and then projects how much of these existing sales can be captured by the new proposed location. Even more significant is recent research that has shown that whatever the *general* relationship between the variables, the strong likelihood of spatial variability in such a relationship ought to be taken into account (Fotheringham *et al*., 2002; Rust and Donthu, 1995). Thus, if a cross-sectional regression analysis provided evidence of a coefficient of say $30 weekly sales per square foot, then a spatially varying parameter might trend significantly with location given the socio-economic patchwork of the city, by analogy with a similar argument in the context of house prices (Fotheringham *et al*., 2002).

One way to make operational estimates is through spatial interaction models. These models are the topic of this chapter, which covers a variety of models largely inspired by several years experience as both an applied and as a theoretical exploration of retail sales and interaction.

22.1.3. The role for models

Among the most basic general questions for spatial interaction modelers are the following: Where do the customers come from? What are the spatial interaction patterns governing the distribution of distance and attraction parameters? What is the probability that a customer at i patronizes a store at j? Conditional upon the location of i what is the probability of being a customer of destination j?

Example

A grocery store has an upper income target consumer. Their research shows that these are very likely to be loyal customers of the produce and fresh foods departments (which in turn are highly profitable assuming that stock can be turned over rapidly to avoid waste/spoilage). In seeking new store locations, where are there sufficient pockets of un-met demand among this target population? In the analysis of existing 'own-brand' stores, which may or may not be currently well-located *vis-à-vis* the standard customer profile, is there any need to modify their type of store to meet consumer needs?

These kinds of questions can be answered with spatial models. Before we get into the

details of how to formulate and apply such a model, it may be very helpful to get a preview of some of the uses to which a model might be put. One common usage is in 'impact analysis.' With a fitted model, purporting to describe the allocation of consumers to demand centers, we can estimate impact on remaining stores if a branch is to be closed, or indeed if we open a new one. Both of these changes have impacts across the system of stores, but of course the 'first law of geography' (Tobler, 1970) which holds that things are more highly interrelated when they are in close proximity, leads us to expect that the impacts are greatest on the centers and competitors closest to the site of the change.

Other uses for fitted models in locational analysis include assessment of the desirability of overhauling various stores or facilities. The applied retail analyst is often asked to estimate the impact of a change on the expected sales of the store: thus having a model which has as its 'independent variables' some measures which can be adjusted to reflect the new attraction of the store can be useful to estimate the change in the retail trade area, expected sales, and so on. By estimating a well-fitted model to these data, we replace the specifics of the data instance with a model that has 'effects' – these are systematic influences on the trends in the levels of spatial interaction, and are likely to include roles for distance and retail attraction (typical basic variables in SI models – see Guy (1991)). In more elaborate settings these models can also include many other independent variables (see especially the Multiplicative Competitive Interaction Models MCI – Nakanishi and Cooper (1974)). Once these models are fitted, the analyst can then dial in various changes in the driver variables, and assuming that the model is reasonably robust to changes in these data, the impact of the changes on the expected sales and interaction levels can be determined.

Models can also be used as a tool in assessment of complex strategic questions. For example, a chain that is considering opening a new branch in a growing suburb might be faced with the question of whether to keep an existing older store in a nearby location. The question is then one of strategy: do stores *A* and *B* together make a better combined profitable solution than the option of closing *B*, presumably giving *A* an even greater new opening sales level, but possibly exposing the chain to the risk that a competitor might take the abandoned site? Not only does the decision hinge on the aggregate sales of the various combinations of open stores but it also must answer questions about the probable impact of competitors. Retailers engage in *strategic behavior*, and open or close locations as part of a system of decisions; such analyses often include issues of pre-emption and blocking competition, and beating competitors to the punch in new areas of expansion (Ghosh and Craig, 1983).

Models are also useful in assessing ongoing measures of store performance and may be used in this way as an early warning of emerging shifts in the market. Assuming that the chain can collect data throughout its system on the performance of each store, and some appropriately calculated variables to describe the stores site and situation the analyst can embark on the kind of 'analog' assessment made popular in the early days of quantitative analysis. This method, in its modern guise, uses the stores sales (as a dependent variable) and a selection of measurements of the trade area characteristics, and develops a multiple regression model to assess the expected (or predicted) sales vs. the actual observed levels. Fundamental to this operation is a meaningful definition of trade area: it makes no sense to include measurement of the 'attributes' of areas far away from a store, if indeed it is known that few if any shoppers

come from that area. So, in other words, the *measurement of the trade area of the store* becomes the first and most important operation. There is no hard and fast consensus on how to define a trade area, and much more will be said on this matter later. For now, suffice it to say that the trade area *could* be objectively defined as an area within 5 min drive time of the store. That leads to a computation of the demand that exists within that area, and that could be one of the independent variables. (Clearly if we use more sophisticated definitions of trade areas, the trade area demand calculation would have to be re-computed.)

Independent variables collected for all the stores are saved as the columns of a table. GIS is especially helpful to calculate features of trade area and give a quantitative descriptive nature of a trade area. The dependent variable is the actual sales performance: there is often a challenge obtaining these data (i.e., weekly sales) in academic research; but it is important to know how these data would be used in an applied case study.

Basing store closing decisions on this kind of result places a lot of faith in the fitted model (and so the importance of regression diagnostics, measures of goodness of fit, and significance levels on the estimated coefficients). What looks like an underperformer may not actually be an instant argument for store closure: for instance a store is projected to draw $400,000 per week, but actual sales come in at $350,000 (i.e., $50,000 below the regression line). While all might agree that it could be doing better (i.e., it is performing *below* potential) there may be good reasons that the store has not yet reached its full potential. It might be under attack from particularly aggressive competitors, be poorly managed, or it might be built 'over-sized' in anticipation of further population growth in the area. The store may suffer from a depressed regional economy, and so a chain may consider shutting it down or attempting to reinvigorate the system by investing a lot of money into the regional advertising campaign. It all boils down to choices, and these choices are best informed by analytic models.

The ease of obtaining a good fit to the model will clearly vary across sectors. Department store sales volumes are notoriously difficult to predict, in that their aggregate sales volume is a combination of the various heterogeneous departments, and the extent of competition for specific categories in these stores could very well vary in an unsystematic way across locations. On the other hand goodness-of-fit for convenience related stores such as grocery chains are likely to be quite acceptable, in that there are a few predictable variables that are very highly correlated with the aggregate performance of the store. For example, the store's size, its population base, and the immediate competitive environment undoubtedly account for the bulk of the store-to-store variation in sales levels. Thus, it is expected that the coefficient of store size, and population and competition will be significant, and that the resulting fitted model will have a strong R-square.[5] Refinements to the model to include regional dummy variables and other more precise measures of target market demand (through surrogates such as parking studies, or traffic flow) are likely to help to improve the model.

Some sectors lend themselves readily to analysis by multivariate regression models (grocery stores) but others require a different approach. If a shoe store, book store, or branch of a chain of clothing stores is typically located in shopping centers, then the analyst might use the center as a surrogate for the size of the market in which the individual store is located (see also Prendergast *et al.*, 1998). Similarly if a chain of this type is planning to enter a new regional market, it could very well limit its attention to

the shopping centers. This type of work is useful because it is frequently necessary to manage thousands of location across many areas/regions.

It is hard to get information on gross sales (what is also called 'turnover' in the British literature) in academic case studies, though practitioners and consultants can of course gain access to their client's data as part of their confidentiality agreement. Many of the ideas in this chapter have been framed as a result of real world experience. In practice, one has access to lots of data; in theory one might have to learn these techniques in a data vacuum, recognizing that the proprietary data would become available to a consultant doing these analyses for a private sector client. This perhaps accounts for the lack of precision in the published literature – a lot of literature in retailing location modeling is quite imprecise mathematically – and the details are often not published in a way that makes verification and validation easy.

22.1.4. Consumer choice

The probabilistic assignment of consumers to retail destinations can be formulated as a production constrained spatial interaction model:

$$P_{ij} = A_i O_i W_j \exp(bC_{ij}).$$

Such models calculate the probability that a user at a specific origin location will select one from a number of available alternative attractive destinations. If these destinations are shopping centers, for example, the attraction of those centers can be represented by a measure of their total retail square feet of selling area. Once a calibrated production constrained spatial interaction model has been formulated for a specific set of destinations, the estimated table of such flows provides an idea of the likely inflow to each of the unconstrained destination trip ends:

$$D_j = \sum\nolimits_i P_{ij} = \sum\nolimits_i A_i O_i W_j \exp(bC_{ij}).$$

The production constrained model leaves the amount and type of flow arriving at each center or store open to calculation. With such calculated inflows, the analyst has an access to a predictive model for the likely composition and size of any centers for its capture area. Think of a column of the spatial interaction matrix that leads to a specific destination as a listing of the contributions to that particular destination. Of all the flows that arrive at the destination, we may estimate the percentage that comes from each one of the surrounding regional sources. From all of those, the core or primary contributors may be determined by sorting the origins from largest to smallest and cumulating their contributions until arriving at a subset that contributes a very significant fraction of the total business of the store of interest. This is none other than Applebaum's (1966) concept of primary trade area being the region from which a particular store draws a high percentage (say 75%) of its business.

22.2. ANALYSIS WITH RETAIL TRADE AREA MODELS

22.2.1. Spatial interaction[6]

Spatial interaction models in general assume that interaction is determined by the attraction of the alternative facilities and by the distance separating the consumer from those alternatives. Huff (1962, 1963, 1964) and Lakshmanan and Hansen (1965) are credited with developing specialized 'retail' variants of the spatial interaction

based allocation model. From an operational perspective, Huff introduced a practical approach to defining the 'attraction' of a center as the amount of floor space, rather than the population of the surrounding area as was commonly used in previous models. This opened up the interpretation of attractiveness and allowed it not only to be determined by a number of variables (e.g., number of functions, parking capacity, etc.) but also allowed attractiveness to be treated as an independent variable that could be estimated in its own right. Another major operational consideration was that Huff fitted the exponent for distance in trip-making behavior (the influence that distance has on a consumer's store choice) to particular circumstances. Finally, he introduced a balancing term that constrained the sum of individual or zonal travel or sales to fit within an overall travel or sales limit.

With respect to the attractiveness or drawing power of a facility, Huff's use of retail floor space has been widely adopted and adapted to include other important characteristics. Most important, though, this model demystified the idea of drawing power or attraction and allowed its direct estimation by focusing on the weight associated with it. Nakanishi and Cooper (1974) were particularly effective at utilizing Huff's probabilistic choice framework and operational perspective to develop a linearization procedure for direct estimates of attractiveness. The MCI model is one of the best tools available for the allocation of consumer demand to facilities. The main advantages of this model is that it can incorporate a variety of attributes of the facilities under consideration by the consumer, yet it is easy to estimate. In cases where more data on the influence of various store attributes are available, the MCI model is apt to provide a more accurate estimation of market share than the original Huff model.

With spatial interaction models, then, facilities no longer have a well-defined geographic market area. Instead each store's market area is a probabilistic surface that shows the probability of a customer from each small geographic area patronizing that facility. The exact nature of this probability surface depends on the parameters of the spatial interaction model. Incorporating spatial interaction models into a location–allocation model represents the state of the art in modeling retail site selection.

22.2.2. Primary trade area

Imagine a store attracting customers from surrounding census tracts or city blocks. Such data have long been analyzed by proponents of the applied school of retail trade area analysis (Applebaum, 1966). As a starting point, examine the distribution of the customers of a particular store, with regard to their origins. If the store has a weekly volume of V, then the customer distribution is used to spread around that demand to the originating areas, in proportion to their draw of customers. That spatially distributed demand in turn can be compared to the potential pot of money that exists in those zones available to be spent somewhere, in order to compute a measure of store penetration of the market. From the data, the top 75% (say) of the sales area may be devised, followed by the next 20% and the rest (all these are hypothetical numbers). Unless some added spatial constraints are added, it is important to note that it is not essential for the top contributing area to a store to be compact (having for example disconnected outliers). Analytically, the primary trade area, P, is defined such that $\sum_{i \in P} P_{i|j} = 0.75$ and the secondary trade area, S, is defined such that $\sum_{i \in S} P_{i|j} = 0.20$. The remaining or 'tertiary' trade area, captures the remainder of the customers, often sparsely dispersed over a very wide area. For most practical purposes in the convenience sector, 'tertiary areas' are

irrelevant to routine operations. On the other hand, significant shopping centers drawing from a large region may well have to treat the marginal sales to the edge of their tertiary area as significant 'icing' on the sales forecast, and may in fact be the key to understanding top-performing locations.

Retail executives are especially interested in market share, strength versus direct competitors, and in the yield of customers from a pool of potential sales dollars. It seems that the only thing worse than a store that has a small sales level is one with a large volume but under-performing its projected potential! These analyses are directed to the question: how well are our stores capturing the market? Are we leaving potential sales untapped? Or are our competitors out-maneuvering us? Penetration of the market area hinges on an assessment of how much demand is available there, and how much our particular branch is capturing.

22.2.3. Characterization of the demography of the trade area

The attributes and weights of demand from the particular types of respondents in the trade area can then be recovered. Say, for example, that the numbers of household in the various tracts that have particular levels of household income are given. Many useful statistics can be computed from these data. Among these are the expected values of customer characteristics over the primary, secondary, and tertiary trade areas respectively. For example, if we have a defined area that encloses the primary trade area, and the total volume of expenditure in that area is X, then the total volume attracted to the store of interest from within that same are is Z, the ratio of X to Z is very useful information about penetration of the market. These analyses provide the tools to diagnose practical issues in the trade area's effectiveness, for example, by indicating untapped sales potential, the need for more intense marketing, or special circumstance arising from unique factors (ethnicity, mobility, etc.).

22.2.4. Connecting retail location models and competing destinations

Retail locational analysis is frequently carried out with the aid of spatial interaction modeling. Many features of the trade area are derived from calculations based on either actual customer origins (from a survey) or from a model of such a distribution that has been fitted from observations. In either case assume that the probability that a customer in area i shops in store j is given by P_{ij}. This joint probability can be further manipulated to give $P_{i\,|\,j}$ and $P_{j\,|\,i}$, respectively these are:

$P_{i\,|\,j} = P_{ij} / \sum_i P_{ij}$ is the conditional probability that a customer who shops in j originates from i,

and:

$P_{j\,|\,i} = P_{ij} / \sum_j P_{ij}$ is the conditional probability that a customer from origin i shops in zone j.

It is this later probability that is highly useful as it allows a prediction from a given zone i, of how much traffic or business might be expected to arrive at a destination in zone j, and this of course can be applied either to pre-existing stores (to check model fit and validity) as well as the use of the

model to forecast the likely patronage of a new or proposed location at j. In that these probabilities are analytically derived from data that are exogenously available (travel times, demand expenditure parameters, and so on) they are quite easily manipulated to give forecasts of 'what if' for cases where there are expected changes in the data or the parameters. This kind of sensitivity analysis can provide a useful cross check on the validity of the model – for example, a sensitivity analysis should predict changes that make sense. Further, extreme values of the parameters often provide consistency checks in that the model collapses to other easily recognized forms in these special circumstances: thus a model with a distance decay parameter collapses to an all-or-nothing nearest center allocation model in the case that the beta parameter is driven to the extreme value. In this case the trade area should take on characteristics such as that seen in the 'Voronoi' diagram or Thiessen polygons.

In macro spatial analysis (e.g., at the scale of interregional interactions) the peripheral areas have, by definition, lower access to the dense cluster of the urban core. So, for a resident of the periphery the number of competitive alternatives in short range is comparatively small, and according to the theory of competing destinations (Fotheringham, 1983), the demand is therefore spread over few alternatives (hence is not divided up so thinly). It would be expected therefore that interaction levels over short distances are enhanced (and comparably the interaction over the longer distances is spread thinly, and hence the slope of the flow vs. distance curve is steeper than it would be expected to be, absent a spatial structure effect). At macro scales then the large beta for peripheral zones results from mis-specification, and does not correctly imply that there are larger distance decay impacts for peripheral residents; in fact, once the mis-specification is corrected, the expectation might be that peripheral residents might show a willingness to travel to distant alternatives at a rate that exceeds those of the comparatively well served central residents.

This notion of a process at one density regime being adapted for other situations was nicely foreshadowed in Berry's (1967) classic work on commercial centers when the expected sales territory size was contrasted in low density rural Iowa with the more commercially dense built up areas of Chicago. Thus there is some interest in whether this theory might be adapted to a more dense urban retail scenario. In the retail scenario the central or core resident has lots of alternatives within short range, and these can provide opportunity for multipurpose trips and shopping on a scale that combines multiple activities. As Eaton and Lipsey have shown, such retail agglomerations then gain more from their collocation than they lose from the presence of intensified competition. Thus the theory of competing destinations developed at a primarily interurban scale might be refined for the case of flows within an urban area, and indeed the opportunity to make multi-purpose trips to clusters of shops in a city might lead to an expected agglomeration effect: what we might coin the 'cooperative destinations' effect arising from spillovers in retail demand (see early theory of Eaton and Lipsey, 1982).

22.3. CALCULATIONS

22.3.1. Data issues

An interesting aspect of retail trade area analysis is that the most commonly collected data (choice-based samples) are not especially well suited to direct manipulation in calibration (see a series of papers on choice based samples by O'Kelly (1999) and Ding and O'Kelly (2008)). Choice based data

from frequent shopper cards at the point of sale or from check based data can tell us the distribution of actual demand around a current store. Clearly the interest in these data from a predictive point of view is to be able to use them to devise some origin based parameters such that the trade area attributes that determine the store success/failure can be studied and translated into parameters that can predict how a proposed new location (assuming that represented stores provide a decent analog for the new operation) might be expected to perform. One could expect to take data about existing operations, and develop a list of those parameters of the trade area that are expected to correlate heavily with good retail performance. The interaction model is simply an improved way to gather data and summarize standardized aspects of these trade areas to provide data about the branches. In applications, these data can then be entered into regression or other models to determine the different aspects of the trade areas that are especially highly correlated with successful operations.

An important step in managing a retail trade area data set is to understand the scope and reach of the center to the areas surrounding the store. In fundamental economic geography we learn concept of the *range* of the good: this is the maximum distance a customer would be willing to travel to reach the store. This maximum radius or reach has relevance for the concept of spatial interaction and trade areas as there is clearly no necessity to include demand from a place that is so far from the store as to be unable to reach that store's trade area. Distance impedance and maximum travel radius are critical to the accurate specification of gravity models. In the case of a maximum travel radius, one has to be sure to set up a spare or 'dummy' destination to allow for demand that has no feasible option within range to be 'parked' there pending either some additional site, or some relaxation of the maximum range.

Very large energy costs cause a contraction in peoples' willingness to travel long distance or make excess discretionary trips; instead one would expect two countervailing forces: to make a smaller number of multipurpose trips to major agglomerations would serve to support the development of a small number of heavily clustered mega malls; on the other hand the smaller willingness to travel might cause a stronger tendency to use the closer alternatives and activate the incentive to build a series of small decentralized regional centers. This trade-off between agglomeration and convenience is an interesting empirical question.

22.3.2. Determination of market effectiveness and penetration

The idea in retail interaction modeling is to use a probabilistic estimate of the demand originating in each sub-area, and its likelihood of being spent at a particular store of interest. It is convenient, though perhaps increasingly less realistic, to assume that the pool of available money is all allocated to 'bricks and mortar' stores, and that the demand is a simple function of the population, its income, and expenditure habits. With that assumption it is possible to take readily available census expenditure data and predict how much would be available for particular product categories in each micro-demographic area. Such micro marketing data have been used with great precision by the package goods industry, car industry, banks, and retailers in general. These applications represent one of the most powerful uses of the gravity model. Some industry specific intelligence is needed with regard to the reasonable range of potential destinations from the point of view of an origin. This is because it is necessary to be able to make an all-inclusive list of the

probabilistic choice sets that exist or that might provide opportunities for the shoppers to make choices. To adapt this base case to the more realistic case of alternative non-spatial alternatives (in competition with conventional alternatives), we need to be able to estimate leakage from an origin area to electronic, catalog, and on line purchases. From the retailer's point of view at a specific location, it is necessary to be able to circumscribe the potential originating zones from which the trip makers might be attracted. For a convenience-oriented store like a supermarket, one can imagine a reasonably compact service area. For department stores, or retailers co-located with attractions that can draw from farther places (think of Mall of America as a destination), it is perhaps a little more difficult to know the universe of the attraction, and hence difficult to make computations of the share of the attraction provided for by local or further away origins.

22.3.3. Performance assessment of existing stores

It is reasonable to assume that the primary trade area, which accounts for say 70% of the branch business is key to characterizing the stores potential customers. In an applied context, working for a retailer, we would *need them to provide us with* some measure for each store of the total retail volume and perhaps some breakdown by product line or class, and also an indication from the stores perspective if the chain regards the branch as successful. With the sales data we can produce measures in the surrounding zip codes for sales/household and this could give some indication of penetration rate. From that we can characterize the trade area make up for the store (Hispanic, middle class, etc.). While these data are a very big part of the puzzle, what *we cannot* do with such data alone is to talk about the residents in a particular subareas and their probability of being a customer. For those who are customers (and for those who are not) we need some additional way to measure reasons as to why or why not. To get at these added questions we either need prior theoretical expectations, *or* to employ a survey to ask residents in a residential area about their reasons for shopping or not shopping at our chain. As surveys tend to be very expensive, a controlled theoretical choice experiment is perhaps a worthwhile future framework for such destination choice problems (see Eagle, 1984).

From these two sources of data detailed intelligence about the trade areas of the various branches can be accumulated and the results used to characterize the stores; if there are added data from the retailer about which stores are under- or over-performing, we could do some correlation analysis, or perhaps data envelopment analysis (Donthu and Yoo, 1998) which allows a gauge of performance *vis-à-vis* peer benchmarks.

22.3.4. Impact assessment

One of the most frequently asked question from an applied perspective is to determine the loss of sales at existing stores to new entrants or competitive analysis for the diversion of existing dollars to the store of interest either from ones own chain (cannibalism) or preferably from competition.

Impacts of changed conditions are quite well accommodated by the gravity model, because the difference between two scenarios may be quite instructive. The impact of new store k on existing store j, from the point of view of zone i, is measured as:

$$I_{i,j,k} = (P'_{ik} / \sum_{\text{over all new sites}} P'_{ik})[P_{ij} - P'_{ij}]$$

where $I_{i,j,k}$ is the impact of new store k on existing store sales in zone i, P'_{ik} is the new allocation to center k from zone i, and P_{ij} is the allocation to center j from zone i.

The types of scenario that can be handled using the methodology are as follows:

- analyze the trade areas of current stores (run with just fixed locations)
- pick sites from candidates (run with fixed and potential locations)
- re-consider current sites (make currently fixed sites flexible or optional)
- examine specific proposed sites (lock in particular new sites)
- analyze specific closings (lock out particular site and see what happens)
- analyze the opening of a known competitor (add fixed locations).

All of these versions of the problem have been deployed in practice with good empirical and quantitative results.

22.3.5. Temporal and seasonal variations in trade areas

Clearly, the volume of business is not simply related to the local demand, and the seasonal adjustment for external visitors is something that would have to be taken into account in developing accurate sales forecasts. Imagine a seaside resort such as Hilton Head, South Carolina: its sales would be quite variable over the seasons, in a cycle tied to the peak tourist demand in the northern winter. One way to do this is to examine sales records and develop a set of monthly seasonal adjustments. Whatever the base level of demand, the modeler could then devise factors to scale up or down the sales for specific months.

A simple time series model, with a set of monthly or seasonal dummy variables can be used to make an empirically fitted set of correction factors. Another way that trade area models need to be corrected is for the excess in demand that often accompanies a new store opening as the novelty of that location is added to the mix of existing stores and, at least initially, there may be large incentives or advertising efforts made to attract customers. Clearly, it would be advisable to temper these initial sales figures with some kind of decay or dilution effect that would bring the stores sales into alignment at moderate levels (see Kaufmann *et al.*, 2000). Rules of thumb abound in this area, and equilibrium sales after opening may settle down to say 60% of the initial week sales.

22.4. LOCATION ALLOCATION MODELS

22.4.1. Introduction to location allocation models

The use of the location allocation model in retail site selection has greatly advanced over the past 15 years. Examples include the use of interaction models to develop optimal site locations for stores in a variety of different types of retailing including supermarkets, department stores, big box retailers, and retail banking.

Successful use of these models led to their commercial acceptability and widespread adaptation in retail outlet location study (see the Thompson site selection book (Buckner, 1998)). Commercial examples in Britain include the G-MAP package (see Longley and Clarke, 1996). Specialized programs in business-GIS packages now provide routine access to methods that were previously only obtainable in customized software and

research publications.[7] This diffusion of the innovation of retail trade area analysis from specialized journals such as *Environment and Planning A*, into many applied sectors has been a major success for analysts. These models serve as a critical underpinning of the site selection analysis that goes into many large format stores in almost every urbanized area in the U.S. and Europe. The reason that such models are widely used is that they are essential to the rapid 'pro-forma' evaluation of numerous site proposals. The models provide the kinds of rapid computations that would ordinarily have taken a great deal of manual computation; and certainly when a chain is screening as many as 10 sites for every actual chosen location, the need for rapid analysis is obvious. For example the early studies by Applebaum (1966), directly predate the computation of trade area penetration models that may now be made using spatial interaction models.

One of the goals of this chapter is to provide the analytical background to the models that are now a commercial fact of life for retail analysis. The idea that a model of retail attraction could be deployed as a model for retail site location is an extension over the simple, earliest work in central place theory, where consumers were assumed to patronize closest centers (see also Ghosh, 1986). In turn the central place approach defined a region in close proximity to the store from which it would be reasonable to expect that the demand would be assigned to that particular store. Following a large amount of study of consumer behavior indicating dispersal of choices over many alternatives beyond just the most convenient (Clark, 1968; Hanson, 1980; O'Kelly, 1981), market researchers and others devised more precise means of estimating likely consumer behavior. The deterministic 'all-or-nothing' allocation of demand to the nearest or most convenient branch is no longer a necessary or indeed acceptable simplifying hypothesis about spatial behavior. Instead, we now expect that consumer behavior may be examined with the same tools that econometricians have devised for the analysis of discrete choice. Databases in turn provide a wealth of data. Geographers have derived a representation of consumer behavior with a model that locates services; this involves a breakthrough in the use of spatial interaction models. The key idea was to replace the nearest center assignment of customers in central place theory, with a more realistic gravitationally based estimate of likely destination choice (O'Kelly, 1987). Thus, the customer might have a certain probability of visiting a large center that is a bit further away than a small center close to the consumer. In gauging these trade-offs, the model makes a carefully calibrated estimate of the impact of size and distance on the consumer's willingness to travel to particular destinations. Once this calibrated model is available to us, the analyst can propose specific new site locations and gauge the expected level of consumer patronage at those sites. So called 'turnover' or retail sales volume is a critical first step in the analysis of any commercial property deal as the sales levels helps to support the go/no go decision on rental, lease, re-model, or closing.

Location–allocation models generally involve the simultaneous selection of locations and the assignment of demand to those locations in order to optimize some specified objective or goal (usually to maximize market share or profit; see, for example, Craig, *et al.*, 1984). These models have several advantages. They can determine the optimal (or near optimal) location of several stores simultaneously by systematically analyzing the system-wide interactions among all stores in the market area. They are capable of utilizing a wide range of objectives that could be used in siting stores. In addition, the models

are flexible in that they can incorporate the behavior of retailers, consumers and/or the retailing environment. Finally, heuristics are available for these models which provide good (optimal or near optimal) solutions and yet are easy to implement. The use of location–allocation models typically involves empirical research to determine the important store attributes for the population within the market area and a mathematical model to determine the optimal locations for retail outlets based on the pattern of market demand, store chains and existing competing outlets.[8]

Even though it is recognized that many consumers engage in multi-purpose, multi-stop shopping, models of multi-purpose shopping behavior have *not* been thoroughly integrated into facility location analysis, though early efforts by O'Kelly (1981, 1983a, b) have been recently reconsidered as the basis for new location models (Leszczyc *et al*., 2004). So the assumption of single-purpose trips is made in order to devise practical (usable) store-location models. Nevertheless, the fact that our analysis is primarily designed around shopping center destinations ensures that the attraction of a destination for a specific store is partly determined by the attraction of the cluster of stores as a whole.

There are several types of retail location models in the literature. Some representative examples include models which combine location–allocation with spatial interaction (for example, the MULTILOC model by Achabal *et al*., 1982); models which can deal with multiple objectives (for example, Min, 1987); models that consider the uncertainty inherent in the retailing environment (such as the scenario planning model by Ghosh and McLafferty, 1982); and models which involve the decision maker in the decision-making process (for example, the STORELOC model by Durvasula *et al*., 1992). No one model is capable of handling all the important aspects of retail site selection which must be addressed in order to provide the decision maker with the best set of locations for any particular market area in which the stores will be located.

Some aspects of these models are developed in more detail in the following section.

22.4.2. Retail location models and spatial interaction

MULTILOC (Achabal *et al*., 1982) was one of the first location–allocation models to simultaneously locate more than one store. The model optimizes the location of stores using the knowledge that consumers will choose among the alternatives according to a probabilistic interaction model (the MCI model). Such models maximize total profit for a retail chain (or a single store) after subtracting the fixed costs of establishing a store at the determined location (i.e., location-specific fixed costs). It has later been given a more mathematical treatment in O'Kelly (1987).

The major problem facing the manager of site selection is the large number of options from which to choose, although the conceptual bases for this model are very simple. A set of potential locations is defined and from this set P facilities are to be chosen. The so-called N choose P problem clearly involves a large number of combinatorial options. Not all of these choices need to be examined, however, in order for the model to make a reasonable estimate of the ideal subset of P facilities. Two major strategies are available. First, if the model can be posed as an optimization task, computer programs using mathematical techniques such as mixed integer programming (MIP) or Lagrangian relaxation to select optimal locations (O'Kelly, 1987). Second, and in many ways more robustly, the modeler can set up

the problem and employ heuristics in order to make a quick and reliable estimate of the core portion of the preferred site selections.

An example may help to make this concept clear. Suppose a clothing retailer is considering siting stores in some of the many available shopping centers in a large metropolitan region such as Atlanta. It is unlikely that the retailer would want to place a store in every available shopping center. Budget constraints would limit this option and simple common sense would indicate that the market could not bear the saturation coverage of 'too many' stores. The question of the optimal number of stores will be addressed presently, for now assume that the retailer has a limited number of sites that are under consideration. Therefore the retailer seeks to prioritize a subset of all the available centers that might be expected to perform well given their products and customer profile. This latter point is a key one. In order for the retailer to prioritize the store locations, the retailer needs to use an accurate model of the underlying demand for the service. Thus many geo-demographic case studies use *profiles* of existing customers to create a measure that reflects the attraction of the store for particular populations. This in essence is a computerized version of the classic idea by Applebaum (1966) of using analogs to project the trade area success of a proposed new store location. If the chain already has a set of stores in a wide variety of different spatial contexts, cross-sectional comparison of the performance of those stores can be used to produce a regression type model for store sales levels. Once these models are estimated, the retailer can then seek new locations where the mix of factors leans heavily towards those variables that have proven to be successful predictors in other locations. The operational version of this idea is to test each of the locational scenarios by projecting the probable trade area of each store, existing or proposed, in the context of the surrounding demographics and competition. These models have become very sophisticated because of the availability of detailed micro demographic profiles of spatial areas that may be assigned to each potential location.

As the model explores the number of locations, the analyst can keep track of the performance of those proposed sites. For example a set of five stores distributed throughout the metropolitan region might very well succeed in capturing the selected demographic submarkets that are sought and desired by this retailer. In contrast, some other combination of five stores could easily be eliminated from consideration because the sites do not deliver the expected mix and density of demand to make this package feasible. A great deal depends on a reasonable and accurate projection of the impact of each new store and its performance both against existing competitors and any stores that the chain might already have located in the district.

22.4.3. Combinatoric issues

A key to the efficient implementation of interaction based location models is a data structure that enables the computerized evaluation of sites to be made relatively quickly. The following notes provide a guide to the collection and organization of data in such a way as to make such computations feasible for quite a large study program. Assume that there are M origin zones. The N locations from which the model will select sites are organized as the columns of the interaction table with an extra column that will be used to store any user demand that is under-served by the solution program. This modification is essential when dealing with site selection models. To see this, imagine that a retailer is planning to site three new outlets in a very large metropolitan area. If the maximum distance a customer would be willing to

travel to the store is set at say 10 miles (equivalent to the concept of the range in CPT) then in a large city, it is quite clear that some consumers will be too far from any of the chosen sites to be able to use this retailer's service. It is important that the model provide a means to calculate such unserved customers and we propose to do this by placing those 'unserved' consumers in a separate 'dummy' destination category as a holding bin for the under-provided origin zones. In the absence of competition, the goal would be to minimize unserved demand. In the presence of competitive alternatives, the goal would be to capture as much unserved demand as possible for the client's chain.

With the exception of the concept of an additional destination, the basic calculation process is identical to that of a production constrained spatial interaction model. The device used to operationalize a particular choice of actively considered facilities is to simply keep a list of certain columns from the interaction matrix to which consumers might be allocated during that particular iteration. As the model proceeds from one locational pattern to another the set of active columns is simply switched on and off to provide an indication of the currently available destination choices. To make these calculations efficiently the computer is provided with lists pointing to various types of columns in the matrix. For example any sites which are required to be provided in all cases may be indicated by placing their column numbers in a vector of open facilities. Such a vector might be the noted by the letter *R* for required centers. A second set of pointers might be used to indicate that in a particular analysis some potential facility locations are to be ignored completely. These, for example might be sites which we wish to lock out of the current set of optionally available sites. Yet another list could maintain a set of pointers to the available remaining unexplored options that are freely available to the model to be chosen or not as the analysis progresses. Once again having an example may help to fix these ideas. Suppose that a city currently has a total of 35 supermarkets from a number of major chain stores. One of these chains is considering a variety of expansion programs in this city. Among the locational options available to it are the acquisition of new sites, the acquisition of existing sites from competitors, and the expansion of some or all of the current stores in its portfolio. In this case it is reasonable to think that the existing stores in the market are in a sense locked in and will occur in all of the comparison scenarios: 35 columns of the interaction matrix are therefore locked in for the purposes of this initial run. Any additional locations are simply tacked on as say the 36th, 37th or 38th columns of this interaction matrix. Depending on how many candidates sites are available from which to pick these three additional locations one can imagine that the model is exploring a finite list of potential new store packages. Common sense dictates that the store chain is unlikely to want all of its new site picks in the same area, as it would make a great deal more sense to spread the chosen sites across a variety of sectors of the city. If it so happened that a pool of presently underserved demand could be found, the model would place a facility in that area. More likely, the model would be making a complex set of trade-offs, trying to eke out a market share from among and between the existing set of competitive centers, and indeed avoiding cannibalizing the existing store already owned by the chain. In this regard the strategy is essentially similar to the well known 'gap in the map' rubric for locating new services. The bulk of the program then would spend time computing the benefits of specific chosen alternatives in, for example, the north, east, and south suburbs. For those with the obvious question of 'how is this done,' it would be realistic to state that the

current practice involves a combination of GIS software to manage the spatial data, customized optimization algorithms coded as executable computer programs, and a report writer to digest the output from the optimization run. While these capabilities may be combined in various customized software environments by consultants, there is probably no prepackaged comprehensive optimization environment for the applied tasks enumerated here, though this situation will undoubtedly change.

22.4.4. Heuristics and other shortcuts[9]

The position of the store relative to the pool of demand and to other complementary and competitive stores is critical in measuring market area and size. If the objective is related to maximizing aggregate market share for our entire chain, and if there is an accurate representation of maximum distance (reservation distance) we can expect that the model will 'naturally' space out our stores giving them somewhat non-overlapping exclusive market areas. Nevertheless, when two stores are close enough to contest a middle ground then the gravity model will do better than the usual deterministic all-or-nothing location models. The gravity model will in fact partition the demand between the centers in proportion to their attraction and weight.

If such a model is to be run in site selection mode, realize that the attraction/repulsion score will have to be computed for prospective as well existing sites – in other words it has to be some calculated feature of sites that are 'prospects': it cannot be simply some observable feature of existing sites.

Required site

These are locations of our own chain that we wish to keep. We can also, in some cases, represent the locations of competitor stores that we know are remaining in operation. These would be treated as fixed sites.

Prohibited site

Areas or store locations that are prohibited from entering the model are equally important to a realistic implementation. If we are sure that the chain does not wish to enter certain malls, or if the location in proximity to some existing stores is strongly discouraged, then candidate locations in the 'no go zone' should be flagged to (a) save computer time; and (b) and to enhance the chance that the model will focus attention in areas that are worth investigating.

Flexible sites

The set of locations from which the model will pick are predefined by the user. These could be the result of selection set operations, query based lists, or geographically delimited regions on the screen. What is important is for an underlying comprehensive data base to be kept up to date in order for the analysts to have meaningful choices from which to derive the set of active alternatives.

22.4.5. Computable Location Models[10]

Location models must be flexible to allow analysis of different scenarios. The model takes as input the required and flexible sites. The existing literature contains several models dealing with joint location and allocation under spatial interaction: these however, need to be modified to handle realistic selection sets of required and prohibited sites.

The best practice at this time is to use a robust vertex substitution method appropriately modified to handle lists of required and prohibited sites, as well as

efficiently managing the introduction of new candidate locations.

The vertex substitution method also needs to include the capability of a maximum service radius for the facilities, and for this radius to be flexible/variable between centers: this is essential if some notion of center hierarchy is to be accommodated. It should be clarified that the vertex substitution method is a local optimally solution in the sense that there may be a better solution that was not reached during the course of the exploration; this possibility can be reduced by trying the method with various starting values. Research experience has shown, however, that the good locations 'stand out' very well and the possibility that the vertex substitution method completely misses the best package of locations is remote. One idea that is suggested to prevent mistakes due to local optimality is to produce a list not only of the best locations but other close contenders discovered in the course of the algorithm's progress.

Research by Church has shown that the introduction of maximum service radii into a median type of problem (which is what we have) disrupts one 'normal' property of the model, making it potentially possible that the 'optimal' locations occur at points other than the nodes of the network. However, the actual problem that we are concerned with realistically limits the feasible locations to the nodes of the network, as this is where the shopping centers are. In other words we ignore the theoretical possibility that the true optimal solution is at an intermediate location along street segments, as in practice this kind of locational solution would not be permissible.

What does experience tell us about the solution of location allocation models? The basic model is conceptually very simple and easy to understand. The idea is to systematically explore alternative locational scenarios. The method takes as input the fixed locations, the candidates, and the prohibited sites (if any). As output the model produces the requested number of additional facility sites, and reports on the area characteristics of both the current and the new sites. The candidates are either a comprehensive list of all feasible shopping centers, stores are generated from a list of 'picks' and potential sites. The user may select the candidates as those sites which meet some criteria, and the detail and realism of these selection criteria are really only constrained by the imagination of the user. All kinds of filters can be used, including center size, or selections can be based on attributes of the centers. Having selected the candidates, the user would have to select the objective function: normally this is driven on the basis of aggregate market share, or demand, or minimizing competitors share. This is potentially extended to include acquisition, lease, closing and opening financial decisions.

Vertex substitution has the great advantage that as a general purpose optimization strategy (i.e., heuristic) it is robust to changes of objective function, in a way, for example, that would not be true of a specialized exact optimization code. In other words, the weakness of an 'exact' method is that it typically has to exploit some aspect of the problem structure and any change in that structure would likely undermine the mathematical formulation. Heuristics (and there are many of these available for combinatorial problems) can frequently be set up to explore a solution space effectively and this can be chosen to evaluate the users choice of objective (and indeed multiple objectives) to achieve the desired goals. Indeed the final great advantage of an exploratory heuristic is that by careful book-keeping many 'runner up' or close alternative solutions can be kept and compared.

22.5. STRATEGIC PLANNING EXAMPLES

22.5.1. Shopping centers

Store location siting is often made from among a predefined set of existing shopping centers, so in a sense the set from which the strategic location is to be chosen is already fixed. Thus, 1747 block groups in Atlanta represent the pools of available demand, which for the purposes of this simple example are weighted by the population or disposable income as a proxy for the demand. Assume a chain has 12 existing stores distributed throughout the Atlanta region in specific shopping centers. There are approximately 230 potential sites in shopping centers. Assume that the reach or 'draw' of the center candidates is a function of the size of the center – in other words the decision to open a new branch in a thriving center with a super-regional draw might be appropriately measured by using the size of the center as a proxy for its suitability. Suppose then that the location allocation model algorithm, such as the Interchange Heuristic, picks four locations as the close-to-optimal added sites. (We are careful not to call them optimal in view of the many simplifications and the use of a heuristic which after all depends on some short cuts to avoid complete enumeration of the many thousands of combinations that are available.)

The impact of each new center on the 12 existing sites is then operationally measured using a formula such as the one discussed above.

22.5.2. Chain combinations

Sales of branches in two existing sets of chain stores can give a good clue as to the best ones to keep in the combined operation, but that still leaves a difficult problem to determine

Where did the impact on store come from?					
NEW1	*NEW2*	*NEW3*	*NEW4*	*Taken from store number*	
0	0	0	0.04	1	0.04
0	0	0	0.47	2	0.47
0	1.02	0	0.01	3	1.03
0	0.65	0	0.41	4	1.06
0	0	0	0	5	0
0	0	0	0	6	0
0.04	0	0	0	7	0.04
0	0	0	0	8	0
0	0	1.11	0	9	1.11
0.22	0	0	0	10	0.22
0	0	0.51	0	11	0.5
0.02	0	0	0	12	0.02
0.28	1.67	1.62	0.93	From existing	
0.82	2.35	2.64	2.78	Total	
0.54	0.68	1.02	1.85	Net added	
Before	Total share	0.9434			
After	Total share	0.9842			
		4.08			

which ones to close. Predicting retention of customers from old stores to re-aligned new branches is also difficult though the managers of such operations may have good insight into the likely levels of customer loyalty.

An interesting question is to determine the diversion of sales or the result of a store/chain closure. Such questions frequently are presented in practice to retailers as they have the option to purchase competitors sites. Which of these sites would make good acquisitions (if the option to cherry pick the best of the available store)? Which would be blended well and open under the new label if the acquiring chain gets the whole suite?

If two chains merge, and there are regulatory concerns that the two chains have to divest some of their branches, or wish to streamline their combined operations, one would have to analyze the closure of branches one by one to determine the package that makes the most sense from the point of view of the combined operations.

22.6. SUMMARY AND CONCLUSIONS

The great strength of the gravity model is its simplicity and its allocation of demand to centers in proportion to their attraction and inversely proportional to distance. It can incorporate center specific attraction and center specific maximum trade area radii.

The strength of the SI based location model is that it provides assistance with all of the following tasks: measuring saturation, impact of changes on current trade areas, assessment of the advantages of certain locations for particular formats, and an estimation of the forecast of sales. In addition the allocation models allow a profile of the demographics of a trade area.

What would take a large amount of extra research effort, but which in my opinion would be well worth while, would be the inclusion of the interaction based model in a multiobjective and multiattribute decision framework. The difficulty would be to elicit from the decision maker a set of trade off parameters that define the relative scales for the attributes of the alternative locational packages.

The mechanism reviewed in this chapter will operate to allocate the sales from the origin zones to the destinations is called the allocation model. It is driven by a gravity based spatial interaction model, and given careful data and careful assessment of the foundation assumptions this is a robust model for trade area delimitation.

ACKNOWLEDGMENTS

Parts of this chapter are based on materials developed over many years in my Retail Location Seminar where comments from Debbie Bryan, Tony Grubesic, and Tim Matisziw are gratefully acknowledged (see specific footnotes). In addition, a great deal of the common sense application flavor of this paper derives from conversations with Jim Stone (GeoVue), Tony Lea (Environics Analytics), and Steve Wheelock. I thank these individuals while taking full responsibility for the product here. Some material originally prepared as a discussion/research memo on location models for Geonomics. See 22.1.2, 22.4.5, 22.4.6, and examples in 5. Other material derived from 'Retail Location. Models and Spatial Interaction' M.E. O'Kelly and D. Bryan. A Review of Modeling in Retail Location Unpublished working paper.

NOTES

1 Introduction is based on Geography 845 Lecture Jan 2, 2001.

2 A major sector using the results from spatial modeling capability is that of businesses with multi-store/branch locations. Home Depot for example has made extensive use of reports from what used to be Thompson Associates, and is now a unit of MapInfo in Ann Arbor, MI. Other well known users include McDonalds and Blockbuster.

3 Based on applications as discussed with Jim Stone and Tony Lea.

4 Some aspects of these following paragraphs have benefited from discussion with Jim Stone.

5 The target level of goodness-of-fit in convenience store forecasting models is for high *r*-square values (about 0.8).

6 Section 22.2.1 is based on 'Retail Location Models and Spatial Interaction' by M.E. O'Kelly and D. Bryan, *A Review of Modeling in Retail Location*. Unpublished working paper.

7 GeoVue has a gravity based software package. ESRI Business Analyst software has a Huff trade area model.

8 This material derived from 'Retail Location Models and Spatial Interaction' by M.E. O'Kelly and D. Bryan, *A Review of Modeling in Retail Location*. Unpublished working paper.

9 Material in section 22.4.5 was originally discussed in an explanatory memo from this author to Jim Stone at Geonomics (now GeoVue). Jim's critique was helpful in framing the discussion.

10 This section benefited from discussion with Jim Stone and Steve Wheelock.

REFERENCES

(Although not all these papers are cited directly, these are however all influential papers in my analysis; they are retained as a general bibliographic resource.)

Achabal, D., Gorr, W. and Mahajan, V. (1982). MULTILOC, A multiple store location decision model. *Journal of Retailing*, **58**(2): 5–25.

Applebaum, W. (1966). Methods for determining store trade areas, market penetration, and potential sales. *Journal of Marketing Research*, **3**: 127–141.

Baker, R.G.V. (2000). Towards a dynamic aggregate shopping model and its application to retail trading hour and market area analysis. *Papers in Regional Science*, **79**(4): 413–434.

Balakrishnan, P.V., Desai, A. and Storbeck, J.E. (1994). Efficiency evaluation of retail outlet networks. *Environment and Planning B*, **21**(4): 477–488.

Beaumont, J.R. (1980). Spatial interaction models and the location–allocation problem. *Journal of Regional Science*, **20**(1): 37–50.

Beaumont, J.R. (1981). Location–allocation models in a plane, a review of some models. *Socio-Economic Planning Sciences*, **15**(5): 217–229.

Berry, B.J.L. (1967). *The Geography of Market Centers and Retail Distribution*. Englewood Cliffs, NJ: Prentice Hall.

Birkin, M., Clarke, G. and Clarke, M.P. (2002). *Retail Geography and Intelligent Network Planning*. New York: Wiley.

Black, W. (1984). Choice-set definition in patronage modeling. *Journal of Retailing*, **60**(2): 63–85.

Boots, B. and South, R. (1997). Modeling retail trade areas using higher-order, multiplicatively weighted Voronoi diagrams. *Journal of Retailing*, **73**(4): 519–536.

Borgers, A. and Timmermans, H. (1986). A model of pedestrian route choice and demand for retail facilities within inner-city shopping areas. *Geographical Analysis*, **18**(2): 115–128.

Borgers, A. and Timmermans, H. (1991). A decision support and expert system for retail planning. *Computers Environment and Urban Systems*, **15**(3): 179–188.

Brown, S. (1989). Retail location theory, the legacy of Harold Hotelling. *Journal of Retailing*, **65**(4): 450–470.

Brown, S. (1992). The wheel of retail gravitation. *Environment and Planning A*, **24**(10): 1409–1429.

Brown, S. (1994). Retail location at the micro-sale – inventory and prospect. *Service Industries Journal*, **14**(4): 542–576.

Buckner, R.W. (1998). *Site Selection, New Advances in Methods and Technology*, 2nd Edn. New York: Lebhar-Friedman Books.

Clark, W.A.V. (1968). Consumer travel patterns and the concept of range. *Annals, Association of American Geographers*, **58**: 386–396.

Congdon, P. (2000). A Bayesian approach to prediction using the gravity model, with an application to patient flow modeling. *Geographical Analysis*, **32**(3): 205–224.

Craig, C.S., Ghosh, A. and McLafferty, S. (1984). Models of the retail location process, a review. *Journal of Retailing*, **60**(1): 5–36.

Current, J.R. and Storbeck, J.E. (1994). A multiobjective approach to design franchise outlet networks. *Journal of the Operational Research Society*, **45**(1): 71–81.

Ding, G. and O'Kelly, M.E. (2008). Choice-based estimation of Alonso's Theory of Movement, Methods and Experiments. in *Environment and Planning A*, **40**(5): 1076–1089.

Donthu, N. and Yoo, B. (1998). Retail productivity assessment using data envelopment analysis. *Journal of Retailing*, **74**: 89–105.

Drezner, T. (1994). Optimal continuous location of a retail facility, facility attractiveness, and market share – an interactive model. *Journal of Retailing*, **70**(1): 49–64.

Drezner, T. and Drezner, Z. (2002). Validating the gravity-based competitive location model using inferred attractiveness. *Annals of Operations Research*, **111**(1–4): 227–237.

Drezner, T., Drezner, Z. and Salhi, S. (2002). Solving the multiple competitive facilities location problem. *European Journal of Operational Research*, **142**(1): 138–151.

Durvasula, S., Sharma, S. and Andrews, J.C. (1992). Storeloc – a retail store location model-based on managerial judgments. *Journal of Retailing*, **68**(4): 420–444.

Eagle, T.C. (1984). Parameter stability in disaggregate retail choice models – experimental-evidence. *Journal of Retailing*, **60**(1): 101–123.

Eaton, B.C. and Lipsey, R.G. (1982). An economic theory of central places. *Economic Journal*, **92**: 56–72.

Erlenkotter, D. and Leonardi, G. (1985). Facility location with spatially-interactive behavior. *Sistemi Urbani*, **1**: 29–41.

Fotheringham, A.S. (1983). A new set of spatial interaction models, the theory of competing destinations. *Environment and Planning A*, **15**: 15–36.

Fotheringham, A.S. (1986). Modelling hierarchical destination choice. *Environment and Planning A*, **18**: 401–418.

Fotheringham, A.S. and O'Kelly, M.E. (1989). *Spatial Interaction Models, Formulations and Applications Studies in Operational Regional Science*. Dordrecht, Netherlands: Kluwer.

Fotheringham, A.S. and Knudsen, D.C. (1986). Modeling discontinuous change in retailing systems – extensions of the Harris–Wilson framework with results from a simulated urban retailing system. *Geographical Analysis*, **18**(4): 295–312.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression, The Analysis of Spatially Varying Relationships*. Chichester: Wiley.

Ghosh, A. (1986). The value of a mall and other insights from a revised central place model. *Journal of Retailing*, **62**(1): 79–97.

Ghosh, A. and Craig, C.S. (1983). Formulating retail location strategy in a changing environment. *Journal of Marketing*, **47**(3): 56–68.

Ghosh, A. and McLafferty, S. (1982). Locating stores in uncertain environments, a scenario planning approach. *Journal of Retailing*, **58**(Winter): 5–22.

Ghosh, A. and McLafferty, S.L. (1987). *Location Strategies for Retail and Service Firms*. Lexington, MA: Lexington Books.

Ghosh, A. and Craig, C.S. (1991). FRANSYS, a franchise distribution system location model. *Journal of Retailing*, **67**(4): 466–495.

Ghosh, A. and Tibrewala, V. (1992). Optimal timing and location in competitive markets. *Geographical Analysis*, **24**(4): 317–334.

Golledge, R. and Spector, A. (1978). Comprehending the urban environment, theory and practice. *Geographical Analysis*, **10**(4): 403–426.

Goodchild, M.F. (1984). ILACS, A location-allocation model for retail site selection. *Journal of Retailing*, **60**(1): 84–100.

Goodchild, M.F. (1991). Geographic information systems. *Journal of Retailing*, **67**(1): 3–15.

Guy, C.M. (1991). Spatial interaction modeling in retail planning practice – the need for robust statistical-methods. *Environment and Planning B*, **18**(2): 191–203.

Hallsworth, A.G. (1994). Decentralization of retailing in Britain – the breaking of the 3rd wave. *Professional Geographer*, **46**(3): 296–307.

Hanson, S. (1980). Spatial diversification and multipurpose travel, implications for choice theory. *Geographical Analysis*, **12**: 245–257.

Hodgson, M.J. (1978). Towards more realistic allocation in location–allocation models, an interaction approach. *Environment and Planning A*, **10**: 1273–1285.

Hodgson, M.J. (1981). A location–allocation model maximizing consumers welfare. *Regional Studies*, **15**(6): 493–506.

Hodgson, M.J. (1986). An hierarchical location–allocation model with allocations based on facility size. *Annals of Operational Research*, **6**: 273–289.

Houston, F.S. and Stanton, J.(1984). Evaluating retail trade areas for convenience stores. *Journal of Retailing*, **60**(1): 124–136.

Hubbard, R. (1978). A review of selected factors conditioning consumer travel behavior. *Journal of Consumer Research*, **5**: 1–21.

Huff, D.L. (1962). *Determination of Intra-Urban Retail Trade Areas*. Real Estate Research Program. University of California at Los Angeles.

Huff, D.L. (1963). A probabilistic analysis of shopping center trade areas. *Land Economics*, **39**: 81–90.

Huff, D.L. (1964). Defining and estimating a trade area. *Journal of Marketing*, **28**: 34–38.

Jain, A.K. and Mahajan, V. (1979). Evaluating the competitive environment in retailing using multiplicative competitive interaction models. In Sheth, J. (ed.), *Research in Marketing*, pp. 217–235. Greenwich, Conn: JAI Press.

Kantorovich, Y.G. (1992). Equilibrium-Models of Spatial Interaction with Locational-Capacity Constraints. *Environment and Planning A*, **24**(8): 1077–1095.

Kaufmann, P.J., Donthu, N.B. and Brooks, C.M. (2000). Multi-unit retail site selection processes, incorporating opening delays and unidentified competition. *Journal of Retailing*, **76**(1): 113–127.

Kitamura, R. and Kermanshah, M. (1985). Sequential model of interdependent activity and destination choices. *Transportation Research Record*, **987**: 81–89.

Kohsaka, H. (1989). A spatial search-location model of retail centers. *Geographical Analysis*, **21**(4): 338–349.

Kohsaka, H. (1992). Three-dimensional representation and estimation of retail store demand by bicubic splines. *Journal of Retailing*, **68**(2): 221–241.

Kohsaka, H. (1993). A monitoring and locational decision support system for retail activity. *Environment and Planning A*, **25**(2): 197–211.

Krider, R.E. and Weinberg, C.B. (1997). Spatial competition and bounded rationality, Retailing at the edge of chaos. *Geographical Analysis*, **29**(1): 16–34.

Kumar, V. and Karande, K. (2000). The effect of retail store environment on retailer performance. *Journal of Business Research*, **49**(2): 167–181.

Lakshmanan, T.R. and Hansen, W.A. (1965). A retail market potential model. *Journal of the American Institute of Planners*, **31**: 134–143.

Langston, P., Clarke, G.P. and Clarke, D.B. (1997). Retail saturation, retail location, and retail competition, An analysis of British grocery retailing. *Environment and Planning A*, **29**(1): 77–104.

Leonardi, G. (1980). A unifying framework for public facility location problems. WP-80-79, IIASA, Laxenburg, Austria.

Leonardi, G. (1983). The use of random-utility theory in building location–allocation models. In: Thisse, J.-F. and Zoller, H. (eds), *Locational Analysis of Public Facilities*, pp. 357–383. Amsterdam: North Holland.

Leszczyc, P. and Timmermans, H.J.P. (1996). An unconditional competing risk hazard model of consumer store- choice dynamics. *Environment and Planning A*, **28**(2): 357–368.

Leszczyc, P., Sinha, A. and Timmermans, H. (2000). Consumer store choice dynamics. An analysis of the competitive market structure for grocery stores. *Journal of Retailing*, **76**(3): 323–345.

Leszczyc, P., Sinha, A. and Sahgal, A. (2004). The effect of multi-purpose shopping on pricing and location strategy for grocery stores. *Journal of Retailing*, **80**(2): 85–99.

Longley, P. and Clarke, G. (1996). *GIS for Business and Service Planning*. New York: Wiley.

McLafferty, S.L. and Ghosh, A. (1986). Multipurpose shopping and the location of retail firms. *Geographical Analysis*, **18**(3): 215–226.

Mercer, A. (1993). Developments in implementable retailing research. *European Journal of Operational Research*, **68**(1): 1–8.

Miller, H. and O'Kelly, M.E. (1991). Properties and estimation of a production-constrained Alonso model. *Environment and Planning A*, **23**: 127–138.

Miller, H.J. (1993). Consumer Search and Retail Analysis. *Journal of Retailing*, **69**(2): 160–192.

Min, H. (1987). A multiobjective retail service location model for fastfood restaurants. *OMEGA*, **15**(5): 429–441.

Munroe, S. (2001). Retail structural dynamics and the forces behind big-box retailing. *Annals of Regional Science*, **35**(3): 357–373.

Nakanishi, M. and Cooper, L.G. (1974). Parameter estimation for a multiplicative competitive interaction

model – least squares approach. *Journal of Marketing Research*, **11**: 303–311.

O'Kelly, M.E. (1981). A model of the demand for retail facilities incorporating multistop multipurpose trips. *Geographical Analysis*, **13**(2): 134–148.

O'Kelly, M.E. (1983a). Multipurpose shopping trips and the size of retail facilities. *Annals of the Association of American Geographers*, **73**(2): 231–239.

O'Kelly, M.E. (1983b). Impacts of multistop multipurpose trips on retail distributions. *Urban Geography*, **4**(2): 173–190.

O'Kelly, M.E. and Storbeck, J.E. (1984). Hierarchical location models with probabilistic allocation. *Regional Studies*, **18**(2): 121–129.

O'Kelly, M.E. (1987). Spatial interaction based location–allocation models. In: Ghosh A. and Rushton, G. (eds), *Spatial Analysis and Location–Allocation Models*, pp. 302–326. New York: van Nostrand Reinhold.

O'Kelly, M.E. and Miller, H.J. (1989). A synthesis of some market area delimitation tools. *Growth and Change*, **20**: 14–33.

O'Kelly, M.E. (1999). Trade-area models and choice-based samples, methods. *Environment and Planning A*, **31**(4): 613–627.

O'Kelly, M.E. (2001). Retail market share and saturation. *Journal of Retailing and Consumer Services*, **8**(1): 37–45.

Oppenheim, N. (1990). Discontinuous changes in equilibrium retail activity and travel structures. *Papers of the Regional Science Association*, **68**: 43–56.

Pirkul, H., Narasimhan, S. and De, P. (1987). Firm expansion through franchising, a model and solution procedure. *Decision Science*, **18**: 631–641.

Prendergast G., Marr, N. and Jarratt, B. (1998). Retailers' views of shopping centres, a comparison of tenants and non-tenants. *International Journal of Retail and Distribution Management*, **26**(4): 162–171.

Roy, J.R. and Thill, J.C. (2004). Spatial interaction modelling. *Papers in Regional Science*, **83**(1): 339–361.

Rust, R.T. and Brown, J.A.N. (1986). Estimation and comparison of market area densities. *Journal of Retailing*, **62**(4): 410–430.

Rust, R.T. and Donthu, N. (1995). Capturing geographically localized misspecification error in retail store choice models. *Journal of Marketing Research*, **32**(1): 103–110.

Seldin, M. (1995). The information revolution and real estate analyses. *Real Estate Issues*, April 1995.

Thill, J.C. (2000). Network competition and branch differentiation with consumer heterogeneity. *Annals of Regional Science*, **34**(3): 451–468.

Timmermans, H., Arentze, T. and Joh, C.-H. (2002). Analysing space–time behaviour, new approaches to old problems. *Progress in Human Geography*, **26**(2): 175–190.

Timmermans, H., Vanderhagen, X. and Borgers, A. (1992). Transportation systems, retail environments and pedestrian trip chaining behavior – modeling issues and applications. *Transportation Research Part B – Methodological*, **26**(1): 45–59.

Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46**: 234–240.

Wee, C.H. and Pearce, M.R. (1985). Patronage Behavior toward Shopping Areas – a Proposed Model Based on Huffs Model of Retail Gravitation. *Advances in Consumer Research*, **12**: 592–597.

Weisbrod, G.E., Parcells, R.J. and Kern, C. (1984). A disaggregate model for predicting shopping area market attraction. *Journal of Retailing*, **60**(1): 65–83.

Wilson, A.G. and Senior, M.L. (1974). Some relationships between entropy maximizing models, mathematical programming models and their duals. *Journal of Regional Science*, **14**: 207–215.

Wilson, A.G., Coelho, J.D. Macgill, S.M. and Williams, H.C.W.L. (1981). *Optimization in Locational and Transport Analysis*, London: Wiley.

Zeller, R.E., Achabal, D.D. and Brown, L.A. (1980). Market penetration and locational conflict in franchise systems. *Decision Sciences*, **11**: 58–80.

23

Spatial Analysis on a Network

Atsuyuki Okabe and Toshiaki Satoh

23.1. INTRODUCTION

In the real world, various types of phenomena occur on or alongside a network; these are termed *network spatial phenomena*. A typical example is illustrated in Figure 23.1, where dots show traffic accidents in Chiba, Japan. As with this example, many types of network spatial phenomena are reported in the related literature: traffic accidents on a road network (e.g., Jones *et al*., 1996; Levine *et al*., 1995; McGuigan, 1981; Nicholson, 1989; Yamada and Thill, 2004), road kills on a road network (e.g., Bashore *et al*., 1985; Clevenger *et al*., 2003; Mallick *et al*., 1998; Saeki and MacDonald, 2004), street crimes on a street network (e.g., Anselin *et al*., 2000; Bowers and Hirschfield, 1999; Ratcliffe, 2002; Ratcliffe and McCullagh, 1999; Painter, 1994), the distribution of seabirds along a coastline (e.g., O'Driscoll, 1998), and the distribution of trees along a road network (e.g., Spooner *et al*., 2004). These are examples of network spatial phenomena that occur directly on a network.

As well as the above network spatial phenomena, another broad class of phenomena occurs alongside rather than directly on a network. A typical example is illustrated in Figure 23.2, where dots indicate parking lots in Kyoto, Japan. There are many facilities in addition to parking lots that are located alongside street networks within densely inhabited areas. In fact, the entrances to almost all facilities in a city are adjacent to a street and users access such facilities through these entrances. Consequently, the location phenomena of almost all facilities within urbanized areas can be regarded as a second class of network spatial phenomena.

Network spatial phenomena are usually analyzed by methods that assume a continuous plane and Euclidean distance (expect for transportation studies). For referential purposes, these types of spatial methods are referred to as *planar spatial methods*,

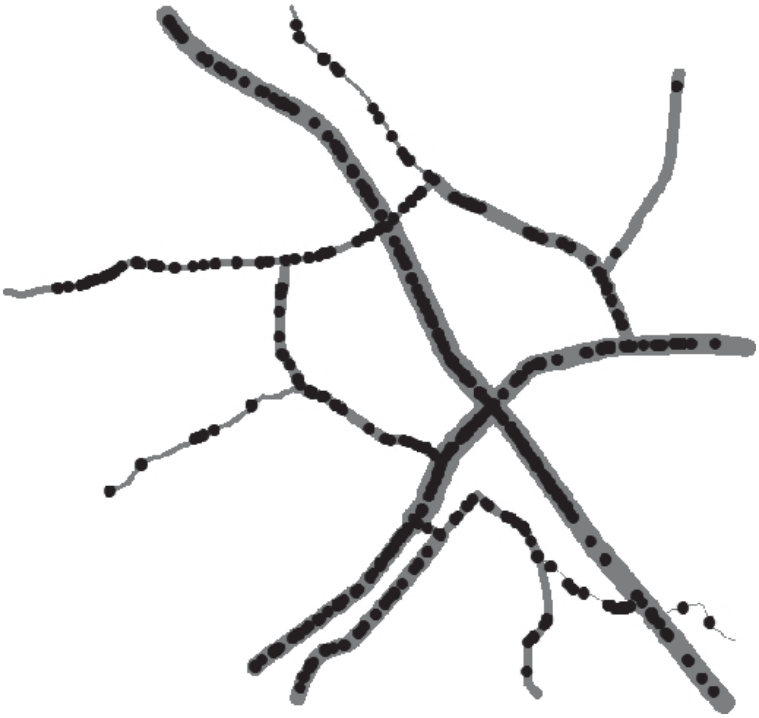


Figure 23.1 Sites of traffic accidents in Chiba, Japan (the width of each line segment represents traffic volume).

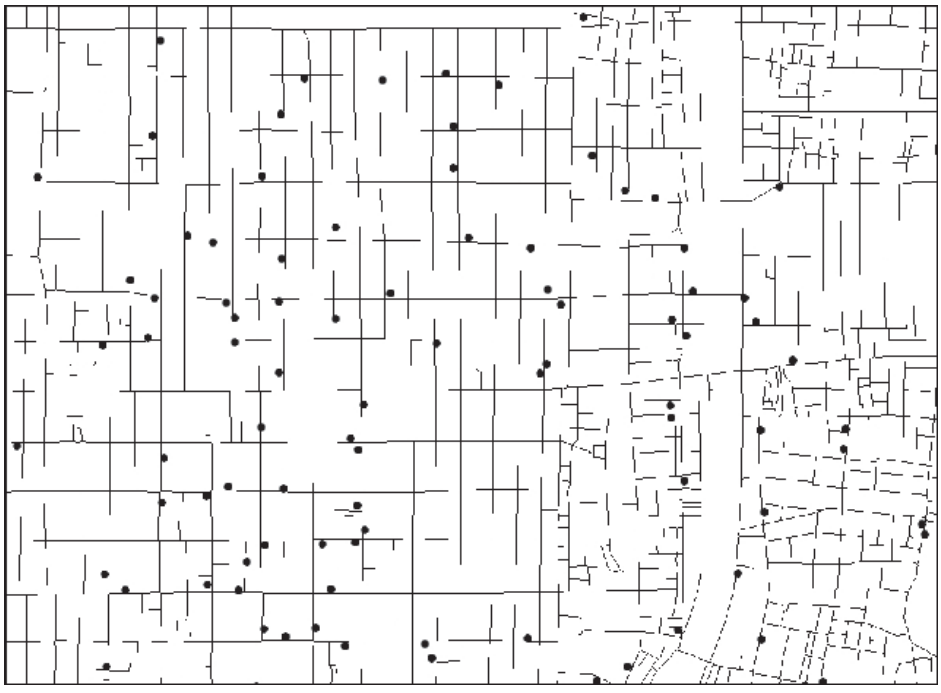


Figure 23.2 The distribution of parking lots in Kyoto, Japan.

and analyses via planar spatial methods are termed *planar spatial analyses*. Planar spatial methods are generally used for the analysis of network spatial phenomena because: (1) it is much easier to compute Euclidean distance on a plane than by the shortest-path distance on a network, and (2) it is believed that the shortest-path distance is approximated by Euclidean distance. The first reason remains true, although the difficulty is reduced these days because the use of Geographical Information Systems (GIS) makes it easy to calculate the shortest-path distance. The second reason might be true over a large region, but its validity is questionable across a small area or within a city. For example, Maki and Okabe (2005) demonstrated that, in Kokuryo, a Tokyo suburb, the difference between the shortest-path distance and Euclidean distance is significant if the Euclidean distance is less than 500 m (see Figure 23.3). Therefore, to analyze spatial phenomena in small areas such as the market areas of convenience stores in a city, planar spatial methods are inappropriate; instead, spatial methods that assume a network space using the shortest-path distance, termed *network spatial methods*, should be used.

The danger in applying planar spatial methods to network spatial phenomena is clearly demonstrated in Figure 23.4. Having assessed the distribution of points in Figure 23.4(a), nobody would consider that the points are randomly distributed. This view is true when points are distributed on a plane; however, this view is false when the points are distributed on the network indicated by the lines in Figure 23.4(b). In fact, the points in the figure are randomly generated on the network.

Figure 23.4 provides the following warning: analyzing network spatial phenomena using a planar spatial method is likely to lead to false conclusions. To avoid such errors, this chapter considers a class of network spatial methods. The chapter consists of seven sections including this introductory section. Section 23.2 describes a method, termed the *uniform network transformation*, that deals with a nonuniform distribution function on a network. Section 23.3 considers a class of network Voronoi diagrams, and section 23.4 discusses a class of network local and global K function methods. Section 23.5 describes a class of network kernel methods, and Section 23.6 outlines a GIS-based toolbox termed SANET, which is used for network spatial analysis. The chapter ends with Section 23.7, which considers network spatial methods that we have not discussed earlier.

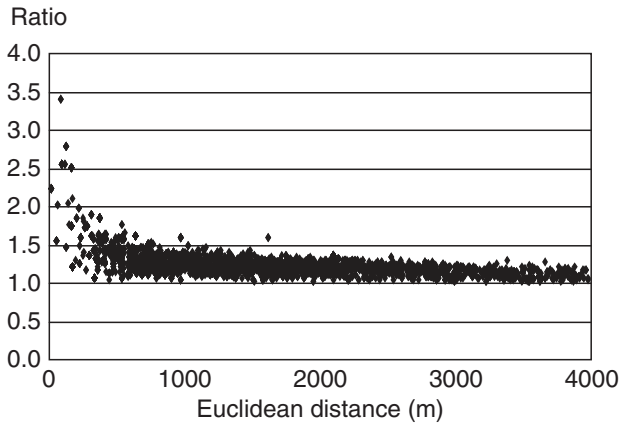


Figure 23.3 Ratio of the shortest-path distance to its corresponding Euclidean distance for the street network in Kokuryo, Tokyo (from Maki and Okabe, 2005).

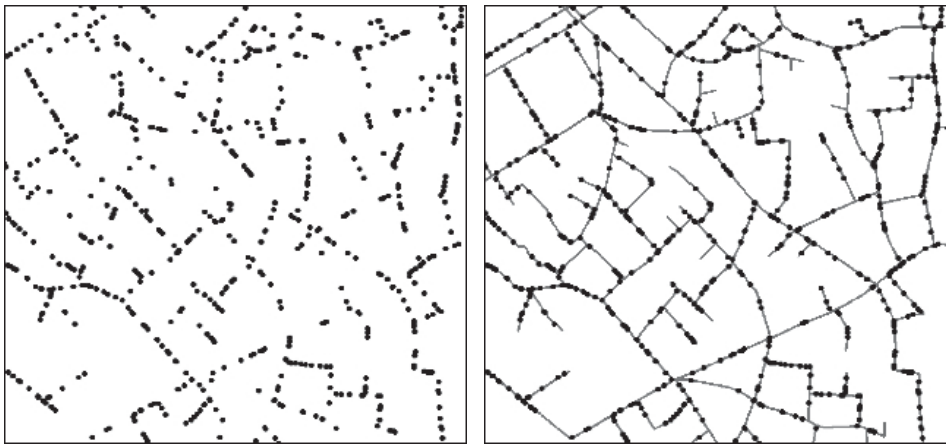


Figure 23.4 Non-randomly distributed points on a plane (a), and randomly distributed points on a network (b). (Note that the two distributions of points are the same.)

23.2. UNIFORM NETWORK TRANSFORMATION

Many spatial methods for analyzing spatial point patterns are designed to test the null hypothesis that points are randomly distributed on a space. In the case of a network space, this null hypothesis means that the probability of a point (say, a traffic accident site) being generated on a unit line segment on a network is the same regardless of the location of the unit line segment; stated differently, the density of points is uniform over the network. Networks that possess this probabilistic property are referred to as *uniform networks*. In the real world, however, uniform networks are unlikely to exist. Rather, the probability of a point being generated on a unit line segment on a network varies according to the location of the unit line segment. Such networks are referred to as *nonuniform networks*. For example, a road network in which traffic accidents occur in proportion to traffic volume is a nonuniform network. Figure 23.1 shows the actual distribution of traffic accidents (as dots) in relation to traffic volume along each line segment (the width represents traffic volume). It is likely that traffic accidents are related to traffic volume, and this relation will be examined in Section 23.4. Applying network spatial methods that assume a uniform network (termed *uniform network spatial methods*) to a nonuniform network is likely to result in false conclusions. Such errors can be avoided by the use of 'uniform network transformation' (Okabe and Satoh, 2006), which is briefly introduced in this section.

First, consider a network L (e.g., a road network) that consists of n line segments (street segments), i.e., $L = \{l_1, \ldots, l_n\}$, and let $c_i \Delta l$ be the probability that a point is located within a unit line segment Δl on l_i, $i = 1, \ldots, n$. Note that this assumption means that the density c_i of points is uniform over l_i, but may vary (as in Figure 23.5(a)) or not vary (as in Figure 23.5(b)) between different line segments. If the density does not vary, i.e., $c_i = c_j$ for all $i,\ j = 1, \ldots, n$, then the network is a *uniform network*; if it does vary, i.e., $c_i \neq c_j$ for at least one pair $i,\ j$, then the network is a *nonuniform network*.

Second, consider a new network $L^* = \{l_1^*, \ldots, l_n^*\}$ whose graph is isomorphic to that

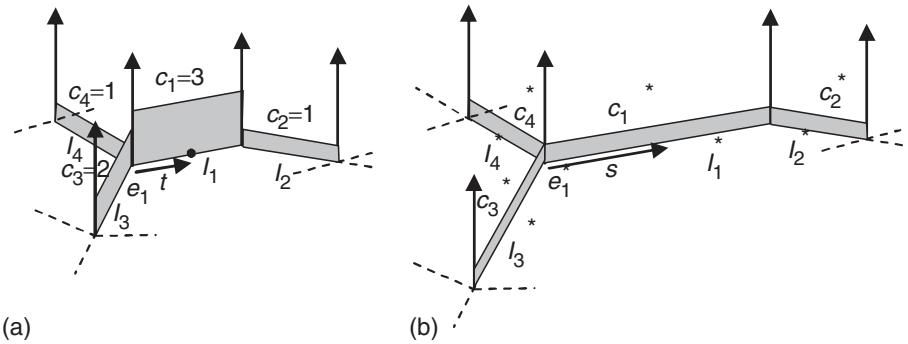


Figure 23.5 A nonuniform network (a) and the equivalent uniform network transformed by the uniform network transformation (b).

of the original network L (Figure 23.5(b)). The location of a point at distance t from one end point e_i of l_i along l_i is mapped on the point at distance s from the end point e_i^* of l_i^* along l_i^* by $s = \alpha c_i t$, where α satisfies $\alpha c_i > 1$ for all $i = 1, \ldots, n$ (Figure 23.5(a, b)).

Third, consider the transformation that satisfies the condition that the probability $c_i^* \Delta l^*$ of a point being placed in a unit line segment Δl^* on l_i^* is the same as the probability $c_i \Delta l$ of a point being placed in a unit line segment Δl on l_i, i.e., $c_i^* \Delta l^* = c_i \Delta l$.

Okabe and Satoh (2006) proved that the above transformation transforms a nonuniform network into a uniform network, and is thus termed the *uniform network transformation*. Note that this transformation is an extension of the probability integral transformation that transforms a univariate nonuniform distribution function to a uniform distribution function. The probability integral transformation is commonly used in statistics to generate nonuniform random variables (Freund, 1998).

The uniform network transformation provides a powerful tool for analyzing nonuniform networks that are commonly found in the real world. Obviously, uniform network spatial methods cannot be used for analyzing a nonuniform network because they are designed for the analysis of a uniform network. However, they can be used if the following simple preprocessing is performed. First, transform a given nonuniform network to a uniform network by the uniform network transformation described above. Second, apply a uniform network spatial method to the transformed network (which is a uniform network). No special development is necessary for dealing with a nonuniform network. Many existing uniform network spatial methods can be utilized for analyzing a nonuniform network without modification through the uniform network transformation. This transformation has the advantage of network spatial analysis, which is not enjoyed by planar spatial analysis.

23.3. NETWORK VORONOI DIAGRAMS

23.3.1. Ordinary network Voronoi diagram

As reviewed by Okabe *et al*. (2000), the ordinary Voronoi diagram, i.e., the Voronoi diagram defined on a plane with Euclidean distance (the ordinary *planar*

Voronoi diagram), is used in many ways in spatial analysis (Figure 23.6). In particular, the ordinary planar Voronoi diagram is commonly used in retail marketing and facilities management as a first approximation of the service areas of stores or facilities.

This approximation, however, is problematic when service areas are small. Table 23.1 shows the average radii of circular market areas in Shinjuku Ward, Tokyo, with respect to store type. In all cases, the distance to the nearest store is less than five hundred meters. Recalling the difference between Euclidean distance and the shortest-path distance shown in Figure 23.3, the data in Table 23.1 suggest that the ordinary planar Voronoi diagram is not appropriate as a first approximation of the service areas.

Instead, a Voronoi diagram defined on a network with shortest-path distance, termed the *network Voronoi diagram*, should be used. To show this clearly, let $d(p, p_i)$ be the shortest-path distance between a point p and a point p_i on a network L, where m generator points (e.g., stores) are located at $p_1, \ldots, p_m$. Let V_i be a set of points on L (a subnetwork) that satisfies

$$V_i = \{p | d(p,\ p_i) \leq d(p,\ p_j), p \in L, \quad j \neq i, j = 1, \ldots, m\}. \tag{23.1}$$

Table 23.1 Average radii of circular market areas in Shinjuku ward, Tokyo (m)

Store types	*Average radius*
Bakery	320
Shoe store	255
Fruit shop	213
Book store	177
Chinese noodle shop	153
Convenience store	150
Beauty parlor	114
Clinic	113

The set of the resulting subnetworks, $V = \{V_1, \ldots, V_m\}$, is termed the (*ordinary*) *network Voronoi diagram* (Okabe *et al.*, 2000); an example is provided in Figure 23.7. It is instructive to compare this network Voronoi diagram with its corresponding planar Voronoi diagram shown in Figure 23.6.

23.3.2. Directed network Voronoi diagrams

In a downtown area, streets are commonly one-way. Pizza delivery stores should consider this fact when dispatching delivery bikes. To take one-way regulations into account, consider a directed network $L_{\rightarrow}$ and let $d_{\rightarrow}(p_i, p)$ be the directed shortest-path distance from p_i (e.g., a pizza delivery store) to p (e.g., a house). Let $V_{i\rightarrow}$ be a set of points on $L_{\rightarrow}$ (a subnetwork) that satisfies equation (23.1), where $d(p, p_i)$ is replaced with $d_{\rightarrow}(p_i, p)$. The set of the resulting subnetworks, $V_{\bullet\rightarrow} = \{V_{1\rightarrow}, \ldots, V_{m\rightarrow}\}$, is termed a *directed network Voronoi diagram* (Okabe *et al.*, 2008); an example is shown in Figure 23.8, where one-way streets are indicated by arrows.

Note that the directed shortest-path distance is not symmetric, i.e., $d_{\rightarrow}(p_i, p) = d_{\rightarrow}(p, p_i)$ does not always hold. Suppose that $p_1, \ldots, p_m$ are parking lots, and a driver at p wants to use the nearest parking lot among $p_1, \ldots, p_m$. The service area of the parking lot at p_i is then defined by the set $V_{\rightarrow i}$ of points on $L_{\rightarrow}$ (a subnetwork) that satisfies equation (23.1), where $d(p, p_i)$ is replaced with $d_{\rightarrow}(p, p_i)$. The set of the resulting subnetworks, $V_{\rightarrow\bullet} = \{V_{\rightarrow 1}, \ldots, V_{\rightarrow m}\}$, is also a directed network Voronoi diagram. To distinguish $V_{\bullet\rightarrow} = \{V_{1\rightarrow}, \ldots, V_{m\rightarrow}\}$ and $V_{\rightarrow\bullet} = \{V_{\rightarrow 1}, \ldots, V_{\rightarrow m}\}$, the former is termed the *outward directed network Voronoi diagram* and the latter the *inward directed Voronoi diagram* (Okabe *et al.*, 2008). Both are directed Voronoi diagrams that

Figure 23.6 The ordinary planar Voronoi diagram generated from parking lots in Kyoto, Japan.

Figure 23.7 The ordinary network Voronoi diagram generated from parking lots in Kyoto, Japan.

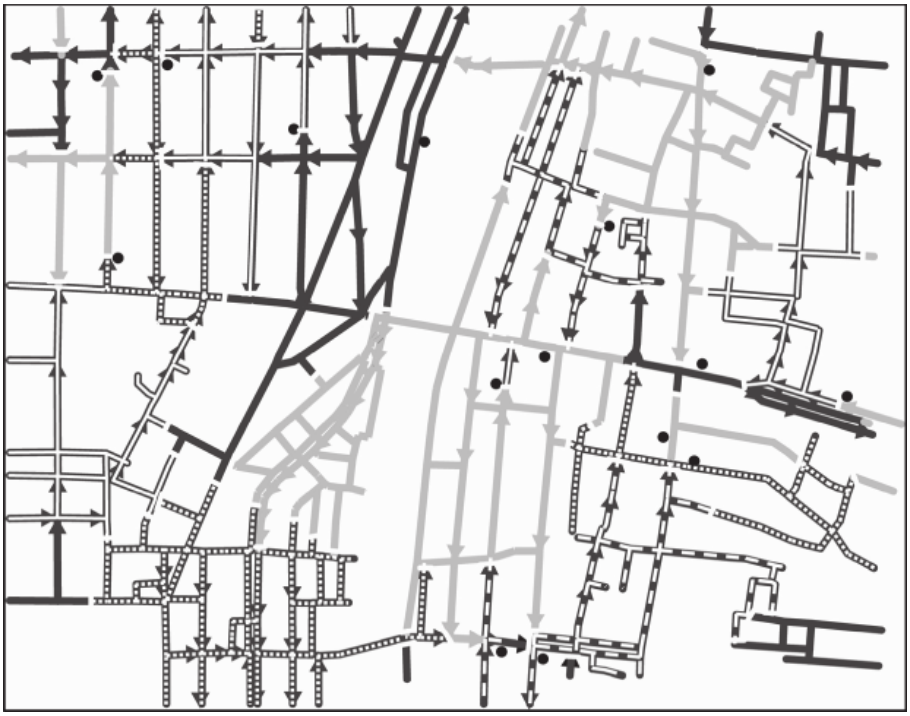


Figure 23.8 The outward directed Voronoi diagram generated from parking lots in Kyoto, Japan.

should be distinguished from the ordinary network Voronoi diagram $V = \{V_1, \ldots, V_m\}$ (alternatively, the ordinary network Voronoi diagram is termed a *nondirected* network Voronoi diagram). An example of an inward directed Voronoi diagram is shown in Figure 23.9.

23.3.3. Weighted network Voronoi diagrams

Consider, for instance, that consumers choose a store by considering prices η_i at alternative stores and the transportation cost $\beta d(p, p_i)$ between their house p and the store p_i, where β is the unit transportation cost. In this case, the market area of a store is defined by the set V_{AWi} of points on L (a subnetwork) that satisfies equation (23.1), where $d(p, p_i)$ is replaced with $d_{AW}(p, p_i) = \beta d(p, p_i) + \eta_i$, termed the *additively weighted network distance*. The set of the resulting subnetworks, $V_{AW} = \{V_{AW1}, \ldots, V_{AWm}\}$, is termed the *additively weighted network Voronoi diagram* (Okano and Okabe, 2004); an example is shown in Figure 23.10, where the dots indicate convenience stores in Kyoto, Japan, and the radius of each circle indicates its weight η_i.

Suppose that the delivery speed β_i of goods is different from store to store. In this case, the multiplicatively weighted distance, $d_{MW}(p, p_i) = \beta_i d(p, p_i)$, is appropriate for estimating market areas. To be explicit, let V_{MWi} be the set of points on L (a subnetwork) that satisfies equation (23.1), where $d(p, p_i)$ is replaced with $d_{MW}(p, p_i) = \beta_i\, d(p, p_i)$. The set of the resulting subnetworks, $V_{MW} = \{V_{MW1}, \ldots, V_{MWm}\}$, is termed the *multiplicatively weighted network Voronoi diagram* (Okano and Okabe, 2004). An example is shown in Figure 23.11, where the dots indicate convenience stores in Kyoto, Japan, and the radius of each circle indicates its weight β_i.

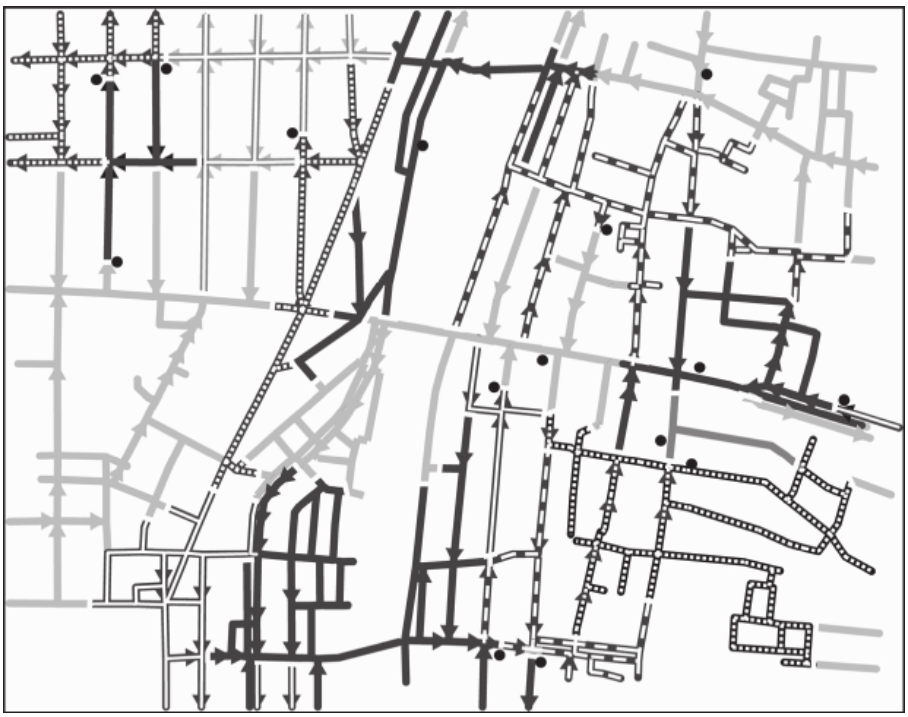


Figure 23.9 The inward directed Voronoi diagram generated from parking lots in Kyoto, Japan.

Figure 23.10 The additively weighted network Voronoi diagram generated from convenience stores in Kyoto, Japan (each circle indicates its weight η_i).

23.3.4. Other network Voronoi diagrams

In addition to the above network Voronoi diagrams, the kth nearest network Voronoi diagram, the network Voronoi diagram for line segments, and the network Voronoi diagram for polygons have also been proposed in the literature. The reader should consult Furuata *et al*. (2005) and Okabe *et al*., (2008) for information on these diagrams.

23.4. LOCAL AND GLOBAL NETWORK *K* FUNCTION METHODS

23.4.1. Global network auto K function

One of the most commonly used techniques in statistical spatial analysis is the K function method. Originally, the K function method was developed for points on a plane, and was termed the *planar K function method* (Ripley, 1976, 1977). Okabe and Yamada (2001) extended the planar K function method to the K function method for points on a network to develop the *network K function method*. To state this method explicitly, consider a network L on which points $p_1, \ldots, p_m$ are placed, and let $D_i(t)$ be a subnetwork of L in which the shortest-path distance between any point on $D_i(t)$ and p_i is less than or equal to t (the heavy lines in Figure 23.12; in the planar case, $D_i(t)$ corresponds to the disk centered at p_i with radius t truncated by a bounded global space). Let $K_i(t)$ be the number of points of $p_1, \ldots, p_m$ that are included in $D_i(t)$. In this term, a *network K function* is defined by

$$K(t) = \sum_{i=1}^{m} K_i(t). \qquad (23.2)$$

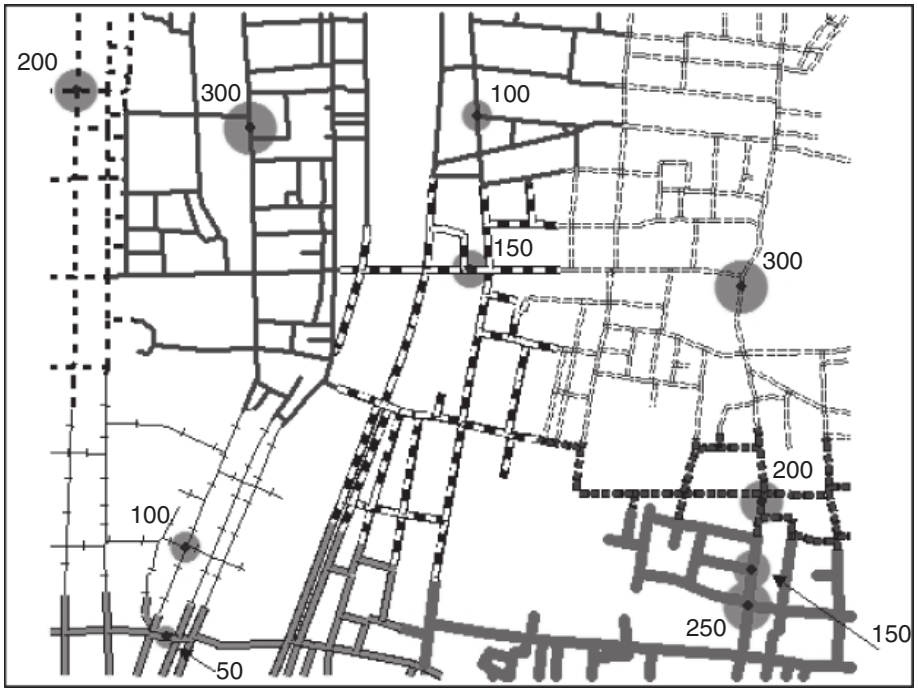


Figure 23.11 The multiplicatively weighted network Voronoi diagram generated from convenience stores in Kyoto, Japan (each circle indicates its weight β_i).

Note that, in contrast to the 'cross' K function, which is defined below, the above function is referred to as the network *auto* K function (as with spatial *auto* correlation); also note that constants (the density and number of points) are omitted here for simplicity.

To show an actual example, the distribution of street burglaries in Kyoto is depicted in Figure 23.12, where the triangle marks indicate sits of incidence. For this distribution, the network auto K function is calculated, and the result is shown in Figure 23.13. The black line indicates the expected value and the gray line indicates the observed value obtained under the null hypothesis that burglaries occur uniformly and randomly distributed on the street network. Because the observed curve is always above the expected curve in Figure 23.13, it is concluded that burglaries tend to cluster themselves.

The difference between the network K function and the planar K function is distinct. Actually, Yamada and Thill (2004) applied both the planar K function method and the network K function method to the same traffic accident data and found that the planar K function method overestimates clustering tendency. The authors concluded that the network K function method should be used for the analysis of traffic accidents.

Figure 23.12 Street burglaries (the triangle marks), railway stations (the circles), and the Voronoi sub-network (the heavy lines) of the station (the large circle) in Kyoto, Japan.

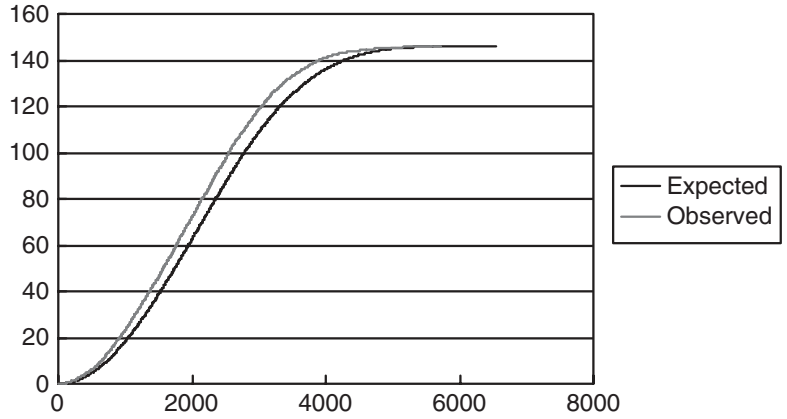


Figure 23.13 The global network auto *K* function for street burglaries on the street network in Kyoto, Japan (Figure 23.12).

23.4.2. Global network cross K function

Another type of network K function method is the network cross K function method (Okabe and Yamada, 2001). Consider two sets of points, $P = \{p_1, \ldots, p_m\}$ and $Q = \{q_1, \ldots, q_k\}$, on a network L. Points of P are stochastically distributed on L, but points of Q are fixed (note that the configuration of the points is arbitrary). For instance, points of P may be crime spots and points of Q may be railway stations. The network cross K function is used for testing whether points $p_1, \ldots, p_m$ (crime spots) tend to cluster around (or apart from) $q_1, \ldots, q_k$ (railway stations) as a whole.

To state the network cross K function explicitly, let $D_{q_i}(t)$ be a subnetwork of L in which the shortest-path distance between any point in $D_{q_i}(t)$ and q_i is less than or equal to t, and let $K_{q_i}(t)$ be the number of points of P that are included in $D_{q_i}(t)$. Then, the network cross K function, $K_{QP}(t)$, is defined by:

$$K_{QP}(t) = \sum_{i=1}^{k} K_{q_i}(t). \qquad (23.3)$$

(Note that a constant is omitted from the above equation for simplicity.) Because this function considers all points of P across the entire network space L (the global space), the function can be regarded as a *global network cross K function*. An actual example of the global network cross K function is shown in Figure 23.14, where points of P are street burglaries and points of Q are railway stations in Kyoto, Japan as shown in Figure 23.12. Because the observed curve is always above the expected curve in Figure 23.14, it is concluded that street burglaries tend to occur around railway stations.

23.4.3. Local network cross K function

The global network cross K function method deals with the average tendency of a point pattern around all fixed points Q; therefore, it cannot detect local tendencies. For example, the global network cross K function cannot detect the specific railway stations around which crime spots tend to cluster. To detect

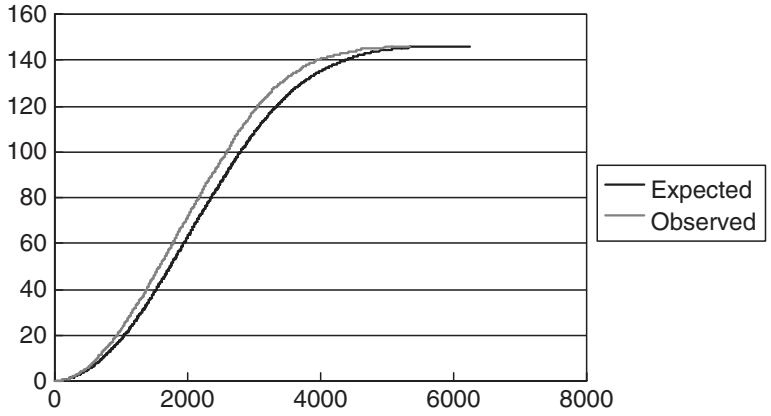


Figure 23.14 The global network cross *K* function for street burglaries in relation to railway stations in Kyoto, Japan (Figure 23.12).

local tendencies, a 'local' network cross K function should be developed.

A simple way of defining a local cross K function is to decompose the global cross K function given by equation (23.3) into each term; that is, a *local cross K function* defined by $K_{q_i}(t)$, $i = 1, \ldots, k$. This local cross K function, however, is not always 'natural', because the local space $D_{q_i}(t)$ of $K_{q_i}(t)$ becomes large as t increases, and eventually the local space includes the global space (the entire space L).

23.4.4. Local network Voronoi cross K function

In the context of crime spots and railway stations referred to above, it is natural to examine whether crime spots cluster in the neighborhood of specific railway stations. If commuters use their nearest railway stations, the neighborhoods of railway stations are given by the ordinary network Voronoi diagram generated from railway stations. To be explicit, let $V = \{V_1, \ldots, V_k\}$ be the ordinary network Voronoi diagram generated by Q. Then, a local space of point q_i (the neighborhood of the ith railway station) is given by the Voronoi subnetwork V_i (a Voronoi subnetwork is indicated by the heavy lines in Figure 23.12).

In terms of this natural local space, an alternative network cross K function $K_{Vq_i}(t)$ can be defined as the number of points of P (e.g., crime spots) that are included in $V_i \cap D_{q_i}(t)$, i.e., the number of crime spots in a local space $V_i \cap D_{q_i}(t)$ whose shortest-path distance to the railway station q_i is less than or equal to t. Because $V_i \cap D_{q_i}(t)$ is bounded by a local space V_i, the local space of $K_{Vq_i}(t)$ remains a local space of the global space even for a large t (this contrasts with the network cross K function $K_{q_i}(t)$ in Section 23.4.3). The function $K_{Vq_i}(t)$ is termed the *local network Voronoi cross K function*, and should be distinguished from the function $K_{q_i}(t)$ in Section 23.4.3, which is referred to as the local network *ordinary* cross K function. Figure 23.15 shows an actual example of the local network Voronoi cross K function for street burglaries in the local space indicated by the heavy lines in Figure 23.12.

23.4.5. Global network Voronoi cross K function

In sections 23.3.2 and 23.3.3, a local K function is obtained from a global K function. Conversely, a global function can also be obtained from a local function. For instance, let $K_{VQP}(t)$ be a function defined in terms of the local network Voronoi K functions as

$$K_{VQP}(t) = \sum_{i=1}^{k} K_{Vq_i}(t). \qquad (23.4)$$

This function deals with all points P in the entire network L (the global space). Therefore, this function can be regarded as a global network cross K function, which is referred to as the *global network Voronoi cross K function*. An example is illustrated in Figure 23.16. Comparison between the local network Voronoi cross K function in Figure 23.15 and the global network Voronoi cross K function in Figure 23.16 reveals local variety.

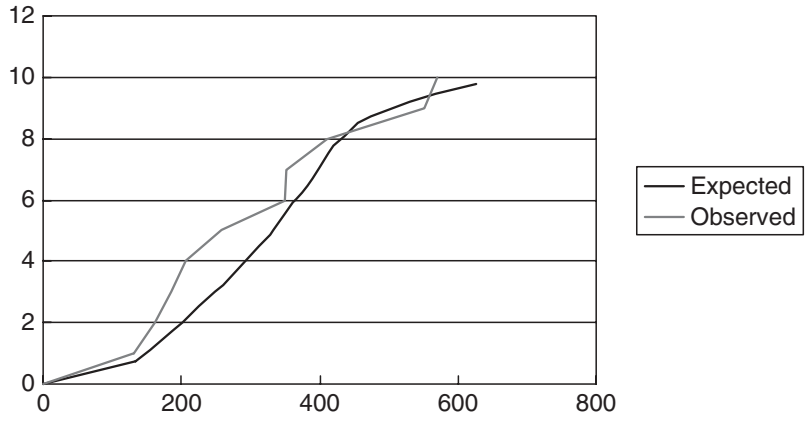


Figure 23.15 The local network Voronoi cross *K* function for street burglaries in the local space (the heavy lines in Figure 23.12) in Kyoto, Japan.

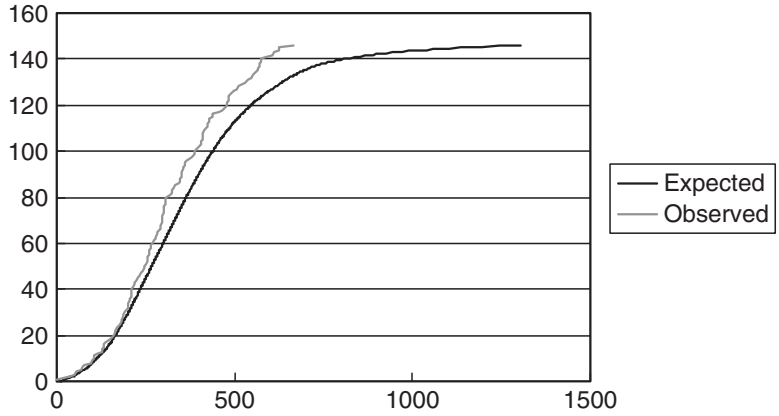


Figure 23.16 The global network Voronoi cross *K* function for street burglaries in relation to railway stations in Kyoto, Japan (Figure 23.12).

Having obtained two global network K functions, namely the global network *ordinary* cross K function in Section 23.4.3 and the global network *Voronoi* cross K function in Section 23.4.4, one might question the difference between them. Figure 23.17 provides an illustration of this difference. Figure 23.17(a) shows $D_{q1}(2)$ (heavy gray lines) and $D_{q2}(2)$ (heavy black lines). Points p_1, p_2, p_3, p_4, p_5, p_6, p_7 are included in $K_{q_1}(2)$, whereas p_3, p_4, p_6, p_7, p_8 are included in $K_{q_2}(2)$; consequently, the global network *ordinary* cross K function $K_{QP}(2) = K_{q_1}(2) + K_{q_2}(2)$ counts points p_1, p_2, p_8 once and points p_3, p_4, p_6, p_7 twice. In contrast, as shown in Figure 23.17(b) where $D_{q1}(2) \cap V_1$ (heavy gray lines) and $D_{q2}(2) \cap V_2$ (heavy black lines) are depicted (broken lines indicate the boundary points of the Voronoi subnetworks V_1 and V_2), the global network *Voronoi* cross K function $K_{VQP}(2) = K_{Vq_1}(2) + K_{Vq_2}(2)$ counts every point of P only once. These two global network cross K functions focus on different aspects of a point pattern. Actually, Figures 23.14 and 23.16 show this difference.

23.5. NETWORK KERNEL METHOD

The kernel method is a nonparametric method for estimating a density function from a given set of observed values (Silverman, 1986). Kernel functions are usually defined for univariate or bivariate density functions. In the context of spatial analysis, the kernel method is applied to the density of points on a plane and is commonly employed to detect 'hot spots' (or 'cold spots'), e.g., highly concentrated areas of crime occurrences within a city. If the crime of interest is street burglaries, the density of street burglaries on streets is to be estimated. This section demonstrates a kernel method for estimating a density function on a network (Okabe *et al.*, 2009).

A simple way of estimating a density function from given observed points is to use kernel functions defined on a plane, referred to as *planar kernel functions*. Suppose that the coordinates (represented by vectors) of observed points on a network L embedded on a plane are $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, and let $k(\boldsymbol{x})$ be a two-dimensional kernel function defined on a plane. Then, the

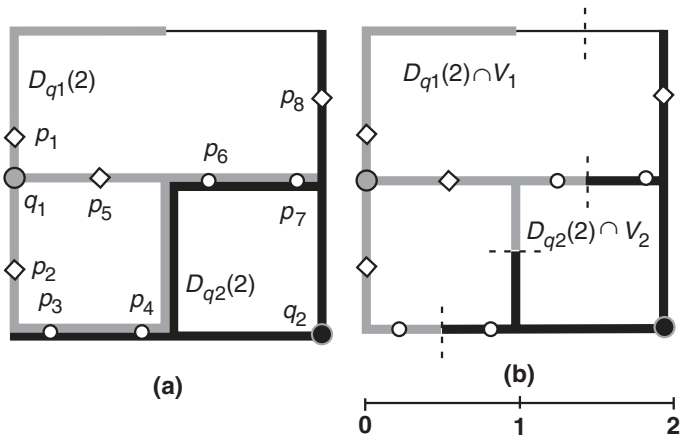


Figure 23.17 Comparison between the global network *ordinary* cross *K* function (a) and the global network *Voronoi* cross *K* function (b).

estimated density function, $f(\boldsymbol{x})$, on a plane is given by:

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} k(\boldsymbol{x}_i). \qquad (23.5)$$

An example is presented in Figure 23.18, where one million points are uniformly and randomly generated and the kernel function is given by the bi-weight function (Silverman, 1986).

One might estimate the density function, $f_L(\boldsymbol{x})$, of points on L from the intersection of $f(\boldsymbol{x})$ with L, i.e., $f_L(\boldsymbol{x}) = f(\boldsymbol{x}),\ \boldsymbol{x} \in L$. This method would be fine if the estimated density function could produce a uniform distribution function for the points that are uniformly and randomly distributed on the network. However, Figure 23.18 shows that $f_L(\boldsymbol{x})$ does not show a uniform distribution for such points (one million). Therefore, this method is inappropriate.

An alternative method is to use a *network kernel function*, $k_L(t) = k_L(\boldsymbol{x}), \boldsymbol{x} \in L$, defined on L. An example is shown in Figure 23.19, where one million points are uniformly and randomly generated and the density function is estimated from those points using the one dimensional bi-weight function.

This appears to be a natural extension of the planar kernel method, but Figure 23.19 proves that this method is inappropriate. As the points in Figure 23.19 are uniformly and randomly generated on L, the estimated density should be uniform; however, the density in Figure 23.19 is not uniform on L, which suggests that this method is inappropriate. One reason that the natural extension of the planar kernel method does

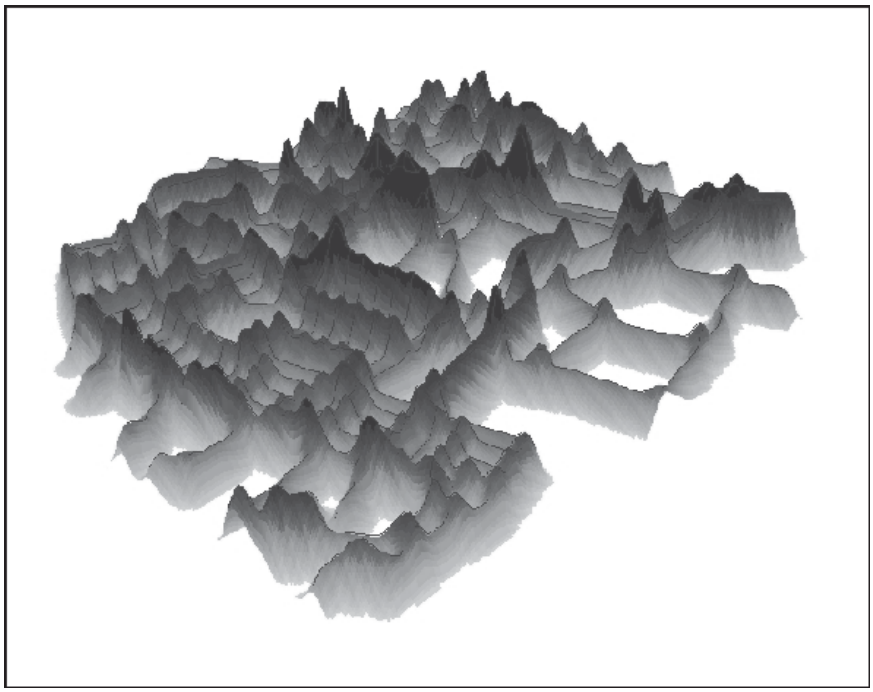


Figure 23.18 The density function for uniformly random points (one million) on the street network in Kyoto estimated by the two-dimensional bi-weight kernel function.

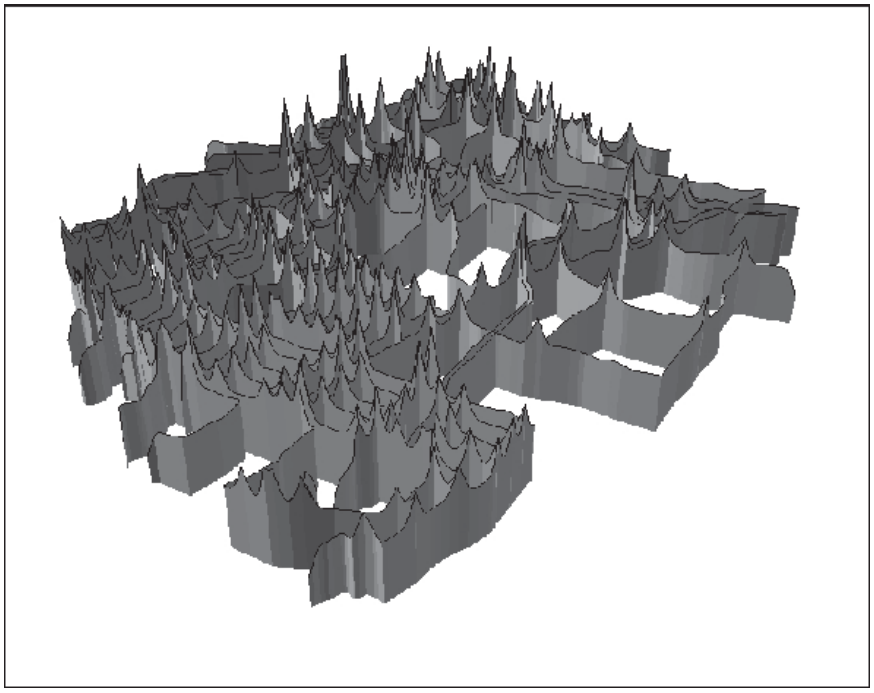


Figure 23.19 The density function for uniformly random points (one million) on the street network in Kyoto estimated by the one-dimensional bi-weight kernel function.

not work is that a plane is isotropic whereas a network is not isotropic in the sense that directions are restricted and is bounded. Okabe *et al*., (2009) provide two kernal functions $K_i(t)$ that produces a uniform density function for uniformly and randomly distributed points.

Once a density function has been estimated, it is easy to find 'hot spots'. Let $L(u)$ be a subnetwork of L that satisfies $f_L(t) \geq u$, and let L_α be the subnetwork $L(u)$ that satisfies:

$$\int_{t \in L(u)} f_L(t)\, \mathrm{d}t \Big/ \int_{t \in L} f_L(t)\, \mathrm{d}t = 100\alpha. \quad (23.6)$$

The subnetwork L_α is the area of *hot spots*; the probability of points occurring on the subnetworks L_α is high at the significance level α.

Figure 23.21 shows hot spots of traffic accidents in Chiba, Japan, determined by using the above method and assuming that the given network is a uniform network, i.e., the probability of an accident occurring in a unit line segment is constant regardless of the location of the unit line segment. As noted in section 23.2, however, it is more likely that traffic accidents tend to occur in proportion to traffic volume, as shown in Figure 23.1 (a nonuniform network). To examine this tendency of accident hot spots, the uniform network transformation in section 23.2 is applied to this nonuniform network, and the network kernel method is applied to the resulting uniform network. Figure 23.22 shows hot spots of traffic accidents for the transformed network. These hot spots indicate the places where traffic accidents tend to occur more frequently than would be expected from the measured traffic volumes.

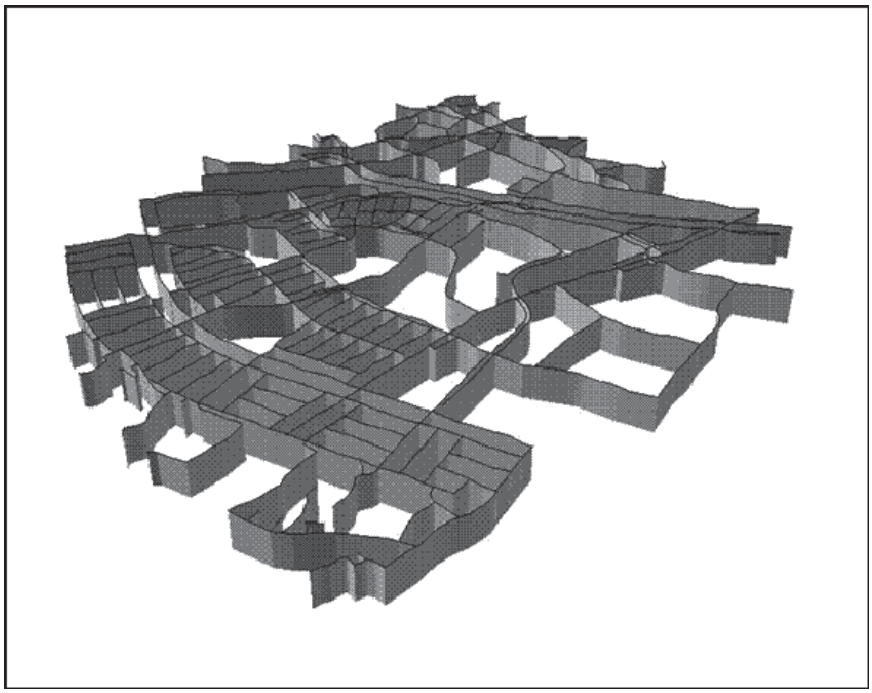


Figure 23.20 The density function for uniformly random points (one million) on the street network in Kyoto estimated by a one-dimensional modified bi-weight kernel function.

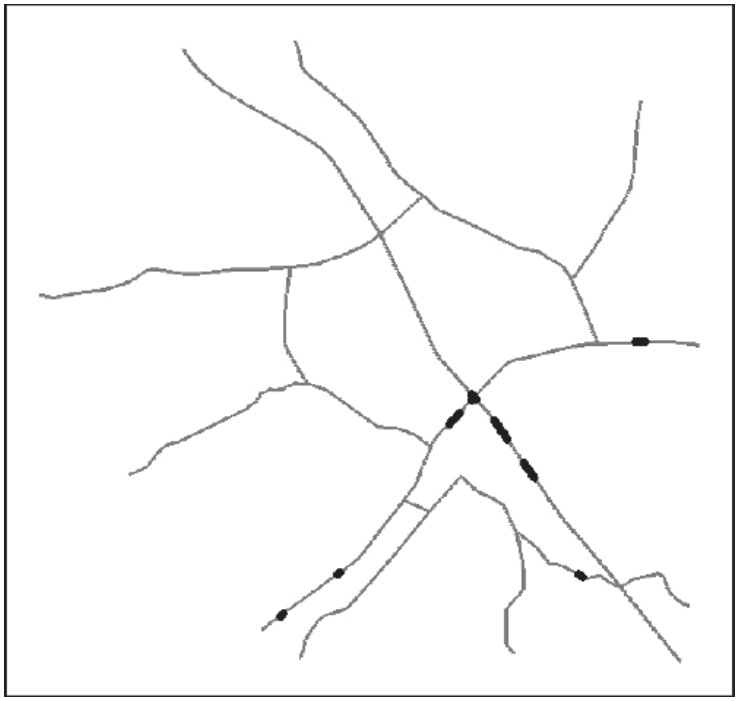


Figure 23.21 Hot spots of traffic accidents on the uniform road network in Chiba, Japan.

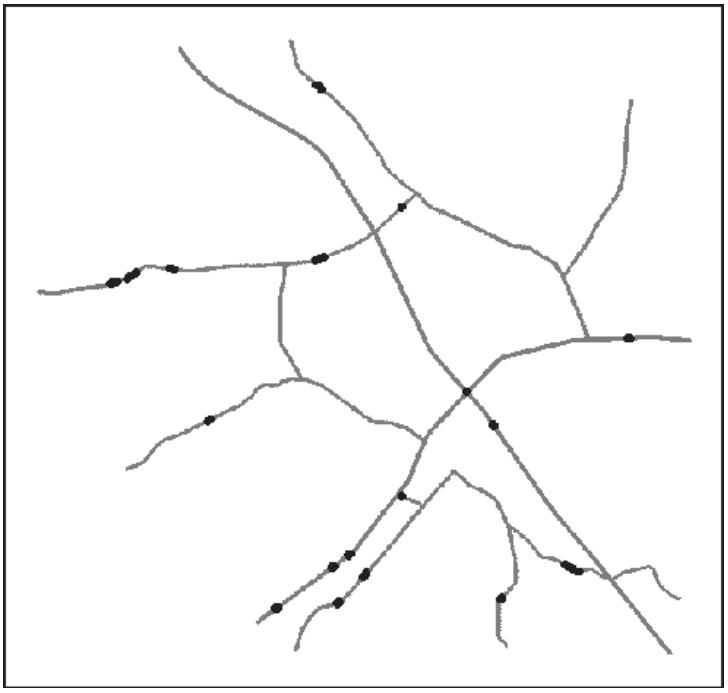


Figure 23.22 Hot spots of traffic accidents on the nonuniform road network in Chiba, Japan that takes account of traffic volume.

23.6. GIS-BASED TOOLS FOR NETWORK SPATIAL ANALYSIS, SANET

To analyze network spatial phenomena in theory, network spatial methods are more appropriate than planar spatial methods. In practice, however, developing computer programs for network spatial methods is more difficult than that for planar spatial methods. Fortunately, this difficulty is overcome by a free software package termed SANET (Spatial Analysis on a NETwork), which can be downloaded from http://okabe.t.u-tokyo.ac.jp/okabelab/atsu/sanet/sanet-index.html.

The functions are outlined in Okabe *et al*. (2006a, b), and the manual is available from the above site (Okabe *et al*., 2004). Note that Version 3 is the current release but functions are updated from time to time.

23.7. CONCLUSIONS

In the real world, there are many network spatial phenomena. Planar spatial methods are inappropriate for analyzing these phenomena because they commonly lead to false conclusions. To avoid such false conclusions, network spatial methods should be used. This chapter considered three classes of network spatial methods: (1) a class of network Voronoi diagrams that includes the nondirected Voronoi diagram, the inward directed Voronoi diagram, the outward directed Voronoi diagram, the additively weighted Voronoi diagram, and the multiplicatively weighted Voronoi diagram; (2) a class of network K functions that includes the global auto K function, the global ordinary cross K function, the local ordinary cross K function, the local Voronoi cross

K function, and the global Voronoi cross *K* function; and (3) a class of network kernel methods that includes a method for detecting hot spots.

In addition to the above network spatial methods, Okabe *et al*. (1995) formulated the network of the (conditional) nearest neighbor distance method; Miller (1994, 1999), Okabe and Kitamura (1996), Okabe and Okunuki (2001), and Morita *et al*. (2001) formulated the network Huff model; Shiode and Okabe (2004a) formulated the network clumping method; Shiode and Okabe (2004b) formulated the network cell count method; and Okabe *et al*. (2006b) proposed the network spatial interpolation method. There are many other planar spatial methods that have not yet been extended to network spatial methods. Hopefully, the readers of this chapter will extend these methods and enrich the field of network spatial analysis.

ACKNOWLEDGMENTS

We express our thanks to Barry Boots, Kei-ich Okunuki, Shino Shiode, Kyoko Okano, Ikuho Yamada and Takashi Maki for their comments on an earlier draft.

REFERENCES

Anselin, L., Cohen, J., Cook, D., Gorr, W. and Tita, G. (2000). Spatial analyses of crime. In: David Duffee (ed.), *Criminal Justice 2000: Volume 4. Measurement and Analysis of Crime and Justice*, pp. 213–262. Washington, DC: National Institute of Justice.

Bashore, T., Tzilkowski, W. and Bellis, E. (1985). Analysis of deer–vehicle collision sites in Pennsylvania. *Journal of Wildlife Management*, **49**(3): 769–774.

Bowers, K. and Hirschfield, A. (1999). Exploring links between crime and disadvantage in north-west England: an analysis using geographical information systems. *International Journal of Geographical Information Science*, **13**: 159–184.

Clevenger, A.P., Chruszcz, B. and Gunson, K.E. (2003). Spatial patterns and factors influencing small vertebrate fauna road-kill aggregations. *Biological Conservation*, **109**: 15–26.

Freund, J.E. (1998). *Mathematical Statistics*, (6th edn), Englewood Cliff: Prentice-Hall.

Furuta, T., Suzuki, A. and Inakawa, K. (2005). The *k*th nearest network Voronoi diagram and its application to districting problem of ambulance systems, *Discussion Paper* No. 0501, Center for Management Studies, Nanzan University.

Jones, A.P., Langford, I.H. and Betham, G. (1996). The application of *K*-function analysis to the geographical distribution of road traffic accident outcomes in Norfolk, England. *Social Science and Medicine*, **42**(6): 879–885.

Levine, N., Kim, K.E., Nitz, L.H. (1995). Spatial Analysis of Honolulu Motor Vehicle Crashes: I. Spatial Patterns. *Accident Analysis and Prevention*, **27**(5): 663–674.

Maki, N. and Okabe, A. (2005). Spatio-temporal analysis of aged members of a fitness club in a suburbs. *Proceedings of the Geographical Information Systems Association*, **14**: 29–34.

Mallick, S.A., Hocking, G.J. and Driessen, M.M. (1998). Road-kills of the eastern barred bandicoot (*Perameles gunnii*) in Tasmania:an index of abundance. *Wildlife Research*, **25**: 139–145.

McGuigan, D.R.D. (1981). The use of relationships between road accidents and traffic flow in 'black-spot' identification. *Traffic Engineering and Control*, **22**: 448–453.

Miller, H.J. (1994). Market area delimitation within networks using geographic information systems. *Geographical Systems*, **1**: 157–173.

Miller, H.J. (1999). Measuring space-time accessibility benefits within transportation networks. *Geographical Analysis*, **31**(2): 187–212.

Morita, M., Okunuki, K. and Okabe, A. (2001). A market area analysis on a network using GIS – A case study of retail stores in Nisshin city. *Papers and Proceedings of the Geographic Information Systems Association*, **10**: 45–50.

Nicholson, A.J. (1989). Accident clustering: Some Simple Measures. *Traffic Engineering and Control*, **30**: 241–246.

O'Driscoll, R.L. (1998). Descriptions of spatial pattern in seabird distributions along line transects using neighbor *K* statistics. *Marine Ecology Progress Series*, **165**: 81–94.

Okabe, A., Boots, B. and Satoh, T. (2006). A class of local and global *K*-functions and cross *K*-functions, *The 2006 Annual Meeting of the AAG*, March 7–11, 2006, Chicago, IL.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S.N. (2000). *Spatial Tessellations*: *Concepts and Applications of Voronoi Diagrams* (2nd edn). Chichester: John Wiley.

Okabe, A. and Kitamura, M. (1996). A computational method for market area analysis on a network. *Geographical Analysis*, **28**: 330–349.

Okabe, A. and Okunuki, K. (2001). A Computational method for estimating the demand of retail stores on a street network using GIS. *Transactions in GIS*, **5**(3): 209–220.

Okabe, A., Okunuki, K. and Shiode, S. (2004). SANET: A toolbox for spatial analysis on a network – Version 2.0 – 040102. Center for Spatial Information Science, University of Tokyo, Tokyo.

Okabe, A., Okunuki, K. and Shiode, S. (2006a). SANET: a toolbox for spatial analysis on a network. *Geographical Analysis*, **38**(1): 57–66.

Okabe, A., Okunuki, K. and Shiode, S. (2006b). The SANET toolbox: new methods for network spatial analysis. *Transactions in GIS*, **10**: 535–550.

Okabe, A. and Satoh, T. (2006). Uniform network transformation for points pattern analysis on a non-uniform network. *Journal of Geographical Systems*, **8**(1): 25–37.

Okabe, A., Satoh, T., Furuta, T., Suzuki, A., Okano, A. (2008). Generalized network Voronoi diagrams: Concepts, computational methods, and applications. *International Journal of Geographical Information Science*, 1–30.

Okabe, A., Satoh, T and Sugihara, K. (2009) A kernel density estimation method for networks, Its computational method and a GIS-based tool, *International Journal of Geographical Information Science* (to appear).

Okabe, A., Satoh. T. and Sugihara, K. (2009). A kernel density estimation method for networks, its computational method, and a GIS-based tool. *International Journal of Geographical Information Science* (to appear).

Okabe, A. and Yamada, I. (2001). The *K*-function method on a network and its computational implementation. *Geographical Analysis*, **33**: 271–290.

Okabe, A., Yomono, H. and Kitamura, M. (1995). Statistical analysis of the distribution of points on a network. *Geographical Analysis*, **27**(2): 152–175.

Okano, K. and Okabe, A. (2004). Algorithms for computing weighted network Voronoi diagrams. *Papers and Proceedings of the Geographic Information Systems Association*, **13**: 311–314.

Painter, K. (1994). The impact of lighting on crime, fear, and pedestrian street use. *Security Journal*, **5**: 116–124.

Ratcliffe, H.J. (2002). Aoristic signatures and the spatio-temporal analysis of high volume crime patterns. *Journal of Quantitative Criminology*, **18**: 23–43.

Ratcliffe, J.H. and McCullagh, M.J. (1999). Hotbeds of crime and the search for spatial accuracy. *Journal of Geographical Systems*, 1: 385–398.

Ripley, B.D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, **13**: 255–266.

Ripley, B.D. (1977). Modeling spatial patterns. *Journal of the Royal Statistical Society, Series B*, **39**: 172–192.

Saeki, M. and MacDonald, D.W. (2004). The effects of traffic on the raccoon dog (*Nyctereutes procyonoides viverrinus*) and other mammals in Japan. *Biological Conservation*, **118**: 559–571.

Shiode, S. and Okabe, A. (2004a). Network variable clumping method for analyzing point patterns on a network. *The 2004 Annual Meeting of the AAG*, Philadelphia, PA.

Shiode, S. and Okabe, A. (2004b). Cell count method on a network with SANET and its application. *International Conference on Geoinfomatics and Geographical Systems Modelling*, Beijng, China.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Spooner, P.G., Lunt, I.D., Okabe, A. and Shiode, S. (2004). Spatial analysis of roadside Acacia populations on a road network using the network *K*-function. *Landscape Ecology*, **19**(5): 491–499.

Yamada, Y. and Thill, J.-C. (2004). Comparison of planar and network *K*-functions in traffic accident analysis. *Journal of Transport Geography*, **12**: 149–158.

24

Challenges in Spatial Analysis

Michael F. Goodchild[1]

24.1. INTRODUCTION

This is a time of unprecedented opportunity for spatial analysis. More people than ever have access to the Global Positioning System for direct measurement of location on the Earth's surface; to the products of high-resolution remote-sensing satellites; and to the manipulative power of geographic information systems (GIS). Some of these technologies are encountered in everyday life, through sites such as Google Earth (earth.google.com), Google Maps (maps.google.com), and Microsoft Windows Live Local (live.local.com), and through the widespread use of in-vehicle navigation systems. Several academic disciplines are recording a *spatial turn*, a new and in some cases renewed interest in space and location as a framework for analysis, understanding, and presentation of results. A recent publication (National Research Council, 2006) has defined and explored *spatial thinking* as a paradigm for primary and secondary education, and projects have been funded around the world to advance *spatial literacy* (e.g., www.spatial-literacy.org).

At the same time the field faces substantial challenges, as it attempts to take advantage of these new opportunities. This chapter addresses four: the challenge raised by the continuing rapid advance in computing and networking technology; the challenge of addressing the temporal dimension through the analysis of dynamic phenomena; the challenge posed by the immense popularity of Web sites that offer rudimentary forms of spatial analysis to a user community that has little or no formal educational background in this area; and the challenge of formulating a new philosophy of science that reflects the actual conditions under which spatial analysis is used in today's research and problem-solving environments.

The four topics by no means exhaust the full set of issues facing the field.

Many readers will have their own ideas, and the chapter on the future of spatial analysis that follows include discussions of additional issues. Meanwhile, the four considered in this chapter are very much a personal list, and reflect the author's own interests and concerns at this point in the long history of spatial analytic methods.

24.2. COMPUTING AND NETWORKING TECHNOLOGY

In the early 1990s a substantial literature accumulated on the opportunities offered by GIS. In 1988 the U.S. National Science Foundation had established the National Center for Geographic Information and Analysis (NCGIA) at three sites: the University of California, Santa Barbara; the State University of New York at Buffalo; and the University of Maine. One of NCGIA's objectives was to advance the use of GIS across the sciences, as a platform for spatial analysis, so it was considered important to assess progress to date, and to identify and remove impediments to the greater use of spatial analysis. NCGIA organized a specialist meeting on the topic that eventually led to a book (Fotheringham and Rogerson, 1994), and several additional papers appeared (Anselin and Getis, 1992; Burrough, 1990; Ding and Fotheringham, 1992; Goodchild, 1987; Goodchild *et al.*, 1992; Openshaw, 1990; and for a later perspective see Goodchild and Longley, 1999).

Underlying this spate of funding and writing was the simple premise that GIS provided an ideal means of implementing the known techniques of spatial analysis, as well as techniques that might be developed in the future. A single package, if sufficiently sophisticated, could offer easy and largely painless access to an abundance of robust, scientifically sound techniques for analyzing and visualizing spatial data. The results of each stage of analysis could be fed into further stages, and data could be managed within a single environment that recognized a range of data formats. Comparisons were frequently drawn with the statistical packages (e.g., Goodchild, 1987), which similarly offered easy access to a multitude of statistical techniques, along with the necessary housekeeping functions.

At the time, each GIS software product was organized into a single, monolithic package. In the 1980s such packages were typically installed on minicomputers such as the VAX or Prime, but in the late 1980s the transition to personal Unix workstations and later to the PC and Mac had opened the possibility of an entirely individualized toolbox installed on the researcher's desk. GIS was likened to a butler – an intelligent assistant working with the user to solve problems, knowing the foibles and preferences of the user, and taking on those tasks that the user found too complex, tedious, time-consuming, or inaccurate if performed by hand. Abler (1987) hailed GIS as geography's equivalent of the microscope or the telescope, a powerful tool that allowed researchers to gain insights that were simply impossible with the normal senses and intuition.

From this perspective, the power of GIS would be judged simply by the proportion of known techniques of spatial analysis that it supported, by the accuracy with which it implemented each method, and by the extent to which it prevented misuse and misinterpretation of results. There were many complaints about this time regarding the success of GIS against these objectives. Commercial software developers were seen as insufficiently interested in supporting advanced spatial analysis, being content instead to direct their efforts at satisfying the needs of their more wealthy corporate and agency customers, whose interests

tended to be more in data management and inventory. GIS designers failed to ground their products in sound theory, preferring intuitive terms and explanations over formal and mathematical ones. Because of this lack of formal grounding, each vendor tended to adopt its own terms, formats, and structures, leading to endless proliferation and an apparently insurmountable lack of interoperability.

It was in this context that the Web appeared on the scene, and the Internet emerged as the dominant and indeed quickly the only network for computer communication. Since 1993 and the release of Mosaic the impact of communications technology has been so profound as to change the entire landscape of GIS and spatial analysis. Sui and Goodchild (2001) have argued that the metaphor of the butler is no longer appropriate – instead, GIS technology now constitutes a medium through which people communicate what they know about the Earth's surface that is comparable to traditional media such as print, radio, and television. As such, its issues are dramatically different from those of earlier decades. Bandwidth, interoperability, and metadata have largely replaced computing speed, storage capacity, and the sophistication of desktop software as major concerns of GIS users. Even the most sophisticated of users no longer program, relying instead on the incredibly abundant resources of the Web, easy mechanisms for sharing code, and new forms of software architecture. The following three sections explore some of these issues, and their implications for spatial analysis.

24.2.1. Server GIS

In the client–server computing paradigm that underlies the Web, the user or client's hardware and software are comparatively simple or *thin*, and most actual computation occurs remotely on a more powerful server. In the extreme, the user needs only a Web browser such as Microsoft Explorer or Netscape. Instead of installing a *thick* piece of software, such as a GIS package, the user obtains many if not all of its services from a remote server. For example, the task of finding the optimum route from an origin to a destination through a street network, the task performed by many Web sites such as mapquest.com, no longer requires the user to obtain a powerful GIS and the necessary database representing roads and streets, and to mount both on his or her desktop machine, since the same service can be obtained free from the server. The user need only specify the origin and destination to the server using a Web browser; the results are then sent back from the server and displayed locally using the same Web browser.

In principle all GIS functions and all types of spatial analysis could be organized in this way. Instead of installing and operating their own software, researchers could send data to sites where sophisticated forms of spatial analysis were performed. Researchers developing new forms of spatial analysis would find it far easier to offer their techniques as Web services than to engage in the time-consuming distribution of software, and users would benefit by not having to spend time obtaining, installing, and maintaining their own copies.

Server GIS is now common among public agencies interested in providing public access to their spatial data, along with simple capabilities for query and visual display. Many local governments provide access to their land-ownership and property taxation databases in this way, allowing users to query details of their own and other properties, using a map interface.

In practice, however, server GIS has had a limited impact to date, particularly for

more sophisticated analysis, for a number of reasons:

- There is no consensus on the appropriate business model for server GIS. Desktop GIS software generates income for its developers through sales and licensing, providing a healthy income stream, and developers of new methods of spatial analysis have sometimes used this same approach. Users of server GIS typically expect services to be provided free, leaving the providers of such services to generate income through advertising and the licensing of services to third parties. Routing services, for example, can be found embedded in the Web sites of on-line travel agencies and real-estate companies, presumably at some cost to them. Moreover, the software for server GIS tends to be more expensive per copy than conventional desktop GIS software (although open-source packages are available, e.g., mapserver.gis.umn.edu).

- Server GIS is most effective when the volume of data that needs to be input by the user is limited, and when the data needed are common to a large number of users and applications. A routing service, for example, requires only an origin and destination, and uses a generic database of streets and roads stored at the server. Moreover, such databases change frequently, and there are enormous economies of scale if all users can rely on a single version. Geocoding or address matching, the task of converting street addresses to coordinates, has become a popular function for server GIS for the same reason.

- Lack of interoperability continues to be an issue for server GIS. There are no standards for the description of services, though several *geo-portals* now provide limited directories (Maguire and Longley, 2005; Goodchild, *et al.*, in press). Extensive reformatting may be needed to make data readable by remote services, and the results returned may similarly need to be reformatted to be useful locally.

The choice between local and server-based computing is a complex one, and developers and implementers of spatial-analytic routines will need to consider the options carefully. However, it is clear that the nature of computing is changing, as many services move to a central, server-based model.

24.2.2. Process scripts

Research tends to proceed in stages, as problems are formulated, data are collected and checked, analysis is performed, and results are scrutinized. Each stage feeds forward to the next, and also back to the previous stages, as projects are rethought and as hypotheses are tested and modified. By the time the project is finally completed, the investigator may well have lost track of some of the stages, and may find it difficult to provide the necessary details in publications and reports. Somewhat paradoxically, the research community has invested heavily in the infrastructure to create and share data, and in the software to process them, but has not made similar investments in the techniques for management of the research process. The problem grows more severe as research becomes more collaborative, with many participants who may or may not communicate in person, and as the tools of research become more complex.

Against this background it is not surprising that many vendors of GIS and spatial-analytic packages have created *macro-* or *scripting* languages that allow researchers to express complex analyses as sequences of operations, and to store, manage, and execute such sequences as simple commands. A script in digital form is immediately more easily shared, managed, and documented than its equivalent in the jotted and invariably incomplete hand-notes of the researcher. Modern scripting languages allow complex hierarchical structures, since a single line in a script can invoke other scripts and programs, and allow sequences of operations to be

repeated many times in such applications as Monte Carlo simulation.

However, the design of an appropriate scripting language is a very sophisticated task, requiring a high level of knowledge of the needs of the research community, across many disciplines and domains. Simple scripting languages merely allow the user to invoke any of the commands of the package, but more sophisticated languages imply a recognition of the fundamental elements from which complex spatial analyses are built. If the granularity of the scripting language is too coarse, researchers will find it too difficult to express the full range of applications – and if it is too detailed, the script will be unnecessarily long.

The work of Tomlin (1990) provided the first successful effort at a generic scripting language for GIS, albeit only for congruent layers of raster data. The language was adopted by several packages, and several extensions were made. Van Deursen (1995) analyzed the operations required to support dynamic modeling in a raster environment, including the implementation of finite-difference models, in what became the scripting language for PCRaster (pcraster.geo.uu.nl), a raster-based package heavily oriented towards environmental modeling. Takeyama and Couclelis (1997) described a sophisticated language for the manipulation of pairs of raster cells, providing support for the analysis of spatial interactions. More broadly, all of these approaches are strongly related to the languages developed in image processing, or *image algebras*.

To date, however, there have been no comparably ambitious efforts to devise languages for vector data, or for the broader framework that spans both discrete objects and continuous fields. Dynamic GIS that addresses both space and time also lacks comprehensive scripting languages. The effectiveness of future spatial analysis clearly depends on the community's ability to devise simple yet comprehensive languages that can be used to describe and share computational methods. In the past, mathematics provided an adequate language, and models were effectively shared using algebraic representation, through the pages of learned journals and books. But today's computational environments present a somewhat different problem, since the language of mathematics lies too far from actual implementation, and cannot readily be used to express the entire algorithmic basis of spatial analysis.

24.2.3. Interchangeable software components

Early computer software was comprised of *programs*, integrated pieces of software that performed well-defined functions. Early GIS developed in this context, and by the early 1990s a fully featured GIS such as ESRI's ARC/INFO included millions of lines of code, all designed to be compiled and executed together to provide a single, integrated computing environment.

This approach to software was both redundant, in the sense that large amounts of code might never be executed by a given user, whose interests might focus only on a small number of functions; and costly, in the sense that it was difficult for programmers to pull pieces of code out of one package to be reused in another. Even today, the average user of a package such as Microsoft Word will likely never have invoked many of the functions in this very large and complex package.

Several attempts to break out of this mold were made in the 1980s and 1990s. One of the more successful was the concept of a *subroutine library*, a collection of standard routines that could be *called* by programs, avoiding the need for repetitive reprogramming. Subroutine libraries became common in areas such as statistics, since they

allowed comparatively sophisticated users to develop new programs quickly, relying on standard subroutines for many of the program's functions. The idea was difficult to implement for less sophisticated users, however, since it required each to possess a substantial knowledge of programming.

Contemporary approaches to software emphasize a rather different approach, in which sections of reusable code, or *components*, can be freely combined during the execution of a program. Standards have been developed by vendors such as Microsoft that allow compliant components to be freely linked and executed. Ungerer and Goodchild (2002) describe one such application, in which ESRI's ArcGIS and Microsoft's Excel have been combined to solve a standard problem in areal interpolation (Goodchild, *et al*., 1993). Functions that are native to the GIS, such as polygon overlay, are obtained from ArcGIS, while operations on tables, such as matrix multiplication, are obtained from Excel. The entire analysis is invoked through commands written in Visual Basic, a form of scripting language, though other general scripting languages such as Python might also be used. Both packages are compliant with the Microsoft COM standard, allowing the components that form the building blocks of each to be freely combined and executed.

Approaches such as these are breaking down the barriers that previously existed between different types of software – in this case, ArcGIS and Excel – and allowing much more flexible forms of analysis. They invite an entirely new approach to software design, in which fundamental components with widespread application are combined to meet the needs of specific applications. They also call for answers to a fundamental question: what are the basic building blocks of spatial analytic software, and to what extent are the operations invoked by each form of analysis common to more than one form? Perhaps they will lead eventually to a new approach to teaching in spatial analysis, in which these fundamental building blocks are the elements of a course, rather than the analytic methods themselves.

24.3. TIME AND DYNAMICS

Many authors have commented on the generally static nature of GIS, and the difficulty of representing time and dynamic phenomena. Most attribute this to the legacy of the paper map, which inevitably emphasizes those aspects of the Earth's surface that remain relatively static, over such dynamic phenomena as events, transactions, and flows. Several comprehensive reviews have appeared, and much progress has been made in building spatial databases that include time (Langran, 1993; Peuquet, 1999, 2001, 2002).

This same emphasis on the static is evident in the toolkit of spatial analysis, with its focus on cross-sectional data. In part this is due to the difficulty of creating and acquiring longitudinal data; to the administrative difficulties that statistical agencies face in funding and maintaining data-collection programs through time; to the changing nature of the Earth's surface, and the impact that this has on data-collection procedures and the definitions of reporting zones; and to the changing nature of human society, and its notoriously short attention span. Efforts such as the National Historic GIS project (www.nhgis.org) have attempted to overcome these difficulties, building systems that allow users to construct longitudinal series from the census for example, but they remain comparatively few and far between.

While much progress has been made, the analysis of spatio-temporal data remains a comparatively underexplored area, and a source of substantial challenges for the

community. The next two subsections address two of these in greater detail.

24.3.1. Fundamental laws

Much of the nature of GIS and many of the architectural choices that have been made over the past several decades are ultimately attributable to the nature of the data themselves – the ways in which spatial data are special. Anselin (1989) has identified two general characteristics, and Goodchild (2003) has discussed several more.

Spatial dependence describes the widely observed tendency for the variance of spatial data to increase with distance. To paraphrase Tobler (1970), nearby things are more similar than distant things, a principle that has become known as the First Law of Geography (Sui, 2004). All of the methods used to represent geographic phenomena in GIS are to some extent reliant on the validity of this principle. For example, there would be no value in representing topography with isolines if elevation did not vary smoothly, and there would be no value in aggregating areas into contiguous regions if the latter could not be designed with relatively low within-region variance.

Anselin's second principle is spatial heterogeneity, the tendency for the Earth's surface to exhibit spatial non-stationarity. All of the various techniques developed over the past two decades for *local* spatial analysis are based on this principle, since they attempt to summarize what is true locally, rather than what is true globally. The Geographically Weighted Regression of Fotheringham, *et al*. (2002) falls into this category, as do the LISA technique of Anselin (1995) and the local statistics of Getis and Ord (1992).

If such principles are generally true of spatial data, and are useful in guiding the development of computational systems, then one might reasonably ask whether similar principles exist for spatio-temporal data, and whether such principles might usefully inform the development of a more dynamic approach to GIS and spatial analysis. What is the spatio-temporal equivalent of Tobler's First Law, for example? Does spatial heterogeneity apply also in time? What relationships exist between the parameters of spatio-temporal and spatial dependence and heterogeneity? Are other general principles of spatio-temporal phenomena waiting to be discovered?

24.3.2. Dynamic form

Spatial dependence and spatial heterogeneity are both properties of how the Earth's surface *looks*, capturing aspects of its form. Studies of form have a long history in science, but have given way in the long term to a desire to understand process – to understand how systems *work*, and the effects of human intervention. In geomorphology, for example, many scientists of the 19th and early 20th centuries were content to describe landforms, devising elaborate systems of morphological classification, and only later did interest develop in understanding how landforms came to be, and the processes that left such characteristic footprints on the surface. Today, of course, such studies of form are largely discredited, as they are in many other disciplines.

Because of its essentially static legacy, much GIS analysis has focused on form, and has been criticized for doing so. It is comparatively difficult to tease insights into process from cross-sectional form, though it is perhaps sometimes possible to eliminate false hypotheses about process. GIS has been accused of being the last manifestation of the quantitative revolution that occurred in geography in the 1960s, when Bunge (1966) and others attempted to draw insights from

the similarity of forms found on the human and physical landscapes (see, for example, the critique of Taylor, 1990).

Very little is known, however, about the characteristic forms that may exist in spatio-temporal phenomena. Hagerstrand (1970) and others have examined the movements of individuals in space and time using three-dimensional displays, in which the two spatial dimensions form the horizontal plane and time forms the vertical axis. Much of this work focuses on similarities that may exist in the forms of such tracks, and the implications they may have for process. We know from the work of many researchers (e.g., Janelle and Goodchild, 1983) that different social conditions lead to dramatically distinct track forms, as for example in the differences between the daily tracks of single mothers, with their orientations to both workplace and daycare, and the tracks of workers in families in which only one of two adults works.

The development of greater support for time in GIS may lead to many other recognizable patterns in spatio-temporal data, and to a rebirth of interest in the study of spatio-temporal form. A new generation of analytic techniques is needed that extracts meaningful pattern from the mass of tracks displayed in the visualizations of Kwan and Lee (2004) and others, and links such patterns to hypotheses about process.

24.4. SPATIAL LITERACY

In the past few years a remarkable series of Web sites have brought the sophisticated functions of GIS and spatial analysis much closer to the general public. While effective use of GIS requires extensive training, and in many cases advanced work at the undergraduate level, technologies such as Google Earth have given every citizen with a computer and a high-speed Internet connection access to many of the data sets and computational functions of GIS, and in some cases have even exposed the more sophisticated functions of spatial analysis. For example, anyone requesting driving directions from one of these sites receives answers that result from the execution of a complex algorithm that was previously the reserve of operations researchers and specialists in spatial optimization.

The methods of cartography and related disciplines are complex, and it is no surprise therefore that sophisticated tools in naïve hands can produce mistakes. A suitable example concerns the Greenwich meridian, and its position when displayed in Google Earth. Many users of this site have noted that the zero of longitude misses the Greenwich Observatory by approximately 100 m, and have posted comments, some of which conclude that a serious mistake has been made by Google, and by extension that the georegistration of imagery on the site is poor. In reality, the WGS84 (World Geodetic System of 1984) datum, now widely adopted around the world, does not place the Greenwich Observatory at exactly zero longitude, despite the international treaty that established it there in 1884 – and the position shown in Google Earth appears to be correct to within a few meters.

Although their support for spatial analysis is extremely limited, these sites have clearly provided the general public with access to a rich resource, and thousands of people have been empowered to create their own applications. The recent publication *Mapping Hacks* (Erle, *et al*., 2005) describes many fascinating examples, but contains not a single reference to the cartographic literature. At the same time students who have endured many hours of lectures and lab exercises to become competent in GIS may be frustrated to realize that a child of ten

can create a computationally complex fly-by using Google Earth in a few minutes.

It seems clear that in part as a result of these developments the demand for basic knowledge of the principles of spatial analysis, GIS, geography, cartography, and related fields – for basic *spatial literacy* – is perhaps two or more orders of magnitude out of alignment with the supply. Education in these topics cannot be confined to a few advanced undergraduates, and to campuses lucky enough to have faculty interest, if it is to be accessible to the numbers of people now exposed to and enthusiastically adopting these tools. In this respect, spatial analysis faces an unprecedented challenge, to make itself known to a much larger community than previously.

There are several ways in which such a challenge might be met, by concerted effort on the part of the spatial-analysis community. One is to bring spatial literacy into the general-education or core curriculum of institutions of higher education, making its material accessible and eligible for credit for the vast majority of undergraduates. Courses in other kinds of literacy are already available in this form; the argument needs to be made that familiarity with spatial analysis and GIS represents another, and arguably a more powerful form of literacy that should be part of the education of every citizen. Another strategy would be to develop a larger and more visible set of courses in the informal education sector, making spatial literacy part of on-line and certificate programs, and exposing its contents through libraries, museums, and other institutions. A third is to work to introduce spatial literacy earlier in the educational hierarchy, in high school and even elementary school. Valiant efforts have been made in this direction in the past, but they remain minimal in comparison with the size of the primary and secondary sectors, and there is much confusion about where such material might fit in the already stove-piped curriculum.

24.5. BEYOND TRADITIONAL PRACTICE IN SCIENCE

When Harvey wrote his well-known and highly influential *Explanation in Geography* (Harvey, 1969) the dominant form of scientific practice centered on the individual investigator, whose methods followed a set of well-defined principles. For example, every experiment was to be reported in sufficient detail to allow its replication by another independent investigator. Every numerical result was to be reported with a level of precision that matched its accuracy. Every search of the literature was to be complete and comprehensive, so that the investigator could demonstrate knowledge of all previous and relevant work and prove the new work's originality. The principle of Occam's Razor – a willingness to adopt the simplest of several competing explanations – was universally accepted, as was the notion that all conclusions could be subject to empirical test and possible rejection. The goal of science was complete explanation, or in statistical terms an R^2 of 1. When sample data were analyzed, all numerical results were to be subject to tests of statistical significance, to prove that they were not likely to be simply artifacts of the particular sample chosen, but properties of the population from which the sample was presumed to be drawn. All terms were to be rigorously defined, and vague terms were to be replaced by ones that met the standard of objectivity – rigorous and shared definition, such that two investigators would always agree on the outcome when the definition was applied.

These standards are of course collectively unattainable in all circumstances. They may

be more attainable in some disciplines than others, and certainly it is possible to imagine a physicist having no difficulty adhering to them, and being fiercely critical of any study that appeared to relax them. But researchers in the general domain of this book clearly encounter situations in which one or more of them is distinctly problematic. This is not to say that one should therefore reject them outright, and follow the lead of those who have looked for alternatives to scientific principles – rather, they constitute goals to which research should attempt always to aspire, while admitting that it may sometimes fall short. This section explores three of these issues in some detail, and then argues for a renewed approach to scientific methodology that better reflects the real conditions under which spatial analysts currently work.

24.5.1. Collaboration, replicability, and the black box

Before the widespread adoption of computing, it was customary for instructors in statistics courses to insist that each student be able to carry out a test by hand, before using any computational aids. Only then, it was argued, would the student fully understand the process involved, and be able to replicate it later. In this simple world it was possible to assume that every researcher knew every detail of every analysis, and that the published version of the research would include sufficient detail to allow others to repeat the experiment and replicate the results.

This principle has come under fire in recent decades, for a number of reasons. Computational aids have advanced to the point where it is not possible for any one individual to comprehend fully all of the algorithms involved. The author recalls passing a threshold, some time around 1990, when it was no longer possible to believe that every aspect of a computational analysis could be replicated by hand, given enough time. Operating systems were perhaps the first such area of computing – by 1990 they had advanced to the point where it was no longer possible to believe they were the work of one person, or that any one individual fully understood every aspect of their operation. Today these failures are commonplace. The documentation of our more sophisticated software, including GIS, is often not sufficient to detail every aspect of an analysis, and it may be impossible to discover exactly how a given system computes a standard property, such as 'slope', from a given input (Burrough and McDonnell (1998) detail some of the options, but many more can be hidden in the details of a given implementation). In effect the developers of software, many of them operating in for-profit commercial environments, have become authorities that must be trusted, and it is difficult to submit their products to rigorous and exhaustive test.

Moreover, researchers now find it increasingly effective to work in teams, each team member providing some specific expertise. Funding agencies often express a willingness to fund research that brings together teams from many disciplines, in the interests of greater collaboration and cross-fertilization of ideas. But such arrangements inevitably lead to situations in which no one individual knows everything about an analysis, and members of the team have little alternative but to trust each other, just as researchers often have little alternative but to trust software.

24.5.2. Keeping the stakeholders happy

Tools such as GIS invite researchers to become involved in the processes of policy formulation and decision making. The very

architecture of GIS, with its database of local details and its procedures representing general principles, invites engagement with the ultimate users of research, since it allows decision makers to investigate the effects of manipulating outcomes in local contexts, and gives them many useful tools for implementing the results of analysis. A new subdiscipline, public-participation GIS, has grown up to study these issues, and to improve the use of GIS and spatial analysis in public decision making.

Many of the arguments for the use of technology in support of decision making – for spatial decision support systems (Densham, 1991) – center on the benefits of these tools in settings that involve the potentially conflicting views of multiple stakeholders. Much has been written about spatial-analytic techniques that support multiple views, and address multiple criteria (Voogd, 1983; Eastman, 1999; Thill, 1999; Malczewski, 1999). GIS may allow stakeholders to express their own views as sets of weights to be given to relevant factors. Saaty's Analytic Hierarchy Process (Saaty, 1980) is a widely used technique for eliciting such weights from stakeholders, and for deriving consensus weights and measures of agreement. Stakeholders benefit from the visualization capabilities of these systems, which allow them to see the effects of decisions in readily understood ways. They gain the impression that decisions are made *scientifically*, with abundant use of mathematics and computation, and are led to believe that these approaches represent a more objective, more desirable approach to debate and conflict resolution.

It is all too easy in such circumstances to see stakeholder satisfaction as the primary goal of the exercise. If stakeholders leave the room believing that a rigorous, scientific process has been conducted then everyone can feel that a useful exercise has come to an acceptable conclusion. None of this guarantees, however, that the results presented to the stakeholders are in fact based on good science. It is easy, with a little thought, to manipulate the outcomes of such processes to achieve hidden objectives. For example, when stakeholders are presented with five alternatives and asked to choose one, it is easy to see how the outcome might be manipulated by presenting a set that includes the desired outcome, plus four obviously unacceptable 'red herrings.' Experience suggests that stakeholders will find no difficulty in assigning relative measures of 'importance' to factors, irrespective of whether the factors are or are not commensurate, and whether or not any definition of 'importance' has been advanced and agreed.

24.5.3. Accuracy, uncertainty, and cost

All measurements are subject to error, and science has developed sophisticated techniques for measuring instrument accuracy, and for determining how accuracy impacts the results of analysis. The basic principles of error analysis have been adapted to the specific needs of geographic data by Heuvelink (1998) and others, and statistical models have been developed for most of the standard geographic data types.

Uncertainty is often defined as the degree to which data leaves the user uncertain about the true nature of the real world. As such it presents a greater problem, because it derives not from errors in measurement, but from vagueness in definitions, lack of detail, and numerous other sources. When definitions are vague, there can be no objective definition of truth, but only the less satisfactory concept of consensus. A scientist steeped in traditional methodology would react by rejecting vague terms entirely, replacing them with terms that have rigorous definition, and are

therefore capable of supporting replicability. Subjective terms such as 'warm,' 'cold,' 'near,' and 'far' would be replaced by well-defined scales of temperature measurement and distance.

Nevertheless, GIS and to a lesser extent spatial analysis clearly exist at the interface between the rigorous, scientific world of well-defined terms and replicable experiments, and the vague, intuitive world of human discourse. Many users of GIS appear happy to work with vaguely defined classes of vegetation or land use, and there has been much interest in building user interfaces to GIS that come closer to emulating human ways of reasoning and discovering. *Naïve geography* has been defined as a field that studies the simplifications humans often impose on the world around them, and writers have speculated about the potential for systems that also simplify – that 'think more like humans do.'

In the past decade or so there has been much interest in the application of fuzzy sets, rough sets, and related ideas in spatial analysis. There seems to be some degree of intuitive appeal in the idea of assigning degrees of membership to a class, even when the class is not itself well defined. Methods have been devised for eliciting fuzzy membership values from professionals, from remotely sensed data, and from other sources, and for displaying these values in the form of maps. All of these methods stretch the norms of science, by arguing that it is possible to observe and measure useful properties despite a lack of agreement on the definitions of those properties. As such, they demand a re-examination of the basic tenets of scientific method.

Finally, spatial analysts find themselves today in a world overflowing with data. Satellite images, digital topographic maps, and a host of other sources provide an unprecedented opportunity for new and interesting research. Massive investments have been made over the past decade in data warehouses, spatial data centers, and geo-portals, with a view to facilitating the discovery and sharing of spatial data. Metadata standards have been devised that support search, by allowing researchers to hunt through catalogs looking for data that might meet their needs.

Yet almost certainly data discovered in this way will fail to meet the exact needs of the researcher. The data set will be too generalized, not sufficiently current, too inaccurate, or inadequate in another of a myriad of possible ways. In these circumstances it is inevitable that research objectives become modified to fit the properties of the available data, if the alternative is an exercise in field data collection that may be impossibly expensive. But the prevailing methodology of science says nothing about such compromises, maintaining instead that data must be exactly fit for purpose, and providing no basis on which users can find compromises between cost on the one hand, and accuracy or fitness for use on the other.

24.5.4. Summary

The previous three sections have presented examples of the ways in which spatial analysts increasingly find the traditional principles of scientific methodology inadequate as a guide to practice. While much of science is concerned with the nomothetic goal of discovering general principles that apply everywhere in space and time, spatial analysis is increasingly concerned also with the variations that exist in such principles from place to place, and in the ways in which such principles are placed in local context to solve problems and make decisions. As Laudan (1996) has argued, there is no longer an effective methodological distinction between science and

problem-solving, since the same principles apply to both. In summary, spatial analysts face an important challenge, to develop a new methodological understanding that is consistent both with the traditional tenets of the scientific method, and with the realities of current practice.

24.6. CONCLUSIONS

The four major sections of this chapter have argued that spatial analysis faces many challenges at this time, but it also faces unprecedented opportunity. More people than ever are aware of its potential, and the tools to implement it are more sophisticated and powerful than ever.

Discussions of the importance of spatial analysis often focus on one or two particularly compelling application domains, and it may well be that by making the case for spatial analysis in support of improved public health, for example, or better response to emergencies, it will be possible at the same time to promote the entire field. On the other hand, one might argue that identifying spatial analysis too clearly with one application domain tends to render the case for other applications more difficult. Essentially, it can be very difficult to promote a set of techniques that are applicable to almost *everything* – the case for spatial analysis is everywhere, and yet at the same time it is nowhere.

The argument for spatial literacy made in section 24.4 seems especially relevant in this context. Many skill areas are important across a vast array of human activities, including skill in language, in mathematics, and in logic. Spatial analysis should not be a highly specialized area of technique that is only accessible to experts, but should be part of every citizen's basic set of skills, and used every day in such basic activities as wayfinding and activity planning.

How the field responds to these challenges remains to be seen, of course. Undoubtedly new and better techniques will be discovered and published in the next few years, new code will be written, and new application areas will be described. But the challenges described in this chapter seem to go beyond such business-as-usual, and to require discussion across the entire community. Such community-wide debate has occurred very rarely in the past, yet is more feasible than ever with today's communications technologies.

ACKNOWLEDGMENTS

Support of the U.S. National Science Foundation through award BCS 0417131 is gratefully acknowledged. The author also benefited from an E.T.S. Walton Fellowship which allowed him to spend much of the 2005–6 academic year at the National Centre for Geocomputation, National University of Ireland, Maynooth.

NOTE

1 National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA. Phone +1 805 893 8049, Fax +1 805 893 3146, E-mail good@geog.ucsb.edu

REFERENCES

Abler, R.F. (1987). The National Science Foundation National Center for Geographic Information and Analysis. *International Journal of Geographical Information Systems*, **1**: 303–326.

Anselin, L. (1989). *What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis*. Technical Report 89–4. Santa Barbara, CA: National Center for Geographic Information and Analysis.

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis*, **27**: 93–115.

Anselin, L. and Getis, A. (1992). Spatial statistical analysis and geographic information systems. *Annals of Regional Science*, **26**: 19–33.

Bunge, W. (1966). *Theoretical Geography*. 2nd edn. Lund Studies in Geography Series C: General and Mathematical Geography, No. 1. Lund, Sweden: Gleerup.

Burrough, P.A. (1990). Methods of spatial analysis and GIS. *International Journal of Geographical Information Systems*, **4**: 221–223.

Burrough, P.A. and McDonnell, R.A. (1998). *Principles of Geographical Information Systems*. New York: Oxford University Press.

Densham, P.J. (1991). Spatial decision support systems. In: Maguire, D.J., Goodchild, M.F. and Rhind, D.W. (eds), *Geographical Information Systems: Principles and Applications*. pp. 403–412. Harlow, UK: Longman Scientific and Technical.

Ding, Y. and Fotheringham, A.S. (1992). The integration of spatial analysis and GIS. *Computers in Environmental and Urban Systems*, **16**: 3–19.

Eastman, J.R. (1999). Multi-criteria evaluation and GIS. In: Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds), *Geographical Information Systems: Principles, Techniques, Management and Applications*. pp. 225–234. New York: Wiley.

Erle, S., Gibson, R. and Walsh, J. (2005). *Mapping Hacks: Tips and Tools for Electronic Cartography*. Sebastopol, CA: O'Reilly Media.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Hoboken, NJ: Wiley.

Fotheringham, A.S. and P. Rogerson, (eds), (1994). *Spatial Analysis and GIS*. London: Taylor and Francis.

Getis, A. and Ord, J.K. (1992). The analysis of spatial association by distance statistics. *Geographical Analysis*, **24**: 189–206.

Goodchild, M.F. (1987). A spatial analytical perspective on geographical information systems. *International Journal of Geographical Information Systems*, **1**: 327–334.

Goodchild, M.F. (2003). The fundamental laws of GIScience. Paper presented at the Summer Assembly of the University Consortium for Geographic Information Science, Pacific Grove, CA, June. Available: http://www.csiss.org/aboutus/presentations/files/goodchild_ucgis_jun03.pdf

Goodchild, M.F., Anselin, L. and Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, **25**: 383–397.

Goodchild, M.F., Fu, P. and Rich, P. (in press). Sharing geographic information: an assessment of the geospatial one-stop. *Annals of the Association of American Geographers*.

Goodchild, M.F., Haining, R.P. and Wise, S. (1992). Integrating GIS and spatial analysis: problems and possibilities. *International Journal of Geographical Information Systems* **6**: 407–423.

Goodchild, M.F. and Longley, P.A. (1999). The future of GIS and spatial analysis. In: Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds), *Geographical Information Systems: Principles, Techniques, Management and Applications*. pp. 235–248. New York: Wiley.

Hägerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association*, **24**: 7–21.

Harvey, D. (1969). *Explanation in Geography*. New York: St Martin's Press.

Heuvelink, G.B.M. (1998). *Error Propagation in Environmental Modelling with GIS*. Bristol, PA: Taylor and Francis.

Janelle, D.G. and Goodchild, M.F. (1983). Transportation indicators of space-time autonomy. *Urban Geography*, **4**: 317–337.

Kwan, M.-P. and Lee, J. (2004). Geovisualization of human activity patterns using 3D GIS: A time-geographic approach. In: Goodchild, M.F. and Janelle, D.G. (eds), *Spatially Integrated Social Science,* pp. 48–66. New York: Oxford University Press.

Langran, G. (1993). *Time in Geographic Information Systems*. London: Taylor and Francis.

Laudan, L. (1996). *Beyond Positivism and Relativism: Theory, Method, and Evidence*. Boulder, CO: Westview Press.

Maguire, D.J. and Longley, P.A. (2005). The emergence of geoportals and their role in spatial data infrastructures. *Computers, Environment and Urban Systems*, **29**(1): 3–14.

Malczewski, J. (1999). *GIS and Multicriteria Decision Analysis*. New York: Wiley.

National Research Council (2006). *Learning to Think Spatially: GIS as a Support System in the K-12 Curriculum*. Washington, DC: National Academies Press.

Openshaw, S. (1990). Spatial analysis and geographical information systems: a review of progress and possibilities. In: Scholten, H.J. and Stillwell, J.C.H. (eds), *Geographical Information Systems for Urban and Regional Planning*. pp. 153–163. Dordrecht: Kluwer.

Peuquet, D.J. (1999). Time in GIS and geographical databases. In: Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds), *Geographical Information Systems: Principles, Techniques, Management and Applications*. New York: Wiley.

Peuquet, D.J. (2001). Making space for time: issues in space–time representation. *Geoinformatica*, **5**(1): 11–32.

Peuquet, D.J. (2002). *Representations of Space and Time*. New York: Guilford.

Saaty, T.L. (1980). *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. New York: McGraw-Hill.

Sui, D.Z. and Goodchild, M.F. (2001). Guest Editorial: GIS as media? *International Journal of Geographical Information Science,* **15**(5): 387–389.

Sui, D.Z. (ed.), (2004). Forum: on Tobler's First Law of Geography. *Annals of the Association of American Geographers,* **94**(2): 269–310.

Takeyama, M. and Couclelis, H. (1997). Map dynamics: integrating cellular automata and GIS through Geo-Algebra. *International Journal of Geographical Information Science*, **11**(1): 73–91.

Taylor, P.J. (1990). GKS. *Political Geography Quarterly,* **9**(3): 211–212.

Thill, J.-C. (1999). *Spatial Multicriteria Decision Making and Analysis: A Geographic Information Sciences Approach*. Brookfield, VT: Ashgate.

Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography,* **46**: 234–240.

Tomlin, C.D. (1990). *Geographic Information Systems and Cartographic Modeling*. Englewood Cliffs, NJ: Prentice Hall.

Ungerer, M.J. and Goodchild, M.F. (2002). Integrating spatial data analysis and GIS: a new implementation using the Component Object Model (COM). *International Journal of Geographical Information Science,* **16**(1): 41–54.

van Deursen, W.P.A. (1995). *Geographical Information Systems and Dynamic Models: Development and Application of a Prototype Spatial Modelling Language*. Nederlandse Geografische Studies 190. Utrecht: Koninklijk Nederlands Aardrijkskundig Genntschap/Faculteit Ruimtelijke Wetenschappen Universiteit Utrecht.

Voogd, H. (1983). *Multi-Criteria Evaluation for Urban and Regional Planning*. London: Pion.

25

The Future for Spatial Analysis

Reginald G. Golledge

25.1. SPATIAL ANALYSIS PAST AND PRESENT

The future of geography is inextricably bound to the future of spatial analysis. Why? Simply because spatial analysis captures the essence of a support system for the science and technology involved in geospatial thinking and reasoning. The latter are the distinct and unique contributions of geography to the universe of academe, government, and business.

For about 50 years, geographers have been slowly but surely building a structure of theories, models, methods, technologies, and vocabulary that anchor the discipline's claim to being a science. This effort has occurred both in the physical and human components of the discipline. A common theme in both efforts has been the search for valid and reliable conclusions from active and innovative research. A variety of exploratory and confirmatory, qualitative and quantitative procedures have been developed or explored for relevance, and relevant procedures and methodologies have been globally termed 'spatial analysis.' While some parts of the discipline are content to imitate the theories, methods, and technologies of other physical or human sciences, or to copy the research designs and practices of the various humanities, parts of geography have vigorously explored the development of unique means of thinking, reasoning, analyzing, and representing geospatial information. Spatial analysis has been perhaps the most vigorous of these throughout the years. Lately, it has been complemented by the enthusiasm for technology – particularly Geographic

Information Systems (GIS). However, most academic practitioners realized quickly that GIS needed a wider base: a base of analysis as well as its forte in representation. To provide this base, Geographic Information Science (GISc) developed. Spatial analysis proved to be a primary support system for GISc, and the two themes have converged to give geographic researchers and teachers powerful new ways to explore the massive data banks of the new digital world.

Many geographers would not agree with my opening statement. I would challenge them to disprove it or to make valid claims for other dimensions of the discipline. One could not support a contrary argument based on geography's traditional role of collecting facts about the earth's physical or human environment. While other aspects of the discipline continue to have much to offer in terms of understanding the relations between people and places, it is not always possible to differentiate the geographic/geospatial component from the more general humanities', political sciences', or social sciences' thinking and reasoning that drives much of this work. Thus, it has the potential to contribute to the accumulation of general social and cultural knowledge more than to geospatial knowledge. This can be viewed as a positive result if one accepts that integrated disciplinary thinking is likely to be of future importance, but does little to support or enhance the image and practice of geography in the real world.

So, why does Spatial analysis hold the key (in my opinion) to the future of geography? To reflect on this, I offer the following thoughts (see also Goodchild, 2001):

- Spatial analysis is a unique and special contribution by geographers to the ongoing trend of integrated science. Here, 'science' is interpreted in both a qualitative and quantitative manner, and covers both physical science ('natural' science) and human science (the science involved in comprehending human–human and human–environment relations). It provides a menu for ensuring valid and reliable reasoning in the forum of knowledge accumulation.

- Spatially referenced data – either in relative (qualitative) or absolute (quantitative) form – has become the currency of today's information processing society. Spatial analysis is exclusively developed for analyzing place-based digital information. It includes the use of topologies, geometry, fuzzy logic, and multidimensional reasoning capabilities, all directed towards the spatial domain. Thus, it is useful at all scales from the nano and micro levels to the gigantic scale of universe-wide exploration, and is being diffused through areas as different as neurological experimentation, archeological reconstructions of past civilizations, and the search for extraterrestrial understanding.

- It is generally agreed that geographers have a unique way of examining problems (Beck, 1967; Uttal, 2000), and that diagrammatic (including map-based) reasoning provides insight into many problems that is unattainable using conventional reasoning such as verbal, text-based, and mathematical procedures. This uniqueness begins with the accepted significance of the spatial domain (something that has been rather neglected by other disciplines and by many parts of the human side of geography), and then expresses itself via its emphasis on visualization and spatialization processes. Data is collected with some form of spatial coding (Klatzky, *et al.*, 1990; Fujita, *et al.*, 1993), and is represented by the spatializations such as flat (2D) paper maps, 3D models, and on-screen image-based representation (graphs and graphics), all of which require a particular form of interpretation. Faithful representation is one of the prerequisites for spatial analysis.

- During the second half of the 20th century, geography matured by borrowing (sometimes wholesale, sometimes modified) theories from other disciplines. As the profession gained more

confidence in its ability to offer innovative, exploratory and confirmatory investigations of spatial and geospatial concerns, there finally emerged a series of spatially explicit theories of the relations that were being uncovered by research in the spatial (and specifically geographical) domains. These theories tended to be investigated and validated using spatial analysis. They included time–space associations, spatial decision making, spatial choice, location theory, location–allocation processes, population density gradients, the form and structure of built environments, geospatial learning, movement behavior at different scales, and other areas that are explicitly spatial (see earlier chapters). And, as the profession learned to think and reason spatially (rather than socially, politically, or economically), the processes involved in spatial analysis continued to grow in importance.

- The majority of geographers (not just those engaging in spatial analysis) use place-based reasoning. In many cases, this provides the only link to the spatial domain in that it spatializes non-spatial phenomena such as social class, political ideology, and financial perspectives. Often, the tie to place is loose and general but still provides the wherewithal to discuss place-to-place differences. But the latter is frequently incidental to the reasoning process and is used largely for illustrative or representational purposes. But one of the strengths of spatial analysis is its explicit focus on place-to-place variation across all scales of investigation; i.e., a principal purpose of spatial analysis is to record and to help explain the existence of such differences and why they occur. In this way, spatial analysis provides a support system that makes spatial thinking paramount, and not incidental.

- Spatial analysis procedures have become part and parcel of GISc software. Part of GISc has been tied to understanding what spatial analysis can do to clarify and validate spatial thinking across all scales of investigation. The interdependence of GISc and spatial analysis has been forged. As the use of GIS has expanded through academe, government, and business, many disciplines have laid claim to being the principal originator and purveyor of GIS technology. But none have been able to dispute geography's claims to the special confluence of GISc's search for relevant spatial theory, its representational capability, and the many procedures of spatial analysis that add meaning and usefulness, validity and reliability to GISc's problem solving activities. The integration of GIS and spatial analysis has been influential in moving GISc-related research beyond mere technology to scientific status. Via this link, spatial analysis has been forming the basis for new theories that incorporate human–environment relations, e.g., spatial knowledge acquisition (Golledge, 1978; Montello, 1998) and new theories of data and data manipulation (Goodchild, 2004; Couclelis, 2003).

25.2. THE ROLE OF SPATIAL ANALYSIS

In the process of re-establishing itself as a viable academic discipline (i.e., after its role in examining 'what' was 'where' on the earth's surface, and pursuing the description of the results of human–environment interactions, was made somewhat redundant by remotely sensed image processing procedures), geographers have had to justify their continual existence or go out of business. Some leading departments such as Chicago and Michigan have, in effect, 'gone out of business,' while many others have been merged with geology, geological science, environmental science, sociology, urban planning/architecture/design, or other groups. Despite these dire warnings, much of the discipline has gone its pedestrian way, virtually ignoring the global change from a partly known and image-based world to a group of information societies and a digital world.

But these later trends have provided a rationale and need for specific spatially-based means of examining, processing, and representing the data that is becoming increasingly available in digital form. The need for such procedures is not confined to geography. Other social, behavioral, political, economic, and health sciences, for example, have discovered that their data banks are being spatialized by geocoding of occurrences and attributes, and that traditional measures of statistical analysis do not account for the effects of spatial coincidence or variation. Hence, the demand for spatial analysis is growing in these disciplinary areas. I predict it will continue to grow. It is the goal of every spatial theorist to see various methods for spatial analysis of data incorporated into every standard statistical package, thus imprinting this contribution by geographers on the domain of every spatially oriented discipline. One recent example of this recognition is the inclusion of a chapter on GIS in a recent *Handbook of Environmental Psychology* (Bechtel and Churchman, 2002) and a decision by the American Psychological Association (APA) to support an advanced institute on GIS and spatial analysis (probably in 2007).

25.3. NEW DIRECTIONS FOR SPATIAL ANALYSIS

The interweaving of GISc and spatial analysis has given to geography a justifiable scientific base that, for most of geography's history, has been lacking. This new basis has:

- increased the public and academic image of geography as a serious scientific discipline;
- improved the standing and reputation of geography as a useful contributor to the examination and solution of problems such as comprehending global climate change or understanding human spatial abilities;
- made geographic training and expertise a valuable commodity in the job market;
- brought the realization that, as globalization of societies and their essential activities occur, geographers have a unique contribution to make in the form of geo-education, spatial concept recognition, and spatial thinking and reasoning;
- encouraged exploring the possibility of enhancing geography in the K-12 system of education.

I anticipate that each of these contributions will become more important in both the near and distant futures.

To speculate about the 'what' and 'where' of spatial analysis' contribution to the future of geography, consider the following:

- Recognition that spatial analysis applies and can be used at all scales – from the nano scale to the universal. We already have evidence that researchers in microbiology, neurology, DNA, and stem cell research (as well as other research areas not traditionally identified with geography) are facing questions concerning representation and analysis of their spatially-based findings. Both GISc and spatial analysis potentially have an important contribution to make in these areas (e.g., via spatialization, representation, and analysis).
- One of the most important frontiers for future research is to investigate how the mind works. Great advances already have been made in discovering how the brain works. Indeed, one of the most intriguing investigations – from a geographer's viewpoint – is the extent to which 'place cells' exist (O'Keefe and Nadel, 1978) and form a basis for internal data manipulations that constitutes the mind's contribution to solving spatial problems. The question arises then, if place cells do exist, what light is shed on how data is sensed and coded and stored in

the brain? What happens when we start to think spatially? Is there a particular pattern of neural excitation when we think spatially? Can spatial analysis help both to investigate this and add a newly emerging area for geospatial investigation?

- The world is digitizing. We already have more data from satellites than can conceivably be analyzed in the present or the near future. The question arises as to whether the existing form of spatial analysis can contribute to performing data mining and, as necessary, add new and valuable components to existing search engines. A question for the future may be: are there yet other levels of spatial analysis we have not yet thought about but which could be an essential part of recovering the spatial relations contained in these massive archival structures?

As disciplines such as psychology and cognitive science experiment more in the real world (in addition to ongoing research in laboratories and virtual systems), and as the importance of scale effects and the significant role of place-to-place variation in forming attitudes and behavior is realized, so too has the demand for spatial analysis started to emerge. There is much room for geographers to both teach about and disseminate spatial analysis procedures within and beyond the profession of geography. For decades, we have been borrowing from measurement theory from math's symbolic thinking strategies, from mathematical models developed in economics, and analytic procedures from psychology and mathematical statistics; it is time to return this favor by encouraging the use of spatial analytic techniques for processing relevant geospatial data and drawing attention to the very specific contributions of space in the construction of knowledge. At the very least, psychologists and cognitive scientists should become aware of both the advantages and disadvantages of spatializing data for graphic, map, image-based, or symbolic representation. For example, a glance at the psychology literature on spatial perception and cognition reveals little comprehension of the role space plays in information gathering and information processing in the large uncontrolled spaces of the real, inhabited world, and various graphic and image-based representations of this world.

There also appears to be a growing demand for applied geography, particularly in government and business domains. We have already seen such a demand within the business community – as with the use of location–allocation models and use of location-based services. Spatial analysis is a key to expanding this demand. The result should be a more widespread acceptance of the contributions that geography can make to everyday life and practice throughout local and global societies.

In my opinion, therefore, spatial analysis, perhaps in conjunction with the use of GIS technology and a GISc search for reliable and valid bases for knowledge accumulation, will provide an avenue for maintaining and expanding the image and acceptance of geography as an integrated science that has a positive capacity to assist the search for new knowledge, and improve our general quality of life.

As a final statement, allow me to raise a question that is critical to the future of geography itself. Are we producing graduates who can compete for jobs in academic, government, or business marketplaces? Sadly, the answer for most of the profession is NO! But spatial analysts and GIS programs *are* doing this, very successfully. To return to my opening statement: the future of geography as a viable discipline is inextricably tied to the continued development and use of spatial analysis. We've already seen the first indicators of this in terms of which students are getting jobs today outside of academe.

As a discipline, we must become more aware of this need and do our best to ensure that those areas contributing most to this pattern are well supported in the near and more distant future.

ACKNOWLEDGMENTS

The research for this chapter was partly supported by NSF Grant # BCS0239883 ('Spatial Thinking') and by UCTC Grant # SA4655 ('Assessing Route Accessibility for Wheelchair Users').

REFERENCES

Beck, R. (1967). Spatial meaning and the properties of the environment. In: D. Lowenthal (ed.), *Environmental Perception and Behavior* (Research Paper No. 109, pp. 18–29). Chicago: Department of Geography, University of Chicago.

Bechtel, R.B., and Churchman, A. (eds). (2002). *Handbook of Environmental Psychology*. New York: John Wiley & Sons.

Couclelis, H. (2003). The certainty of uncertainty. *Transactions in GIS*, **7**(2): 165–175.

Fujita, N., Klatzky, R.L., Loomis, J.M. and Golledge, R.G. (1993). The encoding-error model of pathway completion without vision. *Geographical Analysis*, **25**(4): 295–314.

Golledge, R.G. (1978). Learning about urban environments. In: Carlstein, T., Parkes, D. and Thrift, N. (eds), *Timing Space and Spacing Time, Volume I: Making Sense of Time*, pp. 76–98. London: Edward Arnold.

Goodchild, M.F. (2001). A geographer looks at spatial information theory. In: Montello, D.R. (ed.), *Spatial Information Theory: Foundations of Geographic Information Science. Proceedings, International Conference, COSIT 2001, Morro Bay, CA, September*, pp. 1–13. New York: Springer.

Goodchild, M.F. (2004). GIScience: geography, form, and process. *Annals of the Association of American Geographers*, **94**(4): 709–714.

Klatzky, R.L., Loomis, J.M., Golledge, R.G., Cicinelli, J.G., Doherty, S. and Pellegrino, J.W. (1990). Acquisition of route and survey knowledge in the absence of vision. *Journal of Motor Behavior*, **22**(1): 19–43.

Montello, D.R. (1998). A new framework for understanding the acquisition of spatial knowledge in large-scale environments. In: M.J. Egenhofer and Golledge, R.G. (eds), *Spatial and Temporal Reasoning in Geographic Information Systems*, pp. 143–154. New York: Oxford University Press.

O'Keefe, J. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.

Uttal, D.H. (2000). Seeing the big picture: Map use and the development of spatial cognition. *Developmental Science*, **3**: 247–264.

Index

Page numbers in italics indicate figures and tables