Spatiotemporal Clustering : A Review

Mohd Yousuf Ansari[1] · Amir Ahmad[2] · Shehroz S. Khan[3] · Gopal Bhushan[1] · Mainuddin[4]

Abstract

An increase in the size of data repositories of spatiotemporal data has opened up new challenges in the fields of spatiotemporal data analysis and data mining. Foremost among them is “Spatiotemporal Clustering,” a subfield of data mining that is increasingly becoming popular because of its applications in wide-ranging areas such as engineering, surveillance, transportation, environmental and seismology studies, and mobile data analysis. This review paper presents a comprehensive review of spatiotemporal clustering approaches and their applications as well as a brief tutorial on the taxonomy of data types in the spatiotemporal domain and patterns. Additionally, the data pre-processing techniques, access methods, cluster validation, space-time scan statistics, software tools, and datasets used by various spatiotemporal clustering algorithms are highlighted.

Keywords Data mining · spatiotemporal clustering · patterns · cluster validation

1. Introduction

Large volumes of spatiotemporal data are generated by various technologies including remote sensing, mobile networks, GPS devices and RFID systems. This massive amount of spatiotemporal data poses challenges in terms of storage, management, analysis and knowledge discovery (Li et al. 2004). Data mining is a very powerful technology that has potential to extract nontrivial, implicit, and previously unknown information from huge repositories of data (Larose 2005). Spatial Data Mining is a process of discovering trends or patterns from large spatial databases that hold geographical data (Manjula and Narsimha 2014). Spatiotemporal data mining refers to the extraction of implicit knowledge, spatial and temporal relationships, or similar patterns from spatiotemporal data (Yao 2003). Spatiotemporal data mining has many real world applications encompassing the fields of social, earth, and medical sciences (Chen et al.2015), the internet of things (Gubbi et al.2013), epidemiology (Shekhar et al.2008), and public safety (Leipnik and Albert 2002).

Mohd Yousuf Ansari[1]
[1] Defence Scientific Information & Documentation Centre (DESIDOC), Defence R&D Organisation (DRDO), Metcalfe House, Delhi-110054, India
e-mail: md_ya@yahoo.com
Amir Ahmad[2]
[2] College of Information Technology, United Arab Emirates University, Al Ain, UAE
e-mails: amirahmad@uaeu.ac.ae
Shehroz S. Khan[3]
[3]Toronto Rehabilitation Institute, University Health Network, Toronto, Canada
e-mail : shehroz.khan@utoronto.ca
Gopal Bhushan[1]
e-mail: gopalbhushan@gmail.com
Mainuddin[4]
[4]Department of Electronics & Communication, Faculty of Engineering and Technology,
Jamia Millia Islamia, New Delhi-110025,India
e-mails: mainuddin@jmi.ac.in

Clustering is the process of grouping large data sets based on a certain similarity measure. Cluster analysis is an unsupervised form of machine learning, because it does not require a priori knowledge of the data sets (Ahmad and Dey 2007). Spatial clustering is a type of clustering in which data values are usually in terms of longitude and latitude (Tork 2012). Spatiotemporal clustering is an extension of spatial clustering in which the time dimension is introduced into spatial data (Tork 2012; Birant and Kut 2007). In spatiotemporal clustering, the objects are grouped as per their spatial and temporal similarity (Kisilevieh et al.2010). The combination of geographic dimensions with time introduces several application-dependent challenges (Kisilevieh et al.2010). Spatiotemporal clustering plays a vital role in many areas of engineering, scientific and real-world applications such as image processing and pattern recognition (Li et al. 2004), molecular biology (Sander et al.1998), environmental studies (Birant and Kut 2007), seismology studies (Georgoulas et al. 2013; Wang et al.2006), transportation (Jeung et al. 2008; Spaccapietra et al. 2008), traffic management (Anbaroglu et al. 2015; Spaccapietra et al. 2008; Palma et al. 2008), Geographical Information Systems (Sander et al.1998), identification of terrorist activities (Kalyani and Chaturvedi 2012), surveillance (Vieira et al. 2009), mobility data analysis (Bogorny and Shashi 2010), trajectory outlier detection (Bogorny and Shashi 2010), animal behavior pattern identification (Kalnis et al. 2005; Vieira et al. 2009), vehicle convoy tracking (Kalnis et al. 2005), and automatic discovery of fishing spots (Rocha et al. 2010).

This review paper provides a comprehensive review of spatiotemporal clustering approaches. We present the taxonomy for spatiotemporal clustering based on the spatiotemporal domain, such as event clustering, geo-referenced data item clustering, moving clusters, trajectory clustering, and semantic-based trajectory data mining. Kisilevieh et al. (2010) presented a literature survey of spatiotemporal clustering research and Maciag (2017) reviewed clustering algorithms pertaining to complex spatial objects such as polygons and distance measures for spatial objects. However, the survey of Kisilevieh et al. (2010) only covers spatiotemporal clustering algorithms pertaining to the trajectory data type, has limited discussions of patterns, and does not mention the tools and datasets used by various spatiotemporal clustering approaches; meanwhile, that of Maciag (2017) does not cover patterns, tools, or datasets. Here the attempt has been made to fill these gaps, by discussing the topics mentioned above and rendering this review of use to the practitioner. In this review paper, we expand the scope of the literature review covered by Kisilevieh et al. (2010) and Maciag (2017) by including spatiotemporal clustering algorithms (Georgoulas et al. 2013; Lee 2012; Han et al. 2016; Agrawal et al. 2016; Izakian et al. 2015; Zaghlool et al. 2015; Chen et al. 2015; Rocha et al. 2010) that were not covered in these papers. Our literature review also includes various applications of spatiotemporal clustering approaches. Moreover we present a review of software tools used for spatiotemporal data analysis and data mining, highlight several publicly available data repositories to help accelerate research in this field and encourage comparison of competing algorithms.

Section 2 presents a brief tutorial on the data, its types, basic spatial relationships, and data pre-processing. Section 3 introduces the concept of spatiotemporal clustering algorithms of various categories of the spatiotemporal domain. Section 4 presents an overview of patterns along with the approaches adopted for the extraction of patterns in the spatiotemporal domain. Section 5 addresses certain aspects of cluster validation. Section 6 briefly discusses application-specific and generic spatiotemporal analysis tools, and free data mining tools available in the literature in the context of spatiotemporal clustering, the general characteristics of datasets used in various spatiotemporal clustering approaches, as well as potential applications. Section 7 provides an overview of applications of spatiotemporal clustering approaches in many areas of engineering, scientific, and real-world applications.

Section 8 highlights various issues and challenges in the context of spatiotemporal clustering and section 9 finally concludes the paper.

2. Introduction to Spatiotemporal Data

Spatial data are represented as a list of numbers using a particular coordinate system. The objects of an electronic map are represented using spatial data.

To describe geographic objects, we require spatial and non-spatial attributes as well as spatial relationships (Bogorny and Shashi 2010). The spatial attribute describes the spatial location and representation of the object, considering the geometry and a coordinate system, whereas the qualitative and quantitative properties of a geographic entity are described by the non-spatial attributes. There are three basic spatial relationships (Güting RH 1994): distance, direction (or order), and topological. The foundation of distance relationships lies on the Euclidean distance between two spatial features. Examples of direction relationships include above, below, north_of, southeast_of, etc. Topological relationships describe the type of intersection between two spatial features. The topological spatial relationships (Clementini et al. 1993) are disjoint, touches, overlaps, equals, inside, contains, and crosses.

Spatiotemporal data incorporates temporal information in addition to spatial information, which makes them more complicated (Tork 2012). Spatiotemporal data generally record an object state, an event, or a position in space, over a period of time. Spatiotemporal data could be classified into five types: events, geo-referenced data items, geo-referenced time series, moving objects, and trajectories (Kisilevieh et al.2010; Tork 2012; Kalyani and Chaturvedi 2012; Giannotti et al. 2008; Shekhar et al. 2015; Zhang et al.2003; Izakian et al.2013; Achtert et al. 2008; Rinzivillo et al. 2008; Spaccapietra et al. 2008).

An event is a triplet <longitude, latitude, timestamp> that has an association with the recording location and time (Kisilevieh et al.2010). A typical example of events is seismic activity monitored by sensors or geo-referenced records of an epidemic (Kisilevieh et al.2010). Events usually do not have correlation between data items and also have no identification for each data item (Tork 2012). Figure 1 depicts a set of sampled spatiotemporal events.

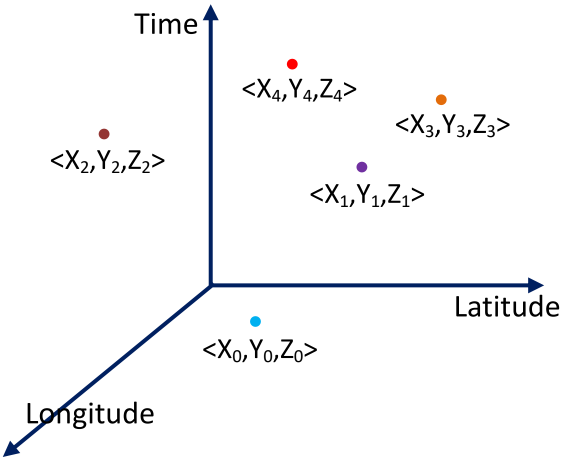

Figure 1. Spatiotemporal events

A geo-referenced data item is usually an observation of a particular phenomenon at a fixed location over a period of time (Kisilevieh et al.2010). It can be defined by a set of spatiotemporal sequences, where each sequence element consists of spatial, temporal, and non-spatial attributes, i.e., <longitude, latitude, timestamp, non-spatial value> (Kalyani and Chaturvedi 2012; Giannotti et al. 2008). A geo-referenced data item represents the most recent value. A typical example of a geo-referenced data item is a weather station location along with most recent temperature value. Figure 2 depicts a sample of geo-referenced data items, where only the most recent values (at t = 3) are kept, as shown by the solid rectangle, and the historical data are represented by the dotted rectangles.

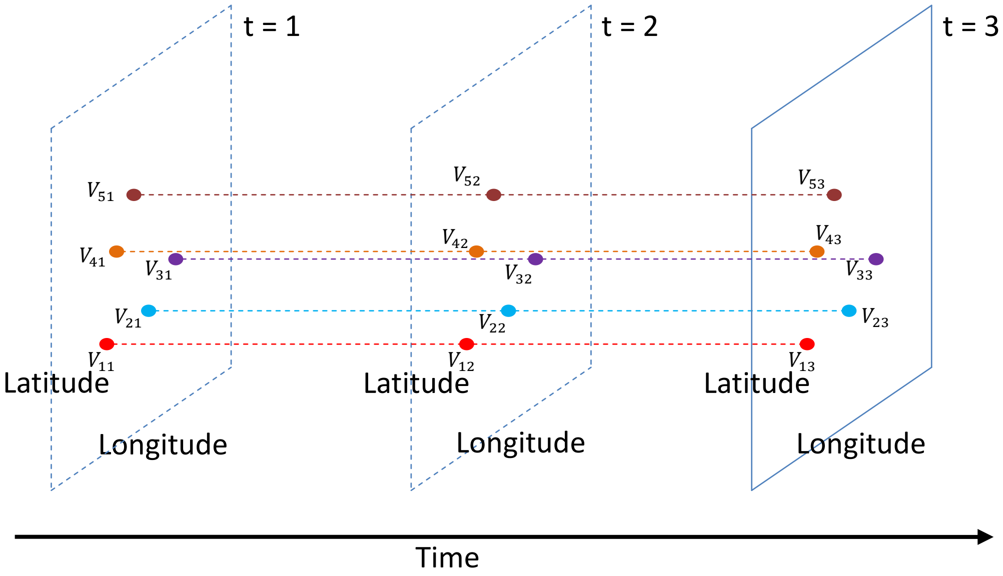

Figure 2. Geo-referenced data items

A geo-referenced time series keeps the whole history of the evolving object over a period of time (Kisilevieh et al.2010). Typical examples include the measurement of the daily temperature at a location over years (Zhang et al.2003), monthly reported cases of disease in different cities, and hourly recordings of air pollution (Izakian et al.2013). Several geo-referenced data form a single multidimensional time series. Here, the challenge is in the detection of correlations among different time series to nullify mutual interference effects among objects owing to spatial proximity (Zhang et al.2003). Figure 3 depicts a sample of a geo-referenced time series. The solid rectangles indicate that the geo-referenced time series keeps the whole history of all the objects and thus represents their evolution.

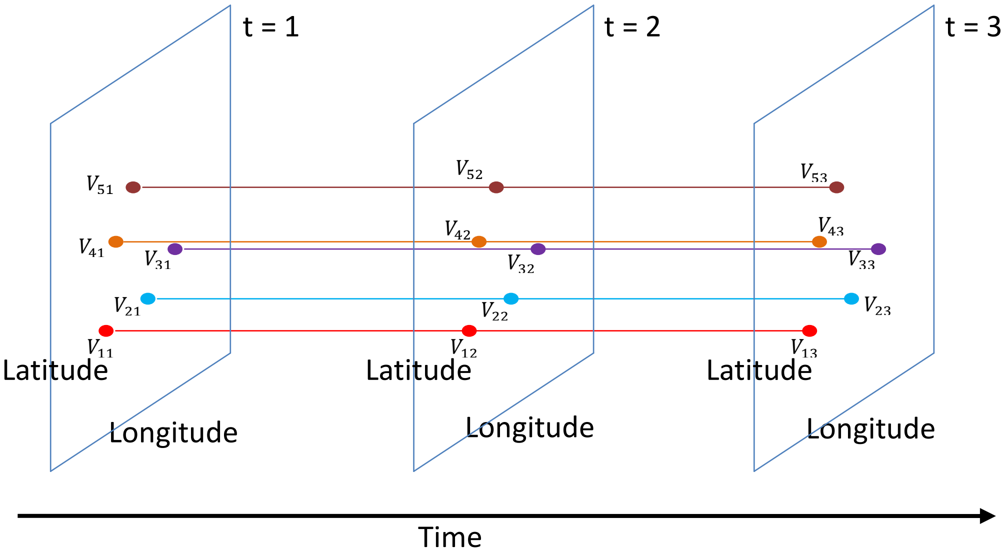

Figure 3. Geo-referenced time series

A moving object changes its spatial location with respect to time and has an identifier to trace the movement over a period of time (Tork 2012). Here, the most recent positions are maintained, and a history of past locations is not required (Kisilevieh et al.2010). An example of a moving object is real-time monitoring of vehicles for security applications. The moving object is defined by a set of sequences of the form <id, x, y, t>, where id is the object identity, and the x and y are spatial attributes of the moving object at time t (Zaghlool et al. 2015). Figure 4 depicts a sample of moving objects, where the dotted lines indicate the recent values only. For example, <2, X_7, Y_7, T_3> denotes current spatial location (at t = 3) of an object with id = 2.

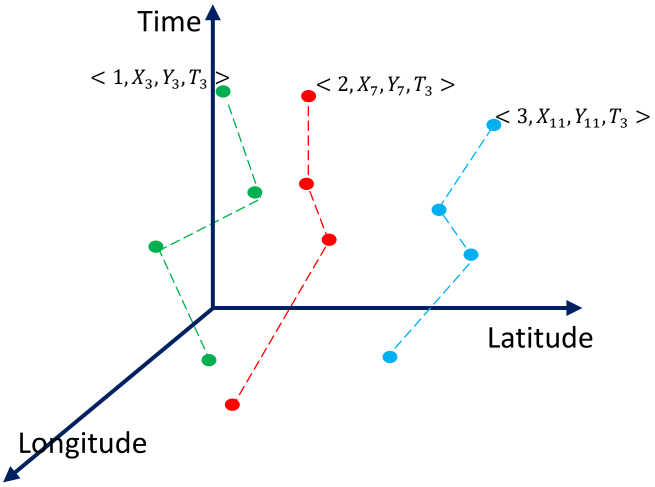

Figure 4. Moving objects

A trajectory of an object is a sequence of spatial locations with time-stamps. In the case of trajectories, the whole histories of the moving objects are kept, which describe the movement behaviors of the objects (Kisilevieh et al.2010). A trajectory can be denoted by a list of space-time points (p_0,p_1,…,p_n), where p_i=(x_i,y_i,t_i); x_i, and y_i are spatial attributes, and t_i is a temporal attribute with t_0<t_1<....<t_n (Rinzivillo et al. 2008). Figure 5 depicts trajectories of objects, where the solid lines indicate the whole histories of moving objects.

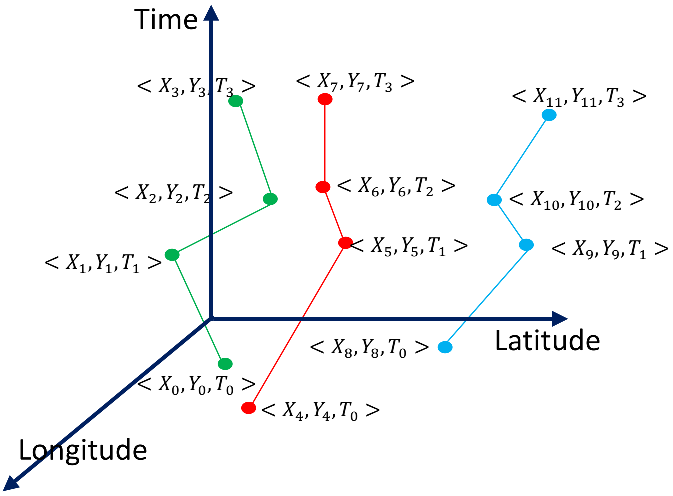

Figure 5. Trajectories

Spatial and spatiotemporal data require data pre-processing, transformation, data mining, and post-processing techniques to find new and useful patterns (Bogorny and Shashi 2010). In case of spatiotemporal clustering, careful data pre-processing is essential, because irrelevant attributes may negatively affect proximity measures and eliminate clustering tendency (Berkhin 2006). The pre-processing activity includes data cleansing, handling missing values, attribute representation and encoding, and generating derived attributes. The selection of the proper subset of attributes is a crucial step in accurate and efficient modeling (Becher et al.2000). To simplify the Knowledge Discovery in Data (KDD) process, Becher et al. (2000) present an automated exploratory data analysis (EDA) approach to the exploration, preprocessing, and selection of the optimal attribute subset. Different attribute selection methods exist for predictive mining, but not for descriptive mining, and as a result, they do not address general-purpose attribute selection for clustering (Berkhin 2006).

To construct clusters of better shapes and attributes, the scaling and assigning of different weights to attributes may be required. Certain algorithms, such as Density Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al. 1996), make use of spatial access methods such as R*-tree (Beckmann et al. 1990) to process very large databases (Ester et al. 1996). The rapid access of data in spatiotemporal databases depends on the structural organization of the stored data and suitable indexing methods. A suitable indexing method and a well designed data structure provide a mechanism to quickly locate single or multiple objects and extract desired information from a database (Li et al. 2004). Well known spatial indexing techniques include R-tree (Guttman 1984), R+-tree (Sellis et al. 1987), R*-tree (Beckmann et al. 1990), Quad-tree (Samet 1985), kd-tree (Bentley 1975), LSD-tree (Henrich et al. 1989), and others described in (Güting RH 1994). R-tree and its variants have capability to support spatial data objects (Guttman 1984). Spatial data objects consist of geometries such as points, polygons, and curves. The Quad-tree structure can be used for points, regions, curves, and volumes (Samet 1985). The kd-tree data structure supports k-dimensional points in space (Bentley 1975). LSD-tree has the ability to handle multidimensional points, multidimensional intervals, and arbitrary geometric objects (Henrich et al. 1989). R-tree, R+-tree and R*-tree are unable to efficiently index moving objects (Cai and Revesz 2000). Because R-Tree uses overlapping bucket regions, its exact match performance is poor (Henrich et al. 1989). R-tree does not utilize space efficiently, where as R+-tree utilizes space more efficiently than R-tree. The implementation cost of R*-tree is greater than those R-tree and R+-tree. Kd-tree is efficient in terms of its storage requirements and is able to handle many types of queries efficiently (Bentley 1975). Kd-tree is an effective index structure for range queries and nearest-neighbor searches, provided the dimension is not too high (Otair 2013). To obtain good performance from Quad-tree, the approximation of geometries needs to be fine-tuned. The tuning is a complex process, and Quad-tree is therefore not recommended for use in particular spatial indexing compared to R-tree (Sardadi et at. 2008). The search performance of LSD-tree may be inefficient, because it is not a balanced tree (Zhou and Salzberg 2008).

An spatiotemporal indexing technique can employ multi-dimensional spatial indexing techniques, with time as an extra dimension (Güting 1994). To handle spatiotemporal information, an improvement was made to the R-tree indexing method by Birant and Kut (2007). The authors created certain nodes for each spatial object in R-tree and linked them in temporal order. The tree is traversed to find the spatial or temporal neighbors of any object. Two objects are considered as temporal neighbors if their values are observed at consecutive times such as on consecutive days in the same year or on the same day in consecutive years.

A summary of spatial index structures, their brief descriptions, and selected literature references are provided in Table1.

Table 1. Overview of spatial index structures

Spatial index structure	**Brief description**	**Selected literature**
R-tree	A multiway tree structure using overlapping bucket regions	Guttman (1984)
R+-tree	A variation to R-tree, which avoids overlapping regions by clipping data rectangles	Sellis et al. (1987)
R*-tree	A variation to R-tree, which incorporates a combined optimization of area, margin and overlap of each enclosing rectangle	Beckmann et al. (1990)
Quad-tree	A tree representing regions by successively subdividing them into four equally sized quadrants	Samet (1985)
Kd-tree	A multidimensional binary search tree for storage of information to be retrieved by associative searches	Bentley (1975)
LSD-tree	A data structure supporting efficient access to geometric objects	Henrich et al. (1989)
Improved R-tree	An R-tree with capability to handle spatiotemporal data	Birant and Kut (2007)

After introducing the concept of spatiotemporal data and data pre-processing requirements, the next section describes the various spatiotemporal clustering techniques present in the literature.

3. Spatiotemporal Clustering Techniques

As noted earlier, spatiotemporal clustering is an extension of spatial clustering in which the time dimension is introduced into spatial data (Tork 2012; Birant and Kut 2007). We identify six categories of spatiotemporal clustering algorithms based on the type of spatiotemporal data. Four categories correspond to events, geo-referenced data items, geo-referenced time series and moving objects, and two categories correspond to trajectories. The six broad categories of spatiotemporal clustering algorithms are:

(i) Event clustering – focus on the discovery of groups of events that is close to each other with respect to space and time.

(ii) Geo-referenced data item clustering – discover the groups of objects that are similar to each other in respect of spatial and non-spatial attributes at any given instance of time.

(iii) Geo-referenced time series clustering – aim to group objects based on spatial closeness and associated time series similarity among objects.

(iv) Moving clusters – aim to detect the behaviour of moving objects. In this case, the identity of a moving cluster does not change although its location and content may change over time.

(v) Trajectory clustering – aim to detect group of objects that have similar movement behaviour.

(vi) Semantic based trajectory data mining – incorporates domain-specific information in the trajectories in a pre-processing step, so that information such as important places, stops and moves etc. can be extracted from the trajectories by applying data mining algorithms.

We present a comprehensive literature review and recent advances in spatiotemporal clustering algorithms within each category in the following subsections.

3.1 Event clustering

Event clustering involves the discovery of groups that are close to each other with respect to space and time, and possibly share other non spatial attributes (Kisilevieh et al.2010). The clustering is performed by considering both spatial and temporal dimensions, as depicted in Figure 6. To analyze sequences of seismic events, Wang et al. (2006) suggest the

ST-GRID spatiotemporal algorithm. The ST-GRID approach employs the partitioning of the spatial and temporal dimensions into a multi-dimensional grid with different precision, allocation of patterns into grid cells, and finally extraction as well as merging of dense spatiotemporal regions into clusters (Wang et al.2006). The ST-GRID algorithm is based on a neighborhood searching strategy and relies on a sorted k-dist graph for the determination of input parameters.

Georgoulas et al. (2013) present an approach, Seismic Mass DBSCAN (SM-DBSCAN), for seismic event clustering. SM-DBSCAN is based on DBSCAN and introduces the notion of accumulated seismic mass for the isolation of clusters of seismic events in time and space. For a second- level spatial clustering, the single linkage agglomerative hierarchical clustering technique is used. This algorithm has the ability to find irregularly shaped clusters, which is helpful in identifying irregular seismic zones. The shortcoming of the approach is that the results heavily depend on the selection of the cumulative seismic mass parameter.

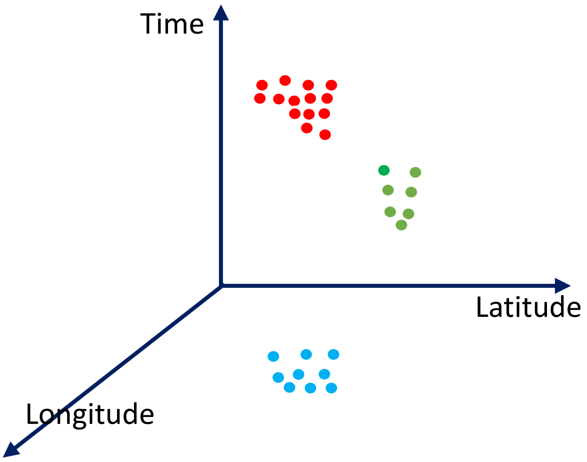

Figure 6. Event clustering

Lee (2012) develops density-based online clustering algorithms for mining micro-blogging text streams, to obtain real-time and geospatial event information. The objectives of situational awareness and risk management in the case of a natural disaster could be achieved by analyzing the spatiotemporal impacts of emerging events. The lack of semantics integrity in tweets poses a challenge to designing weighting and clustering algorithms. The proposed approach addresses this issue by adopting a dynamic weighting method. Events that occur in small spatial areas could not be analyzed by the proposed algorithm.

Han et al. (2016) develop an unsupervised crowd activity discovery algorithm to automatically mine visual patterns that exist in a large number of crowd activities and partition them into meaningful clusters. The algorithm employs an spatiotemporal saliency strategy to capture human-focused components in activities. A graph-based clustering algorithm, Normalized Cut, is applied to cluster the crowd activity dataset. To compute the similarity between activities, Dynamic Time Warping (DTW) techniques are used. This approach does not perform satisfactorily on highly imbalanced data.

Huang et al. (2019) propose an spatiotemporal clustering-based approach to detect human activities in daily traffic congestions from geo-tagged tweets. Space, time, and semantics are associated with human activity. The DBSCAN-based spatiotemporal clustering approach is applied to indentify the space and time of an activity. To infer the topic, modeling is performed on tweets associated with each cluster. This approach is helpful in urban planning and policy making, as well as traffic event detection using social media data.

Martino et al. (2018) propose an Spatiotemporal Extended Fuzzy C-Means (SEFCM) clustering algorithm for detecting spatiotemporal hotspots. The approach adopts the distance function used in the augmented Fuzzy C-Means (FCM) algorithm (Izakian et al. 2013), which considers the spatial and temporal features of patterns. The multiplicative parameter λ is used in the temporal part of the distance function to control the effect of the spatial and temporal parts for the calculation of overall distance. To find the optimal value of λ, the reconstruction error is used. The algorithm is applied to seismic events that occurred in Southern Italy from 2001 to 2014. The results of the algorithm are comparable to those of the ST-DBSCAN algorithm suggested by Birant and Kut (2007).

The statistical analysis for spatiotemporal cluster detection is an important research domain. These clusters mainly refer to hotspots and also have many diverse application domains. The following sub section briefly describes statistical methods for cluster detection problem which is a class of methods of event clustering.

3.1.1 Space-time Scan Statistics

Space-time scan statistics (Kulldorff 1997; Kulldorff 2018) are employed to identify statistically significant hotspots in spatiotemporal data. The cylinder is used to scan space-time to identify candidate hotspots and then hypothesis testing is performed. For each candidate hotspot, the log-likelihood ratio is calculated and the highest likelihood ratio is evaluated using a significance value. The spatiotemporal hotspots have many diverse application domains, ranging from public health to criminology (Shekhar et al. 2015).

Upton and Fingleton (1985) propose distance-based and quadrat-based approaches for the analysis of spatial point patterns and apply the approaches to disease clusters. Kulldorff and Nagarwalla (1995) propose a quadrat method to detect the location of possible disease clusters in a population with inhomogeneous spatial density and at the same time use methods of inference to test for significance. The test is based on the likelihood ratio and can be applied to both aggregated and non-aggregated data.

To detect spatial and spatiotemporal clusters automatically, Neill (2006) propose a flexible, model-based generalized spatial scan statistical framework, which is computationally efficient and applicable to diverse domains. The framework employs a fast spatial scan algorithm and new Bayesian cluster detection methods for faster computation and "expectation-based" scan statistics for timely detection of emerging clusters.

Hotspots indicate the location and time of occurrence of fire and act as an early indicator of fire. Hudjimartsu et al. (2018) employs an spatiotemporal clustering approach with the Kulldorff Scan Statistic (KSS) to identify fire hotspots. Kirana et al. (2016) apply statistical techniques to identify the distribution pattern of hotspot grouping based on spatial and temporal aspects using KSS method.

Anbaroglu et al. (2015) suggest percentile-based and space–time scan statistics (STSS)-based methods to detect non-recurrent congestion (NRC) events in heterogeneous urban road networks for effective management. The detected NRC is based on link journey time estimates.

3.2 Geo-referenced data item clustering

Geo-referenced data item clustering involves the discovery of groups that are similar to each other with respect to their spatial attributes at any given time instant, and whose non-spatial attributes are not constant (Kisilevieh et al. 2010). The cornerstone of distance-based clustering is defining distance functions that make use of similarity among data items (Kisilevieh et al.2010). Density-based approaches utilize the density threshold around each object to distinguish the relevant data items from the noise.

Birant and Kut (2007) propose a density-based clustering algorithm for spatiotemporal data, ST-DBSCAN, which is an extension of DBSCAN (Birant and Kut 2007). ST-DBSCAN has the ability to cluster spatiotemporal data according to non-spatial, spatial, and temporal attributes by employing an spatiotemporal distance function. If there are two points A(x1,y1,t1,t2) and B(x2,y2,t3,t4) where x1,y1 and x2, y2 are spatial values, t1, t3 are the day temperature and t2, t4 are the night temperature, then the closeness of two geographical points can be calculated as $\sqrt{(x1-x2)^2+(y1-y2)^2}$ and the similarity of the non spatial values is computed as $\sqrt{(t1-t3)^2+(t2-t4)^2}$.

The inability of the DBSCAN algorithm to detect certain noise points in the case of different densities is addressed with the ST-DBSCAN algorithm by introducing the notion of a density factor with each cluster. In case the differences in the values of neighboring objects are small and there is significant difference in the values of border objects in a cluster compared to the values of other border objects on the opposite side, the ST-DBSCAN algorithm addresses this issue by comparing the average value of a cluster with a new value. The algorithm needs four input parameters from the user, which heavily influences the quality of the clustering results.

Agrawal et al. (2016) develop and validate an enhanced spatiotemporal clustering algorithm, Spatiotemporal – Ordering Points to Identify the Clustering Structure (ST-OPTICS), by modifying OPTICS. The scalable technique has the ability to identify nested and adjacent clusters and to handle multi dimensional data. The approach first sorts the observation and handles the spatiotemporal data in which spatial and non-spatial attributes are handled using the spatial distance (ε1) and non-spatial distance(ε2), respectively. The temporal dimension is handled by concatenating corresponding spatial and non spatial values while retaining the temporal neighbours (period-wise). To improve the visualization and analysis of the generated micro level clusters, the result of the spatiotemporal clustering algorithm acts as an input to an agglomerative approach. The authors also present quantitative and qualitative performance evaluations of the obtained results. The visualization approach is used to evaluate the quality of the clusters. To quantitatively evaluate the results, different performance indices (i.e., cluster validation indices) are studied by the authors using theoretical principles. The suitable indices for the validation of dense and arbitrarily shaped clusters are selected. The authors also perform an experiment on an spatiotemporal dataset (Agrawal et al. 2016); their results show improved performance in both quantitative and qualitative terms as well as in the run time efficiency of ST-OPTICS compared to the ST-DBSCAN algorithm suggested by Birant and Kut (2007). The limitation of the approach is that it does not support spatial indexing structures.

Liu et al. (2018) propose the dual-constraint spatiotemporal clustering approach (DcSTCA) to mine marine clustering patterns. The approach works in three phases. In the first phase, an spatiotemporal grid cube is generated based on of spatial connectivity and the time evolution process of marine anomaly variations. In the second phase, the spatiotemporal density and clustering cores are obtained using the space, time, and thematic attributes. In the third phase, the spatiotemporal clustering patterns are constructed as per the density connectivity of spatiotemporal neighbors and clustering cores. The authors also perform an experiment on simulated datasets; their results show improved performance in terms of effectiveness compared to the ST-DBSCAN algorithm suggested by Birant and Kut (2007).

3.3 Geo-reference time series clustering

The geo-referenced time series clustering of objects requires comparison of the time series evolution of objects in relation to the object’s spatial positions (Kisilevieh et al. 2010). Izakian et al. (2013) introduce the concept and algorithmic framework of fuzzy clustering for geo-referenced time series data. The approach uses the generic FCM algorithm (Izakian et al.

2013) by modifying its objective function to make it applicable to spatiotemporal data. A modified spatiotemporal distance function is used in the objective function, which is defined as follows:

$d_{\lambda}^{2}(v_i, x_k) = \left|\left|v_i(s) - x_k(s)\right|\right|^2 + \lambda\left|\left|v_i(t) - x_k(t)\right|\right|^2, \lambda \geq 0,$ where $v_i(s)$ and $x_k(s)$ are the spatial parts and $v_i(t)$ and $x_k(t)$ are the temporal parts.

To control the effects of the spatial and temporal parts for the calculation of the overall distance, λ is used. The authors introduce a reconstruction error and a prediction error as optimization criteria, and the same is used to optimize the performance of the clustering technique. The approach is unable to work on multivariate time series.

Izakian et al. (2015) propose three alternatives for fuzzy clustering of time series to reveal the structures within data, using Dynamic Time Warping (DTW) distance. The proposed techniques cluster time series data based on the shape information. The FCM, fuzzy C-medoids, and a hybrid of both are employed by the alternative techniques.

Izakian and Pedrycz (2013) propose a framework to detect amplitude and shape anomalies in time series. The approach generates a set of sub-sequences of time series using a fixed-length sliding window, after which FCM clustering is used to reveal the structure of time series. Dissimilarity is measured using reconstruction criteria. The original representation of time series is used to detect amplitude anomalies, and the autocorrelation representation of time series is used to detect shape anomalies. The approach lacks in estimating an anomaly score based on the nature of the data and the application.

Doborjeh and Kasabov (2015) propose a clustering method for dynamic spatiotemporal brain data. The method is based on the NeuCube spiking neural network (SNN) architecture. The intensity of spike communication within SNN cube is considered as the spatiotemporal similarity measure in the formation of clusters. The clusters reflect the dynamic spatiotemporal process of the brain. The authors illustrate the method using functional magnetic resonance imagery (fMRI).

Doborjeh et al. (2018) propose two clustering methods for spatiotemporal data based on the NeuCube SNN architecture. One method is based on unsupervised learning and the other employs a supervised learning approach. The clusters are generated online in continuous and incremental manner. The authors illustrate the method on spatiotemporal EEG data. The methods are helpful to discover differences in temporal sequences and the involvement of brain regions in response to cognitive tasks.

Husch et al. (2018) propose an spatiotemporal clustering algorithm, Correlation based Clustering of Big Spatiotemporal Datasets (CorClustST), based on empirical correlations of spatial neighbors over time. The spatiotemporal neighborhood is defined on the basis of Pearson's sample correlation (Pearson 1895) between the time series of two spatial locations. The spatiotemporal clustering algorithms, such as ST-DBSCAN (Birant and Kut 2007) and ST-OPTICS (Agrawal et al. 2016), demand a number of parameter settings for optimal clustering solutions, and therefore the clustering results of different scenarios such as with varying time frames are difficult to compare. CorClustST has the ability to compare and interpret clustering results for different scenarios. The authors demonstrate that the algorithm can be extended for large-scale parallelization. The algorithm is applied to the cluster analysis of wind power forecast errors in Europe.

3.4 Trajectory clustering

Trajectory clustering for spatiotemporal data, as depicted in Figure 7, is performed by choosing an appropriate clustering algorithm and distance function in the case of distance based clustering. The shape of the cluster depends on the chosen algorithm. Spatiotemporal trajectory similarity is dependent on the specific application. The distance functions defined in various approaches such as Pelekis et al. (2007) make use of the spatiotemporal characteristics of trajectories, such as direction, velocity, and co-location in space and time. The approach does not support appropriate indexing structures to improve the performance of operators. Auria et al. (2006) exploit the temporal aspect for the calculation of the distance function.

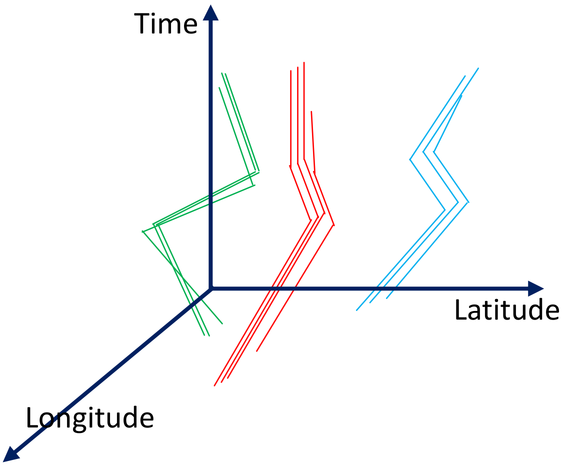

Figure 7. Trajectory clustering

Descriptive and generative model-based approaches aim to derive a global model for describing an entire data set. Gaffney and Smyth (1999) develop a clustering technique on the basis of a mixture model for continuous trajectories. This technique represents trajectories as functional data. The authors employ the expectations-maximization (EM) algorithm to group the Gaussian noise and generate objects from a core trajectory. The proposed approach is a special case of a more general hierarchical framework.

Chudova et al. (2003) present a family of algorithms to simultaneously align the spatial and temporal shifts of trajectories within each cluster. To recover the mean curve shape, and the most likely shifts, offsets, and cluster memberships of each curve, the authors employ EM algorithm. Further, Alon et al. (2003) propose a model-based approach for clustering time-series data in which the cluster representative is expressed by hidden Markov models (HMMs) to estimate the transitions between successive positions. The EM framework is used for parameter estimation of the model. The proposed approach lacks a robust method to handle outliers.

A partition-and-group framework is proposed by Lee et al. (2007) to cluster trajectories. Their Trajectory Clustering (TRACLUS) algorithm, works in two phases: partitioning and grouping. In the partitioning phase, the trajectory is divided into a collection of line segments using the principle of minimum description length. The grouping phase employs a density-based line segment clustering approach, which is based on the DBSCAN algorithm (Lee et al. 2007). The line segment clustering algorithm employs distance function that makes use of three components: the perpendicular distance, parallel distance, and angular distance between two line segments.

Suppose there are two line segments, $L_i=(s_i,e_i)$ and $L_j=(s_j,e_j)$, as shown in Figure 8. The perpendicular distance between line segments L_i and L_j can be computed by equation (1), the parallel distance between the line segments can be computed by equation (2), and the angular distance between the line segments can be computed by equation (3).

$$d_{prp}(L_i, L_j) = \frac{l_{prp1}^2 + l_{prp2}^2}{l_{prp1} + l_{prp2}} \qquad (1)$$

where l_{prp1} is the Euclidean distance between s_j and p_s , and l_{prp2} is the Euclidean distance between e_j and p_e .

$$l_{prl}(L_i, L_j) = Min(l_{prl1}, l_{prl2}) \qquad (2)$$

where l_{prl1} is the Euclidean distance between s_i and p_s , and l_{prl2} is the Euclidean distance between e_i and p_e .

$$d_\theta(L_i, L_j) = \left|\left|L_j\right|\right| sin(\theta) \qquad (3)$$

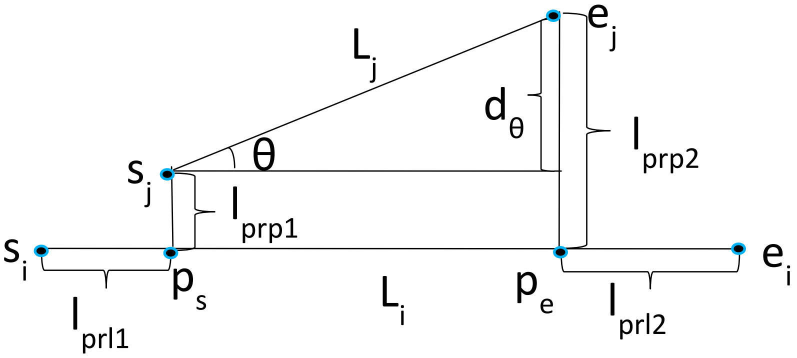

Figure 8. Components of distance function of line segments

The proposed approach is sensitive to parameter values and does not consider temporal information during clustering.

Zaghlool et al. (2015) present an spatiotemporal clustering algorithm that clusters sub-trajectories based on the time dimension. This algorithm is based on TRACLUS (Lee et al. 2007) and employs the spatiotemporal distance function the generalized spatio-temporal locality in-between poly-lines (GenSTLIP) to measure trajectory similarity.

Higgs and Abbas (2015) propose a two-step algorithm for segmentation and clustering to investigate the characteristics of driving behaviors, and link driving states to the driver's actions. Each car-following period is divided into segments of similar driving states and actions after that the repeated segments are clustered by employing a K-means clustering approach. The resultant clusters contain similar sets of driving states and corresponding actions. The methodology reveals the heterogeneity among car drivers and homogeneity among truck drivers.

Zhang et al. (2018) propose a hierarchical trajectory clustering approach based on TRACLUS (Lee et al. 2007) for spatiotemporal periodic pattern mining. The approach considers semantic spatiotemporal information such as direction, speed, and time in the clustering process. The proposed approach extends HDBSCAN (Campello et al. 2013) which is an incremental version of DBSCAN (Ester et al. 1996).

Fiori et al. (2016) propose a Density Consensus Clustering (DeCoClu) approach to mine the topology of a public transport network. The approach uses time series information of positioning signals (i.e., GPS data) to deduce the locations of stops based on consensus clustering by formulating a new distance function and defining a route path.

3.5 Moving clusters

In the moving cluster problem, the identity of a moving cluster does not change although its location and content may change over time. For example, some animals may enter or leave a group of migrating animals. An example of moving clusters is depicted in Figure 9.

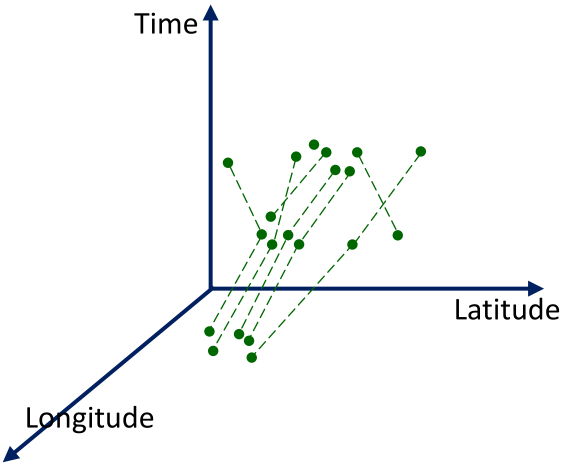

Figure 9. Moving clusters

Kalnis et al. (2005) examine the problem of identifying moving clusters in large spatiotemporal datasets. The authors propose three algorithms to identify moving clusters automatically. The first algorithm is simply an implementation of problem definition, the second one improves the efficiency by avoiding redundant checks, and the final one is an approximate algorithm that trades accuracy for speed. The essence of the algorithms is that at every timestamp t=i, the snapshot $S_{t=i}$ of the object's positions is taken. After that, the DBSCAN algorithm is applied to the snapshot to form a cluster $C_{t=i}$. For two snapshot clusters $C_{t=i}$ and $C_{t=i+1,}$ the moving cluster $C_{t=i}$ $C_{t=i+1}$ is formed if $\frac{|Ct=i \cap Ct=i+1|}{|Ct=i \cap Ct=i+1|} > \theta$, where θ is an integrity threshold for the contents of the two clusters. The efficiency and accuracy of the approach depends on the appropriate selection of parameters and the accurate estimation of errors and it lacks a self-tuning method for parameter selection.

Chen et al. (2015) propose an spatiotemporal clustering approach for identification of dynamic clusters in the presence of noise and missing data. The size, shape, and statistical properties of clusters may change over time. The algorithm is an extension of DBSCAN, and identifies "core points" that have as a minimum *m* neighbors within a distance ε in the feature space. It considers adjacent points in space that have similar feature values over a non-trivial time window. The technique uses both spatial and temporal information for finding Eps neighbors, similar to ST-DBSCAN (Birant and Kut 2007).

The spatiotemporal distance function is defined as

$$d_{st}(x,y) = \begin{cases} d_t(x,y), if\ x\ and\ y\ are\ spatial\ neighbours \\ \qquad 0,\ \ otherwise \end{cases}$$

where $d_t(x,y)$ is a time-series distance function related to the application.

This approach identifies clusters for each time step and associates these clusters across time, regardless of noise and missing data. The authors manually choose the size of the time window and the method lacks a systematic way to choose the time window size.

Li et al. (2004) present the concept of a moving micro cluster by extending the micro clustering approach to spatiotemporal data (Zhang et al. 1996). The authors also propose efficient algorithms to keep moving micro clusters geographically small. In case the segments of different trajectories occur in similar time intervals, they are grouped within a given rectangle. The approach lacks the ability to discover interesting clusters of various forms.

Hwang et al. (2005) present an approach to discover moving object group patterns from a trajectory database. The disconnected moving object behavior is modeled by non-continuous trajectories. The concept of trajectory similarity is based on the duration of time during which trajectories are near. The trajectories that are near within a given threshold are clustered. The proposed approach is unable to handle location data with uncertainty.

Jeung et al. (2008) extend the moving cluster approach of Kalnis et al. (2005) to solve the problem of convoy discovery. The authors propose three efficient algorithms for convoy discovery based on the filter-refinement framework. In the first step, the line simplification techniques are employed on trajectories and distance bounds between the simplified trajectories are established. This allows efficient discovery of convoys without missing any actual convoys. The refinement step further processes the candidate convoys to obtain the actual convoys.

Flock patterns discovery in moving objects is the process of identification of all groups of trajectories that stay together for a given duration. Vieira et al. (2009) demonstrate that the on-line flock discovery problem is polynomial and present a frame work and techniques for the discovery of flock patterns in spatiotemporal data streams.

3.6 Semantic-based trajectory data mining

In semantic-based trajectory data mining, domain-specific information is included in the trajectories in a pre-processing step, after which data mining algorithms are applied to the trajectories.

To extract important places in single trajectory, Kang et al. (2005) propose an incremental clustering algorithm. The algorithm does not require the number of clusters as input, has the ability to exclude unimportant locations, and does not require a significant amount of computation. The distance between positions and time spent in a cluster controls the creation of a cluster. The result of the approach depends on the distance and time thresholds. The user needs to choose theses parameters carefully for better results. The approach also lacks in automatic labeling of extracted places.

Zheng et al. (2009) develop a hypertext-induced topic search (HITS)-based model to extract important places and travel sequences in a geospatial region, using the GPS trajectories generated by multiple users. The approach lacks in grouping users based on location histories.

Palma et al. (2008) propose an spatiotemporal clustering approach called clustering-based stops and moves of trajectories (CB-SMoT), which uses the speed of a trajectory as the criterion to assign stops. CB-SMoT first creates the cluster of the slow speed part, after which the technique matches the clusters with appropriate geographic places. CB-SMoT follows the same principle as DBSCAN; first, it looks for core points, and then expands them by aggregating other points in the neighborhood. The approach addresses only spatial and speed-based semantics and lacks in handling other types of semantics in trajectories, such as acceleration.

Rocha et al. (2010) present an spatiotemporal clustering approach, direction-based stops and moves of trajectories (DB-SMoT), to discover stops of trajectories. The approach uses direction as the main threshold to find the clusters. The approach addresses only spatial and direction-based semantics and lacks in handling other types of semantics in trajectories, such as a combination of speed and direction.

Table 2 summarizes the categories of spatiotemporal data, problems considered in each category, clustering approaches to address the problems and selected literature.

Table 2. Overview of spatiotemporal clustering approaches

Category	Problem	Clustering approach	Selected literature
Event clustering	Analysis of sequences of seismic events	ST-GRID and ST-DBSCAN (based on DBSCAN)	Wang et al. (2006)
	Seismic event clustering	SM-DBSCAN (based on DBSCAN)	Georgoulas et al. (2013)
	Analysis of spatiotemporal impacts of the emerging events	Density-based online clustering algorithm for mining micro-blogging text streams	Lee (2012)
	Discovery and analysis of unsupervised crowd activities	Spatiotemporal saliency strategy and Normalized Cut clustering	Han et al. (2016)
	Detection of human activities on daily traffic congestions from geo-tagged tweets	DBSCAN-based spatiotemporal clustering approach	Huang et al. (2019)
	Detection of spatiotemporal hotspots	Spatiotemporal extended fuzzy C-Means (SEFCM)	Martino et al. (2018)
Geo-referenced data item clustering	Seawater distribution	ST-DBSCAN (based on DBSCAN)	Birant and Kut (2007)
	Forest and vegetation distribution	ST-OPTICS (based on OPTICS)	Agrawal et al. (2016)
	Mining of marine clustering patterns	Dual-constraint spatiotemporal clustering approach (DcSTCA)	Liu et al. (2018)
Geo-referenced time series clustering	Suitable treatment of spatial and temporal components of the data	Based on generic Fuzzy C-means (FCM) algorithm	Izakian et al. (2013)
	Partition-based clustering of time series data using DTW distance	Based on Fuzzy C-Means and Fuzzy C-Medoids	Izakian et al. (2015)
	Anomaly detection in time series	Based on generic Fuzzy C-means (FCM) algorithm	Izakian and Pedrycz (2013)
	Dynamic spatiotemporal process of brain	Based on NeuCube spiking neural network (SNN) architecture	Doborjeh and Kasabov (2015)
	To discover differences on temporal sequences and involvement of brain regions in response to cognitive tasks	Based on NeuCube spiking neural network (SNN) architecture	Doborjeh et al. (2018)
	Cluster analysis of wind power forecast errors	Based on empirical correlations of spatial neighbors over time (CorClustST)	Husch et al. (2018)
Trajectory clustering	Clustering of short sequences of clustering data.	Probabilistic mixture regression model and Expectation-Maximization (EM) algorithm	Gaffney and Smyth (1999)
	Clustering of multidimensional curve data	Expectation-Maximization (EM) algorithm and Bayesian estimation method	Chudova et al. (2003)
	Clustering of similar object motions and estimation of motion time series models	Hidden Markov models (HMM)-based clustering algorithms (EM and k- means)	Alon et al. (2003)
	Trajectory classification and similarity search	Distance function- based approaches that uses spatiotemporal characteristics of trajectory	Pelekis et al. (2007)
	Discovery of common sub trajectories from trajectory database (hurricane track data and animal movement data)	TRACLUS (based on DBSCAN)	Lee et al. (2007)
	Cluster and analyze spatio-temporal data (hurricane track data and animal movement data)	Extension of TRACLUS	Zaghlool et al. (2015)
	To investigate the characteristics of driving behaviors, linking driving states to the driver's actions	K-means clustering	Higgs and Abbas (2015)
	Spatiotemporal periodic pattern mining	Extension of TRACLUS	Zhang et al. (2018)
	To mine topology of public transport network	Density Consensus Clustering (DeCoClu) approach	Fiori et al. (2016)
Moving clusters	Automatic discovery of moving clusters	Based on DBSCAN	Kalnis et al. (2005)
	Identification of dynamic clusters in the presence of noise and missing data	Extension of DBSCAN	Chen et al. (2015)
	Identification of moving micro clusters	Extension of micro clustering approach	Li et al. (2004)

	Discovery of mobile group patterns	Modeling of disconnected behavior of moving objects as non-continuous trajectories	Hwang et al. (2005)
	Convoy discovery	Based on filter refinement approach and DBSCAN	Jeung et al. (2008)
	Discovery of flock patterns	Based on time joins and filter and refine paradigm	Vieira et al. (2009)
Semantic based trajectory mining	Extracting important places in single trajectory	Incremental clustering approach where location measurements are clustered together	Kang et al. (2005)
	Extracting important places and classical travel sequences within a given geospatial region	Hypertext-induced topic search (HITS)-based model	Zheng et al. (2009)
	Adding semantics to trajectory based on speed	CB-SMoT (Based on DBSCAN)	Palma et al. (2008)
	Adding semantics to trajectory based on direction	DB-SMoT (Based on DBSCAN)	Rocha et al. (2010)

3.7 Analysis of algorithms

This sub section provides the analysis of time complexity of the representative spatiotemporal clustering algorithms in each category. The representative algorithms of event clustering, geo-referenced data item cluster and trajectory clustering are based on DBSCAN algorithm, therefore the analysis of DBSCAN algorithm is also provided in this sub section.

Suppose there are n objects in a database and the DBSCAN algorithm (Ester et al. 1996) makes use of spatial access methods such as R*-trees (Beckmann et al. 1990). The average run time complexity of an Eps-neighborhood query is $\mathrm{O}(\log n)$. There is at most one Eps-neighborhood query for each object, and therefore the average run time complexity of the DBSCAN algorithm is $\mathrm{O}(\mathrm{n} \times \log n)$.

one of the spatiotemporal clustering algorithms in the event clustering category is SM-DBSCAN (Georgoulas et al. 2013), which is based on DBSCAN. Therefore, the average run time complexity of the SM-DBSCAN algorithm is $O(n \times \log n)$. The ST-DBSCAN (Birant and Kut 2007) is the representative algorithm in geo-referenced data item clustering and is based on the DBSCAN algorithm. The modifications by the authors do not affect the run time complexity of the algorithm, which is $\mathrm{O}(\mathrm{n} \times \log n)$.

The augmented FCM algorithm (Izakian et al. 2013) is a geo-referenced time series clustering method and uses the generic FCM algorithm by modifying its objective function to make it applicable to spatiotemporal data. The run time complexity of FCM (Bezdek 1982) is $O(nc^2di)$ where n is number of objects, c is the number of clusters, d is the number of dimensions, and i is the number of iterations. The modification does not affect run time complexity, and therefore the time complexity of the augmented FCM algorithm is $O(nc^2di)$.

The TRACLUS (Lee et al. 2007) algorithm is a representative algorithm of the category of trajectory clustering and is based on the DBSCAN algorithm; therefore, the time complexity of TRACLUS is $\mathrm{O}(\mathrm{n} \times \log n)$, where n is the number of line segments.

The extended micro clustering approach (Li et al. 2004) lies in the category of moving clusters. Suppose there are N moving objects and M micro clusters where $M \ll N$. Let U be the number of object updates, SE the number of split events, and CE the number of collision events. Considering the split events, update handling and maintenance of the kinetic structure, the time complexity is $O((SE \times M \times \log N + \mathrm{CE} + \mathrm{U}\log N)\log(SE + CE + U) + N\log_3 N)$. Considering $SE = O(N),\ CE = O(M^2)\ and\ M \ll N,$ the time complexity become $O((N + U)\log(N + U)\log N + N\ \log_3 N)$.

Table 3. Overview of time complexities of representative spatiotemporal clustering algorithms

Category	Clustering Algorithms	Run time efficiency	Selected literature
Event clustering	SM-DBSCAN	$O(n \log n)$, where n is number of objects.	Georgoulas et al. (2013)
Geo-referenced data item clustering	ST-DBSCAN	$O(n \log n)$, where n is number of objects.	Birant and Kut (2007)
Geo-referenced time series clustering	Augmented Fuzzy C-means algorithm	$O(nc^2di)$, where n is number of objects, c is number of clusters, d is number of dimensions, and i is number of iterations	Izakian et al. (2013)
Trajectory clustering	TRACLUS	$O(n \log n)$, where n is number of line segments	Lee et al. (2007)
Moving clusters	Extended micro clustering approach	$O((N+U)\log(N+U)\log N + N\log_3 N)$, where N is number of moving objects and U is number of object updates	Li et al. (2004)
Semantic-based trajectory mining	CB-SMoT	$O(N + N \log C)$, where N is number of points in trajectory and C is number of candidate stops	Palma et al. (2008)

The CB-SMoT (Palma et al. 2008) spatiotemporal clustering approach lies in the category of semantic-based trajectory mining and uses the speed of the trajectory as the criterion to assign stops. The approach performs the clustering step and also matches every point with background geography. The time complexity of matching with the background geography is $O(N \log C)$, where N is number of points in trajectory and C is number of candidate stops. The time complexity of CB-SMoT in the worst case is the sum of both steps, which is $O(N + N \log C)$.

Table 3 provides an overview of the time complexities of representative spatiotemporal clustering algorithms in each category.

After discussing spatiotemporal clustering techniques, the next section describes the various spatiotemporal patterns present in the literature.

4. Patterns

Different spatiotemporal clustering algorithms exist in mining patterns. The spatiotemporal clustering algorithm produces the pattern, which is interpreted by the domain experts to find novel insights. For example, an spatiotemporal clustering algorithm can mine trajectory patterns from trajectories. This section provides an overview of patterns along with the approaches adopted for pattern extraction in the spatiotemporal domain.

A pattern is a frequent arrangement or regularity, a rule or law, a major direction, or a trend and a prediction (Bogorny and Shashi 2010). Examples of spatial patterns include co-location, outliers, classification, location prediction, spatial association rules, and clustering (Bogorny and Shashi 2010). Pattern mining schemes find existing patterns in data. Spatiotemporal patterns that concisely show the cumulative behavior of a set of moving objects are valuable abstractions to appreciate mobility-related phenomena.

Trajectory patterns are those patterns that are mined from trajectory data. They characterize interesting behavior of single or groups of moving objects (Giannotti and Pedreschi 2008). In the literature, several trajectory pattern mining approaches have been proposed, which can be categorized as repetitive pattern mining, frequent pattern mining, and group pattern mining. Repetitive pattern mining is applicable to the trajectory of single moving objects. Frequent pattern mining and group pattern mining are applicable to the trajectories of multiple objects. In frequent pattern mining, the objects may not move at the same time but need to visit approximately same places in the same sequence, where as in group pattern mining, objects move together (Mazimpaka and Timpf 2016).

Frequent patterns can be defined using spatial or spatiotemporal characteristics of the trajectory. The spatiotemporal sequential patterns are helpful in the analysis and future prediction of the object. The problems in discovering sequential patterns in the trajectory are

(i) the location coordinates do not repeat exactly in every instance of the pattern, and (ii) the identification of non-explicit pattern instances (Cao et al. 2005). To mine frequent spatiotemporal sequential patterns, Cao et al. (2005) define pattern elements as spatial regions around frequent line segments. The authors propose a method to extract pattern elements and a pattern mining algorithm to discover longer patterns.

Giannotti et al. (2007) introduce a trajectory pattern (T-Pattern) as a crisp description of frequent behaviors, in terms of the visited regions and durations, during movements. The trajectory pattern characterizes the interesting behavior of a single or group of moving objects. Two notions central to the determination of T-patterns are the Regions of Interest (RoIs) in the given space, and the travel time of moving objects from region to region. A place visited by several objects forms a RoI. To discover RoIs, the authors propose an algorithm for the detection and identification of popular regions. The working regions are divided into cells and the trajectories that pass through each cell are counted. The approach accepts a grid with cell densities and density threshold as input.

Kang and Yong (2009) present an algorithm for retrieving spatiotemporal patterns in the trajectory data. The approach is based on the approximation of the original trajectory into line segments by preserving spatial information, and then clustering the segments into spatiotemporal regions incorporating temporal constraints. First, the trajectory is simplified using the Douglas-Peucker (DP) algorithm by dividing into segments. The segments are then normalized by adopting min-max normalization. After that, clustering of segments is performed using the BRICH (Zhang et al. 1996) algorithm. Finally, the frequent patterns are extracted by employing a depth-first search algorithm on the clustered regions.

Vieira et al. (2009) introduce a flock pattern, in which the same set of objects stay together in a circular region of a predefined radius. A trajectory satisfies the flock pattern as long as "enough" other trajectories are within the circular region for specified time interval. The authors propose five online spatiotemporal algorithms that work in an incremental fashion. The algorithms use a grid-based structure. Once the grid-based structure is built for a specific time instance, the disks can be processed by one of the algorithms. The candidate disks generated for each time instance are combined into flock patterns by the pattern evaluation algorithm. Moreover, the authors also propose different heuristics to limit the number of generated candidate disks, to improve the performance of the pattern evaluation algorithm. One of the proposed heuristics, cluster filtering evaluation, employs the DBSCAN (Ester et al. 1996) algorithm.

Table 4. Overview of patterns in the spatiotemporal domain

Models and patterns	**Adopted approach**	**Selected literature**
Trajectory pattern (T-Pattern)	Spatiotemporal clustering algorithm development for detection and identification of popular regions	Giannotti et al. (2007)
Spatiotemporal patterns (ST-patterns)	Algorithm development for retrieving ST-patterns in trajectory data by approximating original trajectory into line segments and spatiotemporal clustering it into spatiotemporal regions	Kang and Yong (2009)
Flock pattern	Development of online algorithms that use a grid structure built for a specific time instance; the disks are processed and then combined into flock patterns by using spatiotemporal clustering	Vieira et al. (2009)
Convoy pattern	Spatiotemporal clustering algorithm development by extending moving cluster approach; algorithm development that uses line simplification techniques by adopting filter-refinement framework	Jeung et al. (2008)
Chasing pattern	Algorithm development to extract chasing trajectory behavior pattern, considering time, distance, and speed as the main thresholds	De Lucca and Bogorny (2011)
Avoidance pattern	Algorithm development to detect avoidance pattern that identifies when a moving object avoids a specific spatial region	Alvares et al. (2011)

A convoy is a set of objects that have travelled together for some time. A convoy pattern is defined by Jeung et al. (2008), where groups of objects stay together in a region of arbitrary extent and shape. The authors propose the coherent moving cluster (CMC) algorithm by extending the moving cluster approach (Kalnis et al. 2005). To reduce the overall computational cost incurred by the CMC algorithm, the authors provide convoy discovery mechanism using trajectory simplification (CuTS) algorithms (Jeung et al. 2008), which discover convoys by adopting a filter-refinement framework. In the filter step, line simplification techniques are used to effectively reduce the amount of data that need further processing. The distance bound between simplified trajectories is established to ensure that the filters do not eliminate convoys. The refinement step further processes the candidate convoys to obtain the actual convoys.

Alvares et al. (2011) introduce a new type of behavior pattern called avoidance and propose a new algorithm to detect avoidance patterns that identifies when a moving object avoids a specific spatial region, such as those around security cameras.

De Lucca and Bogorny (2011) introduce a novel behavior pattern called chasing between trajectories, where an object chases another one for a certain period of time. The authors present the main characteristics of chasing and suggest a new approach that extracts chasing the behavior pattern by considering time, distance, and speed as the main thresholds.

Table 4 summarizes various patterns, approaches adopted for pattern extraction, and the selected literature.

To evaluate the goodness of the resulting clusters of the clustering process, some validation mechanism is required. The subsequent section provides an overview of cluster validity methods and indices.

5. Cluster Validation

Cluster validity methods use indices to measure the quality of obtained clusters (Salazar et al. 2002). Cluster validity is associated with inherent features of the dataset (Halkidi et al. 2001). Halkidi et al. (2001) classify validity criteria into external, internal, and relative categories. Salazar et al. (2002) propose a new index called the Q index based on Karl Pearson statistics. Weingessel et al. (1999) analyze the performance of validity indices to determine the number of clusters in binary data set. To evaluate the performance of indices, the authors employ k-means and hard competitive learning approaches. Milligan (1981) evaluate several internal criteria measures for cluster analysis, which are applicable in the context of hierarchical clustering algorithms.

Arbelaitz et al. (2013) investigate the validity of several stopping rules present in the literature and show that noise and cluster overlapping influences cluster validation indices. Halkidi et al. (2001) review cluster validity indices for external, internal, and relative criteria. The authors observe that the external and internal techniques are based on statistical tests and are computationally intensive; whereas relative criteria do not involve statistical tests.

The spatiotemporal clustering algorithms usually produce dense and arbitrarily shaped clusters (Agrawal et al. 2016). In the literature, various cluster validity techniques exist but these techniques are unable to measure arbitrarily shaped clusters (Kovács et al. 2005). Therefore there is a need to define novel validity indices to measure arbitrarily shaped clusters. Kovács et al. (2005) propose validity indices to measure arbitrarily shaped clusters.

To validate the results of the SM-DBSCAN algorithm, Georgoulas et al. (2013) adopt a parameter interplay method and some general patterns that guide the selection of parameter values to achieve reasonable results. Han et al. (2016) adopt Purity, Entropy, Precision Recall, and F scores matrices to measure the performance of the Normalized Cut clustering approach. Birant and Kut (2007) and Agrawal et al. (2016) have adopted a K- nearest neighbors approach to validate the results of the ST-DBSCAN and ST-OPTICS algorithms.

Izakian et al. (2013) use reconstruction error to assess the quality of clusters produced by FCM algorithm. The log-likelihood scores and classification error rate measures are employed by Gaffney and Smyth (1999) for validation of the EM algorithm for mixtures of linear regression models. Alon et al. (2003) use a classification accuracy measure in the presence of ground-truth to validate HMM-based clustering algorithms. Lee et al. (2007) use entropy-based and Sum of Squared Error (SSE) with noise penalty approaches to assess the quality of clusters for the TRACLUS algorithm, where as Zaghlool et al. (2015) use the SSE with noise penalty approach to validate of their algorithm, which is based on TRACLUS. The F scores measure is employed by Kalnis et al. (2005) and Chen et al. (2015) for the validation of their algorithms, which are based on DBSCAN.

Li et al. (2004) use SSE to validate their algorithm, which is an extension of the micro clustering approach. Kang et al. (2005) use precision and recall measures for an incremental clustering approach, where location measurements are clustered together and Palma et al. (2008) evaluated the CB-SMoT algorithm through visual representation.

The Table 5 summarizes the adopted cluster validity indices for various spatiotemporal clustering algorithms, along with selected literature. In the table, we mention many validity indices. A few of these are explained below. Explanations of the other validly indices can be found in papers by the authors Rendón et al. (2011), Liu et al. (2010), Weingessel et al. (1999) , Milligan (1981), Kovács et al. (2005), Halkidi et al. (2001), and Han et al. (2016).

The Calinski and Harabasz (Calinski and Harabasz 1974) index can be computed using two statistics, the sum of squares within (SSW) and sum of squares between (SSB), of obtained clusters. Consider a particular partition with C clusters, N samples in the dataset, and N_i samples in cluster $C_{i.}$. The term x_j is a sample in cluster C_i, v_i is the center of cluster C_i, $d(x_j, v_i)$ is the Euclidean distance between object x_j and cluster center v_i, and $\overline{X}$ is the centroid of the dataset. The statistics can be computed by equation (4), equation (5), and equation (6). The Calinski and Harabasz index is defined by equation (7).

$$SSW = \sum_{i=1}^{C} \sum_{x_j \in C_i} d(x_j, v_i) \qquad (4)$$

$$SSB = \sum_{i=1}^{C} N_i \, d(v_i, \bar{X}) \qquad (5)$$

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (6)$$

Table 5. Overview of cluster validity index

Cluster validity index	**Selected literature**
Q index based on Karl Pearson statistics	Salazar et al. (2002)
Calinski and Harabasz	Weingessel et al. (1999)
Hartigan	Weingessel et al. (1999)
Ratkowsky and Lance	Weingessel et al. (1999)
Ball and Hall	Weingessel et al. (1999), Agrawal et al. (2016)
C Index	Milligan (1981), Weingessel et al. (1999)
Gamma	Milligan (1981), Agrawal et al. (2016)
Davies and Bouldin	Weingessel et al. (1999)
Entropy	Han et al. (2016), Lee et al. (2007)
Precision Recall	Han et al. (2016), Kang et al. (2005)
F scores	Han et al. (2016), Kalnis et al. (2005) , Chen et al. (2015)
Dunn	Agrawal et al. (2016)
Log-likelihood scores	Gaffney and Smyth (1999), Weingessel et al. (1999)
Sum of Squared Error (SSE) with noise penalty	Lee et al. (2007), Zaghlool et al. (2015)
Reconstruction error	Izakian et al. (2013)
TraceW	Weingessel et al. (1999), Agrawal et al. (2016)
Det_Ratio	Agrawal et al. (2016)

$$CH(C) = \frac{SSB/(C-1)}{SSW/(N-C)} \qquad (7)$$

The Ball and Hall (Ball and Hall 1965) index validates the optimal number of clusters produced by the clustering algorithm. The validation of the clusters is performed based on inherent features and quantities in the dataset. This technique measures the dispersion of objects within the cluster, also known as the sum-of-squares within cluster (SSW) (Zhao et al. 2009).The number of clusters is denoted as C and there are N samples in the dataset; x_j is a data element of cluster C_i , v_i is the center of cluster C_i, and $d(x_j, v_i)$ is the distance between object x_j and cluster center v_i. The SSW can be calculated by equation (8), and the Ball and Hall Index is defined as SSW/C.

$$SSW = \sum_{i=1}^{C} \sum_{x_j \in C_i} d(x_j, v_i) \qquad (8)$$

The Dunn index (Dunn 1974) is mainly used to evaluate the compactness and separation of clusters. A large value of this index indicates that the clusters are compact and well separated. The aim is to maximize the inter-cluster distance while keeping the intra-cluster distance as low as possible. The Dunn index is defined as the ratio of the smallest distance between two objects from different clusters to largest distance between two objects of the same cluster, and it is sensitive to outliers. The Dunn index is defined in equation (9), for a specified number of clusters. The function $dist(C_i, C_j)$ defined in equation (10), is used to calculate the distance between the clusters C_i and C_j. The diameter of the cluster C_k is calculated by eqaution (11). $diam(C_k)$ is the maximum distance between any two elements of the cluster and d(u,v) is the distance between two objects.

$$\min_{i=1,\ldots,C} \left\{ \min_{j=i+1,\ldots,C} \left(\frac{dist(C_i,C_j)}{\max_{k=1,\ldots,C}(diam(C_k))} \right) \right\} \qquad (9)$$

$$dist(C_i, C_j) = \min_{u \in C_i, v \in C_j} d(u, v) \qquad (10)$$

$$diam(C_k) = \max_{u,v \in C_k} d(u, v) \qquad (11)$$

To identify appropriate input parameters to produce high-quality clusters for a dataset, the entropy theory can be adopted. Lower entropy indicates better clustering results. Let n be the number of objects and p(x) be the probability. The entropy definition formula (12) is used to compute the appropriate input parameters that minimize H(X).

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \qquad (12)$$

To evaluate patterns in the spatiotemporal domain and present the developed knowledge, visualization and analysis tools are required. The following section briefly describes the latest spatiotemporal analysis tools.

6. Software Tools and Datasets

This section briefly discusses application-specific and generic spatiotemporal analysis tools, including geographic information systems (GIS), spatial database management systems, spatial big data platforms, and free data mining software tools.

6.1 Visualization and Analysis Tools

The analysis of spatiotemporal clustering algorithms results needs deep understanding of the occurring phenomena. The discovered pattern may be trivial with respect to the phenomenon under observation (Kisilevieh et al.2010). This issue can be addressed by tools that implements visualization techniques of spatiotemporal data and propose various methods of analysis, including trajectory clustering, generalization (Andrienko and Andrienko 2010), and aggregation (Andrienko and Andrienko 2008). These tools are usually application domain-specific and support various types of movement data (Andrienko et al. 2007). However, generic and state-of-the-art visualization and tools include GIS software, spatial database management systems, and spatial big data platforms (Shekhar et al. 2015). Visual analytics software provides an environment to control the computational process by providing input parameters, interpreting results, and directing the algorithm to obtain solutions that better describe the observed phenomena (Kisilevieh et al.2010). Rinzivillo et al. (2008) propose a progressive c1ustering approach supported by visualization and interaction techniques to analyze the movement behavior of objects. In this approach, a simple distance function is applied to each step to gain understanding of the underlying data.

Spatiotemporal clustering algorithms usually make use of spatial access methods to process very large databases (Ester et al. 1996). Therefore spatiotemporal data need to be stored into a spatial data base management system, which provides an indexing structure. Flat data sources lack the support of an indexing structure. An example of a popular commercial spatial database system is Oracle Spatial (Murray 2013). PostGIS (Postgres Spatial) (Obe and Hsu 2015) is the most popular open-source spatial database management system and is an extension to PostgresSQL.

6.2 Spatiotemporal Clustering Tools

The development and application of data mining algorithms including spatiotemporal clustering algorithms require the use of software tools. In the present era, a family of spatiotemporal clustering algorithms have been developed and integrated into existing open-source or commercial packages. This section provides an overview of free data mining tools available in the literature, in the context of spatiotemporal clustering.

Waikato Environment for Knowledge Analysis (WEKA) consists of open-source libraries, and its components have been integrated in several open-source tools such as RapidMiner and KNIME (Mikut and Reischl 2011). WEKA provides free access to the source code; as a result, many projects have incorporated or extended it (Hall et al. 2009). Alvares et al. (2010) propose a Weka-STPM (semantic trajectories pre-processing module), to support automatic raw trajectory data pre-processing to transform them into semantics trajectories. Weka-STPM is an extension of the Weka tool kit. The system adheres to open geospatial consortium (OGC) specifications, which makes it interoperable with all OCG-compliant databases. The authors have tested Weka-STPM with geographic databases and trajectory data stored in PostGIS.

R (The R Project for Statistical Computing) is a free software programming language and software environment for statistical computing and graphics. Space time point patterns (Stpp) is an R package for spatiotemporal clustering. Stpp encompasses several models needed in applications of point process methods to the study of spatiotemporal data.

Connections to flat external data sources are supported by Weka and RapidMiner. The flat data representation restricts the assessment of the index structure on the performance of data mining algorithms (Achtert et al. 2008). Achtert et al. (2008) develop a java-based framework **E**nvironment for Deve**L**oping **K**DD-Applications Supported by **I**ndex Structures (ELKI), which implements many popular algorithms in the field of clustering and especially subspace clustering. The implemented clustering algorithms include, among others, k-means

(Macqueen 1967), DBSCAN (Ester et al. 1996), OPTICS (Ankerst et al. 1999), and CLIQUE (Agrawal et al. 1998). The framework allows the incorporation of new spatiotemporal clustering algorithms, with a capability for a fair algorithmic evaluation. Bernárdez (2016) uses the ELKI framework for comparing spatiotemporal clustering approaches. ELKI comprises a set of data mining algorithms such as clustering, classification, outlier-detection, and item-set mining, and also has support for arbitrary index structures including large and high-dimensional data sets.

To provide distinctive definitions to different types of movement behavior, Baglioni et al. (2009) propose an ontology-based approach for the semantic explanation of trajectory patterns. The approach adopts an incremental process to enrich raw trajectories with perspective geo-information. The first step is the definition of a semantic trajectory as sequence of stops and moves. The next step makes use of the knowledge provided by the ontology to integrate contextual geo-information with the semantic trajectory patterns. Finally, a reasoning step allows inferring new knowledge (e.g., the inferred movement behavior) that can be added back to the ontology. To better understand the feasibility of the approach, the authors develop a system, called Athena (Baglioni et al. 2009), which allows a user to construct a query by means of the ontology concepts.

SaTScan (Kulldorff 2018) is a free software tool to analyze spatial, temporal, and spatiotemporal data using spatial, temporal, or space-time scan statistics. The tool is designed to detect spatial or spatiotemporal disease clusters and to evaluate the statistical significance of alarming clusters. The tool is helpful in geographical surveillance of disease and can also be used for similar problems in other fields.

A list of general characteristics of four spatiotemporal clustering tools is provided in Table 6.

To render this review of practical use, the next sub section provides information on data sets used by various spatiotemporal clustering algorithms available in the literature.

6.3 Spatiotemporal Datasets

A large number of spatiotemporal data sets have been collected in different domains. This section discusses the general characteristics of some of the data sets used in various spatiotemporal clustering approaches along with its potential applications.

Birant and Kut (2007) design an spatiotemporal data warehouse system containing environmental data about the Black Sea, the Marmara Sea, the Aegean Sea, and the east of the Mediterranean Sea. At various geographic locations, environmental data regarding the temperature and height of the surface of the sea, wave height, and sea winds are collected. The geographic coordinates of the work area include 30^{o} to 47.5^{o} north latitude and 17.0^{o} to 42.5^{o} east longitude.The temporal dimension can be grouped into year, month, and day. The ST-DBSCAN algorithm uses this data to evaluate their results.

Table 6. General characteristics of the free spatiotemporal clustering tools

Characteristic	**Weka**	**Elki**	**R**	**Athena**
Developer	Univ. of Waikato, New Zealand	Achtert et al. (2008)	Chambers (The R Project for Statistical Computing)	Baglioni et al. (2009)
Main purpose	General data mining	clustering	Statistical Computing	Semantic interpretation of trajectory patterns
Language	Java	Java	C , Fortran, and R	Ontology-based
Current version	3.9.0	0.7.1	3.3.1	---
Spatiotemporal clustering support	Weka-STPM	Yes	Stpp	Yes
License	GNU General Public License	AGPL	GNU General Public License	--

The spatiotemporal clustering algorithm, ST-OPTICS (Agrawal et al. 2016) uses a dataset provided by the Indian Space Research Organisation (ISRO). The spatial dimensions of dataset are latitude and longitude; the temporal dimensions are day, month and year; and the non-spatial dimension is normalized difference vegetation index (NDVI). The specifications of the data set are (i) Satellite: MODIS (Moderate Resolution Imaging Sepctroradiometer); (ii) Measure: NDVI (16 day composite); (iii) Area : Entire Country (India); (iv) Grid Size : 5×5 km^2 (v) Total No of Grids : 1,30,307 (includes different states and their districts).

The spatiotemporal clustering approach proposed by Chen et al. (2015) autonomously extracts all water bodies from 166 lake areas on a global scale. The approach makes use of SRTM Water Body Dataset (SWBD) (SRTM Water Body Dataset | The Long Term Archive). The dataset contains a majority of water bodies ranging from 60^o S to 60^o N. The SWBD data can be obtained through the publicly available MODIS repository as the MOD44W product (MOD44W | LP DAAC :: NASA Land Data Products and Services). The authors compare the output of the algorithms for the February 18th 2000 snapshot of TCWETNESS against the SWBD data.

The trajectory clustering algorithm, TRACLUS, developed by Lee et al. (2007) uses of two real data sets: hurricane track data (UNISYS, Atlantic Tropical Storm Tracking by Year), called Best Track and animal movement data. Best Track has the latitude, longitude, maximum sustained surface wind, and minimum sea level pressure of hurricanes at a frequency of 6-h. The authors use the Atlantic hurricanes from 1950 to 2004. It contains 570 trajectories and 17736 points. The animal movement data set was generated through the Starkey project (Y - U.S. Forest Service). It contains the radio-telemetry locations along with other information of elk, deer, and cattle from the years 1993 through 1996. The authors used Elk Movements in 1993 and Deer Movements in 1995. Elk1993 has 33 trajectories and 47204 points; Deer1995 has 32 trajectories and 20065 points.

The spatiotemporal algorithm proposed by Zaghlool et al. (2015) make use of Hurricane and Animal Movement Data Sets by adding time dimensions. The Hurricane Data Set contains Atlantic hurricanes from 1950 to 2006, and consists of 608 un-partitioned trajectories. Each trajectory consists of a set of points, where each point has x, y coordinates and time. The authors used Elk Movements in 1993 and Deer Movements in 1995 data sets of animal movement. The elk data sets contain 33 un-partitioned trajectories while the deer data set contains 32 un-partitioned trajectories.

As real data set was not available, Kalnis et al. (2005) make use of synthetic data set to evaluate proposed algorithms developed to identify moving clusters automatically. The authors develop a generator to synthetically generate data sets with various distributions. The input parameters of the generator include the number of clusters per time slice, average number of objects per cluster, neighborhood radius ε, density MinPts, average velocity of clusters, and change probability P_c. The output of the generator is a series of time slices. At each time slice, each cluster may move from its previous position; the velocity vector changes with probability P_c. With the same probability, a cluster may rotate around its center. The objects are inserted or deleted with probability P_c. The size of each data set ranges from 500K to 5M objects.

The seismic event clustering approach, SM-DBSCAN, proposed by Georgoulas et al. (2013), makes use of all the seismic events having a magnitude greater than 1 during the period from 2000 to 2010 in the Hellenic seismic arc region, to identify seismic zones. The region is the most seismologically active part of Europe.

CB-SMoT (Palma et al. 2008) uses trajectory data collected in Rio De Janerio, by the Traffic Engineering Company of Rio De Janerio. The dataset contains more than 2000 trajectories.

Table 7. General characteristics of the spatiotemporal data sets

Application	**Features**	**Data classification**	**Spatial dimensions**	**Temporal dimension**	**Number of points / size**	**Selected literature**
Marine environmental studies	NOAA-Series satellites data	Geo-referenced data item	Latitude and longitude	Days, months and year	Approximately 1.5 million rows (0.8 GB)	Birant and Kut (2007)
Study and analysis of forest and vegetation distribution	MODIS satellites data Non-spatial dimensions : NDVI	Geo-referenced data item	Latitude and longitude	Days, months and year,	1,30,307 grids Grid size: 5×5 km^2	Agrawal et al. (2016)
Identification of dynamic clusters in the presence of noise and missing data	SRTM Satellite Water Body Dataset, MOD44W product	Moving object	--	--	318 tiles each tile contains a 4800×4800 pixel grid for 250 m resolution	Chen et al. (2015)
Identification of seismic regions and seismic activities	Region : Hellenic seismic arc	Event	Latitude and longitude	5 min to 30 min	Seismic events having a magnitude greater than 1 during period 2000–2010	Georgoulas et al. (2013)
Discovery of common sub trajectories from trajectory database	Hurricane track data set	Trajectory	Latitude and longitude	6 hourly	570 trajectories , 17736 points	Lee et al. (2007)
	Animal movement data set	Trajectory	Universal Transverse Mercator Coordinates: easting- first six digits; northing- second seven digits.	hours	Elk Movements in 1993 : 33 trajectories and 47204 points; Deer Movements in 1995: 32 trajectories and 20065 points	
Clustering and analysis of spatiotemporal data	Hurricane track data set	Trajectory	latitude and longitude	6 hourly	608 un-partitioned trajectories	Zaghlool et al. (2015)
	Animal movement data set	Trajectory	Universal Transverse Mercator Coordinates: easting- first six digits; northing- second seven digits.	hours	Elk Movements in 1993 : 33 un-partitioned trajectories Deer Movements in 1995: 32 un-partitioned trajectories	
Tracking of convoys of vehicles, Identification of suspicious activities of individual's moving similarity, Identification of moving dense regions of traffic	Synthetic data sets were generated by inserting or deleting objects with a probability	Moving object			The size of each data set ranges from 500K to 5M objects	Kalnis et al. (2005)
Throughput planning o delivery trucks, Collective	-50 truck trajectories were measured for	Trajectory	Latitude and longitude	Seconds in case of cattle, Minutes for trucks, cars, and	-267 trajectories of trucks, 59894 points	Jeung et al. (2008)

transport Scheduling and carpooling	33 days, −13 trajectories were measured for several hours through GPS-enabled ear– tags every second. -183 private car trajectories were measured for one week. - 500 taxi trajectories were recorded in one day			taxis	-13 trajectories of cattle, 2283268 points. -183 trajectories of cars, 82590 points -500 trajectories of taxi, 41144 points	
Surveillance applications (Identification of suspicious behavior of people in the crowd, convoys of vehicles)	Four real and one synthetic datasets of trajectories have been used	Trajectory	Latitude and longitude	--	-276 truck trajectories, 112,203 points -145 bus trajectories, 66,096 points -183 car trajectories, 134,263 points -43 caribous trajectories, 15,796 points -synthetic data set: 50,000 trajectories, 2,548,084 points	Vieira et al. (2009)
Applications where speed plays a key role (traffic management)	Data collected in Rio De Janerio, by the Traffic Engineering company of Rio De Janerio	Trajectory	Latitude and longitude	hour	More than 2000 trajectories	Palma et al. (2008)
Automatic discovery of places for oceanic fishing	Real trajectories of fishing boats have been used	Trajectory	Latitude and longitude	30 minutes	Two trajectories	Rocha et al. (2010)

The algorithms for convoy discovery proposed by Jeung et al. (2008) use numerous real datasets of vehicles and animals. The different object types have different trajectory characteristics, such as sampling frequency of location and data distributions. The authors obtain 267 trajectories of 50 trucks, by measuring locations during 33 days in the Athens metropolitan area, Greece. The movements of 13 cattle are obtained by GPS-enabled ear-tags, which measure the locations every second for several hours. Jeung et al. (2008) also obtain 183 car trajectories from one week. The car trajectory data set has very different lengths. The authors also use the trajectories of 500 taxis, which are recorded by GPS-enabled devices during the day. The locations of the trajectories are sampled irregularly.

The methods proposed by Vieira et al. (2009) to discover flock patterns, uses four real and one synthetic dataset of trajectory. The Truck dataset contains 112,203 locations from

276 trucks and the Buses dataset contains 66,096 locations generated from 145 buses. Both datasets are collected in Athens, Greece. The Cars dataset comprises 134,263 locations generated from 183 private cars in Copenhagen, Denmark. The Caribous dataset is the outcome of an analysis of the migration of 43 caribous in the north-western provinces of Canada. The dataset contains 15,796 locations. The synthetic dataset SG is created by simulating the movement of 50,000 vehicles on the road network of Singapore. The moving objects have different velocities, and their starting locations are randomly placed on the road network. The synthetic dataset comprises 2,548,084 object locations.

DB-SMoT (Rocha et al. 2010) uses real trajectories of fishing boats. Each trajectory contains information such as the fishing period, average fishing time for settings and haulings, number of settings and hauling, and the amount of captured fish. The GPS location is collected every 30 minutes in the trajectories.

The general characteristics of the data sets used in various spatiotemporal clustering approaches along with their potential applications are summarized in the Table 7.

The good sources of spatiotemporal data sets are Plenar.io (Plenar.io-A spatio-temporal open data platform), New York City taxi dataset (NYC Taxi & Limousine Commission) and Google Dataset Search (Dataset Search – Google).

As many researchers use different data sets in their research papers, we believe that the researchers should agree on some benchmark spatiotemporal data sets so that it will be easy to compare different spatiotemporal clustering algorithms.

The following section provides an overview of applications of spatiotemporal clustering approaches.

7. Applications

This section provides an overview of applications of spatiotemporal clustering approaches in many areas of engineering, scientific, and real-world applications. Spatiotemporal clustering techniques can be applied in numerous fields such as image processing and patterns recognition (Li et al. 2004), molecular biology (Sander et al.1998), environmental studies (Birant and Kut 2007), earth quake studies (Georgoulas et al. 2013; Wang et al.2006), transportation (Jeung et al. 2008; Spaccapietra et al. 2008), traffic management (Anbaroglu et al. 2015; Spaccapietra et al. 2008; Palma et al. 2008), GIS (Sander et al.1998), identification of terrorist activities (Kalyani and Chaturvedi 2012), surveillance (Vieira et al. 2009), mobility data analysis (Bogorny and Shashi 2010), trajectory outlier detection (Bogorny and Shashi 2010), pattern identification of animal behavior (Kalnis et al. 2005 ;Vieira et al. 2009), tracking of vehicle convoys (Kalnis et al. 2005), and automatic discovery of fishing spots (Rocha et al. 2010).

In the field of GIS, spatiotemporal clustering plays a vital role. A clustering algorithm is used to cluster areas of geographic entities by using the concept of geometric space similarity (Manjula and Narsimha 2014). Spatiotemporal movement patterns are useful in traffic management, traffic flow, and animal tracking. These approaches coupled with GIS are extremely helpful for traffic analysis and management.

Alatrista-Salas et al. (2015) applied knowledge discovery process to hydrological data to monitor river water quality. The spatiotemporal clustering techniques are applied to group sequential patterns mined using a water course approach. GIS have been used as a post processing step in the KDD process.

Birant and Kut (2007) develop and apply the spatiotemporal clustering algorithm, ST-DBSCAN, for marine environmental studies. The applications of the algorithm include the discovery of similar sea water characteristics regions, cyclone detection by extracting locations of cyclones and tracks followed by cyclones.

Agrawal et al. (2016) develop, apply, and validate an enhanced spatiotemporal clustering algorithm, ST-OPTICS, for the study and analysis of forest and vegetation distribution. The authors use the data of different areas from different states of India to apply and validate the algorithm using selected performance indices.

The application of spatiotemporal clustering in the field of seismic studies includes the identification of seismic regions and seismic activities. The observed earth quake epicenters could be clustered along continent faults. Wang et al. (2006) suggest spatiotemporal clustering algorithms to analyze sequences of seismic events. Georgoulas et al. (2013) develop and apply the spatiotemporal clustering algorithm, SM-DBSCAN, to seismic event clustering.

The objectives of situational awareness and risk management in cases of natural disaster could be achieved by analyzing the spatiotemporal impacts of emerging events. Lee (2012) apply density-based online clustering algorithms for mining micro-blogging text streams, to obtain real-time geospatial event information

Han et al. (2016) apply an unsupervised crowd activity discovery algorithm to automatically mine and cluster visual patterns in a large number of crowd activities. The algorithm employs spatiotemporal saliency strategy and the Normalized Cut clustering algorithm.

Kalnis et al. (2005) apply exact and approximate algorithms in various fields, to identify moving clusters automatically. In the field of ecology, the automatic discovery of moving clusters helps to study the evolution of moving group of animals. The tracking of vehicle convoys could be another application. Military applications of automatic discovery of moving clusters include monitoring of movements of troops in space and time. To assist intelligence and counterterrorism agencies, the approach could be employed to identify suspicious activities of individual’s movement similarity. Traffic surveillance systems can benefit from these techniques to identify moving dense areas of traffic.

Jeung et al. (2008) propose convoy discovery algorithms. The real-life applications of convoy discovery include throughput planning of delivery trucks, scheduling of collective transport, and identification of cars for carpooling.

Vieira et al. (2009) propose strategies to discover flock patterns in streaming spatiotemporal data. Such patterns are useful to study patterns of animal behavior including the migration of birds, whales, sharks, etc. Flock patterns may be useful in surveillance applications, for example the identification of suspicious behavior of certain people within a large number of people, convoys of vehicles, etc.

The CB-SMoT approach (Palma et al. 2008) has potential applications where speed plays a key role, such as traffic management. The DB-SMoT approach (Rocha et al. 2010), has been applied to real trajectory data associated with oceanic fishing vessels, with the goal to automatically discover the real places where vessels develop fishing activities.

The stops and moves model proposed by Spaccapietra et al. (2008) has several applications, such as bird migration, where stops could be the places where birds feed or rest; traffic management, where stops could be red lights, traffic jams, speed controllers, etc.; military mobilization applications, where stops could be planned halt locations and engagements; and business travel of employees to conduct business with distant customers.

Hu et al. (2013) propose an spatiotemporal topic (STT) model to address the problem of recommending the right locations to users at the right time. The STT model captures the spatiotemporal aspects of user check-ins and use profiles in a single probabilistic model for location recommendation.

A summary of spatiotemporal clustering approaches, their applications, and the selected literature is provided in Table 8.

Table 8. Overview of spatiotemporal clustering approaches and applications

Category	**Clustering approach**	**Application**	**Selected literature**
Environmental	Spatiotemporal clustering technique	Monitoring of river water quality	Alatrista-Salas et al. (2015)
	ST-GRID	Analysis of sequences of seismic events	Wang et al. (2006)
	ST-DBSCAN (based on DBSCAN)	Marine environmental studies	Birant and Kut (2007)
	ST-OPTICS (based on OPTICS)	Study and analysis of forest and vegetation distribution	Agrawal et al. (2016)
	SM-DBSCAN (based on DBSCAN)	Identification of seismic regions and seismic activities	Georgoulas et al. (2013)
	Extension of DBSCAN	Identification of dynamic clusters in the presence of noise and missing data to monitor land water bodies	Chen et al. (2015)
	TRACLUS (based on DBSCAN)	Discovery of common sub trajectories from trajectory database (hurricane track data)	Lee et al. (2007)
	Extension of TRACLUS	Clustering and analysis of spatiotemporal data (hurricane track data)	Zaghlool et al. (2015)
	DB-SMoT (Based on DBSCAN)	Automatic discovery of places for oceanic fishing	Rocha et al. (2010)
Traffic management	Space-time Scan Statistics	Non-recurrent traffic congestion detection on heterogeneous urban road networks	Anbaroglu et al. (2015)
	Based on DBSCAN	Tracking of convoy of vehicles, monitoring of troop movements with respect to space and time in military-domain, identification of suspicious activities of individual's movement similarity, identification of moving dense areas of traffic	Kalnis et al. (2005)
	Stops and Moves Model	Bird migration, traffic management, military mobilization application	Spaccapietra et al. (2008)
	CB-SMoT (Based on DBSCAN)	Applications where speed plays a key role (traffic management)	Palma et al. (2008)
	Based on DBSCAN	Throughput planning of delivery trucks, collective transport scheduling and carpooling	Jeung et al. (2008)
Surveillance	Methods to discover flock patterns based on DBSCAN	Surveillance applications (Identification of suspicious behavior of some people in the crowd, convoys of vehicles)	Vieira et al. (2009)
	Normalized Cut (graph theory-based clustering algorithm)	Discovery of visual patterns in crowd activities	Han et al. (2016)
Disaster management	Density-based online clustering approach	Situational awareness and risk management in case of natural disaster	Lee (2012)

The following section highlights various issues and challenges in the context of spatiotemporal clustering, which includes the acquisition of correct data, development of efficient approaches, large-scale analysis of available data, and development of testing and validation strategies.

8. Issues and Challenges

In this section, we discuss the issues and challenges in handling spatiotemporal data and hence performing clustering. They range from the acquisition of correct data to the analysis, development of efficient approaches, development of testing and validation strategies, availability of good clustering indices, lack of benchmark data sets, and challenges posed by big data.

Affordable sensors and technological infrastructure have made fine-grained geospatial data available at small time intervals. Spatiotemporal data characteristics further adds complexity to the data mining process in general and to the clustering in particular. The complexity of spatiotemporal data types, data representation, and data structures introduce new challenges to spatiotemporal data mining (Yao 2003).

Spatial and temporal resolution plays a vital role in spatiotemporal clustering. A change in resolution may lead to an inappropriate interpretation of the phenomena being investigated. The literature does not provide any guidelines for the selection of spatial and temporal resolution; however, ad hoc approaches are proposed to handle the issue in a specific domain. This necessitates the involvement of domain experts in the spatiotemporal clustering process (Kisilevieh et al.2010). In addition, to overcome the limitations of the automatic process, the involvement of a domain expert is essential for data inspection, parameter tuning, controlling the analysis process, and revising the obtained results (Shekhar et al. 2015; Compieta et al. 2007). The integration of domain knowledge with spatiotemporal clustering is an important research area as it has capability to improve clustering accuracy and cluster interpretation. To achieve this objective, new spatiotemporal clustering algorithms need to be developed by making use of domain knowledge.

GIS technology coupled with spatiotemporal clustering can aid the domain expert with visualization, analysis, and decision making. The distance plays a vital role in spatiotemporal clustering algorithms. The distance function in GIS is not a simple Euclidean distance but depends on map projections, datum, and physical constraints. The challenge in the integration of GIS with spatiotemporal clustering algorithms is to define the distance function considering these criteria (Wang and Wang 2010). Spatiotemporal distance functions need to be formulated considering aforementioned criteria, so that GIS can be integrated with spatiotemporal clustering algorithms.

The data quality often varies in time in the case of many real-world applications (Chen et al. 2015). The quality of spatiotemporal data also depends on the data generation process (Miller and Han 2005). Heterogeneous data sources lead to unexpected errors and data conflicts (Yao 2003). The conflict reconciling strategies for spatiotemporal data need to be developed or the existing strategies need modification. As a pre-processing step, noise, outliers, missing values, duplicate data, or inaccurate parts of data need to be handled, combined with an appropriate handling strategy (i.e., to ignore, interpolate, or tolerate) (Miller and Han 2005) after that the existing spatiotemporal clustering methods can be employed. Alternative approach is to develop spatiotemporal clustering algorithms that have capability to address the problem of varying data quality. Another challenge is heterogeneity in space and time (Faghmous and Kumar 2014). In spatial heterogeneity, the data belonging to different clusters may have the same feature values. Temporal heterogeneity refers to the case where the feature values that uniquely discriminate a cluster change over time for the same cluster. To address the heterogeneity in space and time, spatiotemporal clustering algorithms need to be developed.

The pace of computational power does not match the pace of generation and storage of data. The big spatiotemporal data demands scalable and efficient clustering algorithms. The parallelization of spatiotemporal clustering algorithm is an approach (Husch et al. 2018) that allows scalability with increasing data size and maintains low time complexity. To meet the big spatiotemporal data challenges, the research in this direction is required.

In an application scenario such as a traffic jam, objects can interact and interfere with each other. As a result, proper analysis of mobile objects requires a population of objects along with geographical features and physical constraints. The challenge is the automatic detection of such interaction and incorporation in the spatiotemporal clustering process (Kisilevieh et al.2010). To evaluate the quality of clusters, many performance metrics are available. A major issue is the choice of the performance metric. As spatiotemporal clustering is unsupervised form of machine learning, therefore cluster labels should not be present. However, in some cases, true cluster label may be known. In such cases the performance metric such as accuracy, F-score, and normalized mutual information can be used. In real world scenario, where true cluster labels are not present, the performance metrics such as

Davies-Bouldin index, Dunn index, and Silhouette index (Rendón et al. 2011) are the choices. The spatiotemporal clustering algorithms usually produce dense and arbitrarily shaped clusters (Agrawal et al. 2016). Few researchers have worked to validate dense and arbitrary shaped clusters generated from spatiotemporal data. More research is required to identify suitable performance indices for dense an arbitrary shaped spatiotemporal clusters.

Several spatiotemporal clustering algorithms demand a number of parameter settings by the user, which influences the results of the clustering procedure largely; therefore, it is difficult to identify appropriate input parameters to produce high-quality clusters for a dataset. Moreover, the identification of the clustering technique and accurate number of clusters for a dataset is also difficult. Some parameter insensitive spatiotemporal clustering algorithms have been developed (Auria et al. 2006; Agrawal et al. 2016). More research in this field is desirable.

The lack of availability of benchmark spatiotemporal datasets poses a challenge to compare existing clustering algorithms and evaluate new algorithms. The creation of benchmark spatiotemporal dataset repositories would be helpful to compare existing spatiotemporal clustering algorithms and motivate the development of new spatiotemporal clustering algorithms.

9. Conclusions

This paper provides a comprehensive overview of spatiotemporal clustering research by reviewing spatiotemporal data in the context of the spatiotemporal domain. It summarizes space-time scan statistics, spatiotemporal patterns, and the approaches adopted for their extraction, along with free data mining tools, as well as generic and application-specific spatiotemporal analysis tools. It establishes that spatiotemporal clustering has broad applications, including image processing, pattern recognition, molecular biology, environmental and seismology studies, transportation and traffic management, GIS, surveillance, and mobility data analysis. The general characteristics of data sets used in various spatiotemporal clustering approaches and their potential applications have been mentioned. Finally, the pervasive issues and challenges, encompassing data acquisition to data analysis, efficient approaches, testing and validation of strategies, availability of good clustering indices, lack of benchmark data sets, and challenges posed by big data have been discussed in the context of spatiotemporal clustering.

Acknowledgment

This work is supported by Ministry of Electronics and Information Technology, Government of India under the Visvesvaraya Ph.D. scheme.

References

Achtert E, Kriegel H, Zimek A (2008) ELKI: a software system for evaluation of subspace clustering algorithms. *Scientific and Statistical Database management* (Ssdbm), 580–585.

Agrawal KP, Garg S, Sharma S, Patel P (2016) Development and validation of OPTICS based spatio-temporal clustering technique. Inf Sci (Ny) 369:388–401. doi: 10.1016/j.ins.2016.06.048

Agrawal R, Gehrke J, Gunopulos D, Raghavan P (1998) *Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications*. Proc. ACM SIGMOD’98 Int. Conf. on Management of Data, Seattle, WA, 94-105.

Ahmad A, Dey L (2007) A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl Eng 63:503–527. doi: 10.1016/j.datak.2007.03.016

Alatrista-Salas H, Azé J, Bringay S, et al (2015) A knowledge discovery process for spatiotemporal data: Application to river water quality monitoring. Ecol Inform 26:127–139. doi: 10.1016/j.ecoinf.2014.05.011

Alon J, Sclaroff S, Kollios G, Pavlovic V (2003) Discovering clusters in motion time-series data. 2003 IEEE Comput Soc Conf Comput Vis Pattern Recognition, 2003 Proceedings 1:I--375. doi: 10.1109/CVPR.2003.1211378

Alvares LO, Palma a. L, Oliveira G, Bogorny V (2010) Weka-STPM: from trajectory samples to semantic trajectories. Proc Work Open Source Code 1:1–6.

Alvares LO, Loy AM, Renso C, Bogorny V (2011) An algorithm to identify avoidance behavior in moving object trajectories. J Brazilian Comput Soc 17:193–203. doi: 10.1007/s13173-011-0037-3

Anbaroglu B, Cheng T, Heydecker B (2015) Non-recurrent traffic congestion detection on heterogeneous urban road networks. Transportmetrica A: Transport Science 11: 754–771. http://dx.doi.org/10.1080/23249935.2015.1087229

Andrienko G, Andrienko N (2010) Interactive cluster analysis of diverse types of spatiotemporal data. ACM SIGKDD Explor Newsl 11:19-28 doi: 10.1145/1809400.1809405

Andrienko G, Andrienko N (2008) Spatio-temporal Aggregation for Visual Analysis of Movements. Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium

Andrienko G, Andrienko N, Wrobel S, Augustin S (2007) Visual analytics tools for analysis of movement data. ACM SIGKDD Explor Newsl 9:38–46. doi: 10.1145/1345448.1345455

Ankerst M, Breunig MM, Kriegel H, Sander J (1999) OPTICS : Ordering Points To Identify the Clustering Structure. SIGMOD '99 Proc 1999 ACM SIGMOD Int Conf Manag data 28:49–60. doi: 10.1145/304182.304187

Arbelaitz O, Gurrutxaga I, Muguerza J, et al (2013) An extensive comparative study of cluster validity indices. Pattern Recognit 46:243–256. doi: 10.1016/j.patcog.2012.07.021

Auria M, Nanni M, Pedreschi D (2006) Time-focused dentisty-based clustering of trajectories of moving objects. JIIS Spacial Issue on Mining Spatio-Tamporal Data

Baglioni M, de Macêdo JAF, Renso C, Trasarti R, Wachowicz M (2009). Towards semantic interpretation of movement behavior. In *Advances in GIScience* :. 271-288.

Ball GH, Hall DJ (1965) ISODATA, a novel method of data analysis and pattern classification. Menlo Park, Calif, Stanford Research Institute.

Becher J, Berkhin P, Freeman E (2000) Automating exploratory data analysis for efficient data mining. In *Proceedings of the 6th ACM SIGKDD*, 424-429

Beckmann N, Begel H-P, Schneider R, Seeger B (1990) The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. *ACM SIGMOD Record*. 19(2): 322–331.

Bentley JL (1975) Multidimensional Binary Search Trees Used for Associative Searching. *Communications of the ACM 18* , 509-517.

Berkhin P (2006) Survey of Clustering Data Mining Techniques. Group Multidimens Data Recent Adv Clust 25–71. doi: 10.1007/3-540-28349-8_2

Bernárdez, F. D. P. (2016) "Extraction of user's stay and transitions from GPS logs: a comparison of three spatiotemporal clustering approaches."

Bezdek JC, Trivedi M, Ehrlich R, Full W (1982) Fuzzy clustering; a new approach for geostatistical analysis. International Journal of Systems, Measurement and Decision. Doi: 10.1016/0098-3004(84)90020-7

Birant D, Kut A (2007) ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data Knowl Eng 60:208–221. doi: 10.1016/j.datak.2006.01.013

Bogorny V, Shashi S (2010). Spatial and spatio-temporal data mining. *2010 IEEE International Conference on Data Mining*.

Calinski T, Harabasz J (1974) A Dendrite Method for Cluster Analysis. Communications in Statistics Theory and Methods 3(1):1-27.

Campello RJGB, Moulavi D, Sander J (2013) Density-Based Clustering Based on Hierarchical Density Estimates. Pei J., Tseng V.S., Cao L., Motoda H., Xu G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 7819:160-172.

Cao H, Mamoulis N, Cheung DW (2005) Mining frequent spatio-temporal sequential patterns. In *ICDM: Proc. Fifth IEEE International Conference on DataMining* (2005), IEEE Computer Society Press :82–89. doi:10.1109/icdm.2005.95.

Cai M, Revesz PZ (2000) Parametric R-Tree: An Index Structure for Moving Objects. Proc. of the COMAD Conf.

Chen XC, Faghmous JH, Khandelwal A, Kumar V (2015) Clustering dynamic spatio-temporal patterns in the presence of noise and missing data. IJCAI Int Jt Conf Artif Intell 2015–Janua:2575–2581

Chudova D, Gaffney S, Mjolsness E, Smyth P (2003) Translation-invariant mixture models for curve clustering. Proc ninth ACM SIGKDD Int Conf Knowl Discov data Min - KDD '03 79. doi: 10.1145/956750.956763

Clementini E, Felice PD, Oosterom PV (1993) A Small Set of Formal Topological Relationships Suitable for End-User Interaction. Advances in Spatial Databases.

Compieta P, Di Martino S, Bertolotto M, Ferrucci F, Kechadi T (2007) Exploratory spatio-temporal data mining and visualization. Journal of Visual Languages and Computing 18: 255–279

Dataset Search – Google https://toolbox.google.com/datasetsearch. Accessed on 05 Feb 2019

De Lucca Siqueira F, Bogorny V (2011) Discovering Chasing Behavior in Moving Object Trajectories. Trans GIS 15:667–688. doi: 10.1111/j.1467-9671.2011.01285.x

Doborjeh MG, Kasabov N, Doborjeh Z G (2018) Evolving, dynamic clustering of spatio/spectro-temporal data in 3D spiking neural network models and a case study on EEG data. Evolving systems 9(3):195-211.

Doborjeh M G, Kasabov N (2015) Dynamic 3D clustering of spatio-temporal brain data in the NeuCube spiking neural network architecture on a case study of fMRI data. In International Conference on Neural Information Processing, 191-198. Springer, Cham. DOI: 10.1007/978-3-319-26561-2_23

Dunn JC (1974) Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4:95-105.

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of Second International Conference on Knowledge Discovery and Data Mining , 226–231

Faghmous JH, Kumar V (2014) A big data guide to understanding climate change: The case for theory-guided data science. Big data 2(3):155–163. doi: 10.1089/big.2014.0026

Fiori A, Mignone A, Rospo G (2016) DeCoClu: Density consensus clustering approach for public transport data. Information Sciences 328: 378–388.

Gaffney S, Smyth P (1999) Trajectory clustering with mixtures of regression models. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining KDD 99 10:63–72. doi: 10.1145/312129.312198

Georgoulas G, Konstantaras A, Katsifarakis E, et al (2013) “Seismic-mass” density-based algorithm for spatio-temporal clustering. Expert Systems with Applications 40:4183–4189. doi: 10.1016/j.eswa.2013.01.028

Giannotti F, Kujpers B, Raffaeta A, Manco G, Baglioni M, Renso C (2008) Querying and reasoning for spatio-temporal data mining. In [23] chapter 12, page 335, 2008.

Giannotti F, Nanni M, Pinelli F, Pedreschi D (2007) Trajectory pattern mining. InProc. ACM KDD, 330–339. https://doi.org/10.1145/1281192.1281230

Giannotti F, Pedreschi D (2008) *Mobility, data mining and privacy: Geographic knowledge discovery*. Springer Science & Business Media.

Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of Things (IoT): A vision, architectural elements, and future directions. Futur Gener Comput Syst 29:1645–1660. doi: 10.1016/j.future.2013.01.010

Güting RH (1994) An Introduction to Spatial Database Systems. VLDB J *3*(4):357-399. doi: 10.1007/BF01231602

Guttman A (1984) R-trees: a dynamic index structure for spatial searching. Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data - SIGMOD ’84, 47-57. https://doi.org/10.1145/602259.602266

Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. J Intell Inf Syst 17:107–145. doi: 10.1023/A:1012801612483

Hall M, Frank E, Holmes G, et al (2009) The WEKA data mining software. ACM SIGKDD Explorations 11:10–18. doi: 10.1145/1656274.1656278

Han T, Yao H, Sun X, et al (2016) Unsupervised discovery of crowd activities by saliency-based clustering. Neurocomputing 171:347–361. doi: 10.1016/j.neucom.2015.06.048

Henrich A, Six H-W, Widmayer P (1989). The LSD-Tree: Spatial Access to Multidimensional Point- and Non-Point-Objects. Proc. 15th Intl. Conf. on Very Large Data Bases, Amsterdam, 45-53.

Higgs B, Abbas M (2015) Segmentation and clustering of car-following behavior: Recognition of driving patterns. IEEE Transactions on Intelligent Transportation Systems 16(1):81-90.

Hu B, Jamali M, Ester M (2013) Spatio-temporal topic modeling in mobile social media for location recommendation. Proceedings - IEEE International Conference on Data Mining, ICDM 1073–1078. doi: 10.1109/ICDM.2013.139

Huang W, Xu S, Yan Y, Zipf A (2019) An exploration of the interaction between urban human activities and daily traffic conditions: A case study of Toronto, Canada. Cities 84:8-22. https://doi.org/10.1016/j.cities.2018.07.001

Hudjimartsu SA, Djatna T, Ambarwari A, Apriliantona (2018) Spatial temporal clustering for hotspot using kulldorff scan statistic method (KSS): A case in Riau Province. IOP Conf. Ser.: Earth Environ. Sci. 54 012056. http://iopscience.iop.org/article/10.1088/1755-1315/54/1/012056

Husch M, Schyska BU, Bremen LV (2018) CorClustST - Correlation-based Clustering of Big Spatio-Temporal Datasets. Future Generation Computer Systems (In Press). https://doi.org/10.1016/j.future.2018.04.002

Hwang SY, Liu YH, Chiu JK, Lim EP (2005) Mining mobile group patterns: A trajectory-based approach. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 713-718). Springer Berlin Heidelberg.

Izakian H, Pedrycz W (2013) Anomaly Detection in Time Series Data using a Fuzzy C-Means Clustering. . *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint*. IEEE 1513–1518.

Izakian H, Pedrycz W, Jamal I (2013) Clustering spatiotemporal data: An augmented fuzzy c-means. *IEEE Transactions on Fuzzy Systems* 21: 855-868.

Izakian H, Pedrycz W, Jamal I (2015) Fuzzy clustering of time series data using dynamic time warping distance. Eng Appl Artif Intell 39:235–244. doi: 10.1016/j.engappai.2014.12.015

Jeung H, Yiu ML, Zhou X, et al (2008) Discovery of convoys in trajectory databases. Proceedings of the VLDB Endowment 1:1068–1080. doi: 10.14778/1453856.1453971

Kalnis P, Mamoulis N, Bakiras S (2005) On Discovering Moving Clusters in Spatio-temporal Data. Advances in Spatial and Temporal Databases SE - 21 3633:364–381. doi: 10.1007/11535331_21

Kalyani D, Chaturvedi SK (2012) A Survey on Spatio-Temporal Data Mining. International Journal of Computer Science and Network, 1(4).

Kang JH, Welbourne W, Stewart B, Borriello G (2005) Extracting places from traces of locations. ACM SIGMOBILE Mobile Computing and Communications Review 9:58. doi: 10.1145/1094549.1094558

Kang J, Yong H (2009) Mining Trajectory Patterns by Incorporating Temporal Properties. ". Proceedings of the 1st international conference on emerging databases. 1-6

Kirana AP, Sitanggang IS, Syaufina L(2016) Hotspot pattern distribution in peat land area in Sumatera based on spatio temporal clustering," The 2nd International Symposium on LAPAN-IPB Satellite for Food Security and Environmental Monitoring 2015, LISAT-FSEM 2015, 33:635–645.

Kisilevich S, Keim D, Rokach L (2010) A novel approach to mining travel sequences using collections of geo-tagged photos. In: The 13th AGILE International Conference on Geographic Information Science. 163–182.

Kisilevieh S, Mansmann F, Nanni M, Rinzivillo S (2010) Spatio-temporal c1ustering. In Data mining and Knowledge Discovery Handbook. 855–874.

Kovács F, Legány C, Babos A (2005) Cluster Validity Measurement Techniques. Proc 6th Int Symp Hungarian Res Comput Intell 2006:1–11. doi: 10.7547/87507315-91-9-465

Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: Detection and inference. Statistics in medicine, 14:799-810. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780140809

Kulldorff M (1997) A spatial scan statistic. Commun. Stat.-Theory Methods, 26:1481–1496.

Kulldorff M (2018) SaTScan User Guide for Version 9.6; Available online: https://www.satscan.org (accessed on 20 June 2018)

Larose D T (2005) Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons, Inc.

Lee CH (2012) Mining spatio-temporal information on microblogging streams using a density-based online clustering method. Expert Systems with Applications 39:9623–9641. doi: 10.1016/j.eswa.2012.02.136

Lee J, Han J, Whang K-Y (2007) Trajectory clustering: a partition-and-group framework. Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD '07 593. doi: 10.1145/1247480.1247546

Leipnik MR, Albert DP (2002) GIS in Law Enforcement: Implementation Issues and Case Studies;CRC Press: Sacramento, CA, USA.

Li Y, Han J, Yang J (2004) Clustering moving objects. Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04 617–622. doi: 10.1145/1014052.1014129

Liu J, Xue C, He Y, Dong Q, Kong F, Hong Y (2018) Dual-Constraint Spatiotemporal Clustering Approach for Exploring Marine Anomaly Patterns Using Remote Sensing Products. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11(11); 3963-3976. DOI: 10.1109/JSTARS.2018.2873216

Liu Y, Li Z, Xiong H, Gao X, Gao J (2010) Understanding of Internal Clustering Validation Measures. Proceedings of the 2010 IEEE International Conference on Data Mining :911-916. DOI 10.1109/ICDM.2010.35

Maciag PS (2017) A Survey on Data Mining Methods for Clustering Complex Spatiotemporal Data. International Conference: Beyond Databases, Architectures and Structures 115–126. doi: 10.1007/978-3-319-58274-0

Macqueen J (1967) Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1:281–297. doi: citeulike-article-id:6083430

Manjula, A, Narsimha G (2014) A review on spatial data mining methods and applications. International Journal of Computer Engineering and Applications, VII: 208–218

Martino FD, Pedrycz W, Sessa S (2018) Spatiotemporal extended fuzzy C-means clustering algorithm for hotspots detection and prediction. Fuzzy Sets and Systems 340:109–126.

Mazimpaka JD, Timpf S (2016) Trajectory data mining: A review of methods and applications. Journal of Spatial Information Science 13:61-99. doi:10.5311/JOSIS.2016.13.263

Mikut R, Reischl M (2011) Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1:431–443. doi: 10.1002/widm.24

Milligan GW (1981) A monte carlo study of thirty internal criterion measures for cluster analysis. Psychometrika 46:187–199. doi: 10.1007/BF02293899

Miller HJ, Han J (2005). Geographic data mining and knowledge discovery. Vol. 2. CRC Press.

MOD44W | LP DAAC :: NASA Land Data Products and Services https://lpdaac.usgs.gov › MODIS › MODIS Products Table. *Accessed 22 Dec 2016*

Murray C (2013) Oracle Spatial Developer's Guide, 11*g* Release 2 (11.2) E11830-15.

Neill DB (2006) Detection of Spatial and Spatio-Temporal Clusters. *PhD Thesis*. School of Computer Science,Carnegie Mellon University, Pittsburgh. https://www.cs.cmu.edu/~neill/papers/thesis-final.pdf

NYC Taxi & Limousine Commission http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml. *Accessed 21 June 2018*

Obe RO, Hsu LS (2015) *PostGIS in Action*. Second Edition, Manning Publications Co., Shelter Island, NY 11964

Otair M (2013) Approximate k-nearest neighbor based spatial clustering using kd- tree. International Journal of Database Management Systems (IJDMS) Vol.5, No.1, 97-108.

Palma AT, Bogorny V, Kuijpers B, Alvares LO (2008). A clustering based approach for discovering interesting places in trajectories," in ACMSAC. New York, NY, USA: ACM Press, 863–868.

Pearson K (1895) Note on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London Series I 58: 240–242.

Pelekis N, Kopanakis I, Marketos G, et al (2007) Similarity search in Trajectory Databases. Proceedings of the International Workshop on Temporal Representation and Reasoning 129–140. doi: 10.1109/TIME.2007.59

Plenar.io-A spatio-temporal open data platform http://plenar.io. Accessed on 05 Feb 2019

Rendón E, Abundez I, Arizmendi A,Quiroz EM (2011) Internal Versus External Cluster Validation Indexes. International Journal of Computers and Communications 5(1):27–34.

Rinzivillo S, Pedreschi D, Nanni M, et al (2008) Visually Driven Analysis of Movement Data by Progressive Clustering. Information Visualization 7:225–239. doi: 10.1057/palgrave.ivs.9500183

Rocha JAMR, Times VC, Oliveira G, et al (2010) DB-SMoT: A direction-based spatio-temporal clustering method. 2010 5th IEEE International Conference Intelligent Systems 114–119. doi: 10.1109/IS.2010.5548396

Salazar G EJ, Velez AC, Parra M CM, Ortega L O (2002) A Cluster Validity Index for Comparing Non-hierarchical Clustering Methods, EITI, 2002, pp. 1-5.

Samet H (1985). Using quadtree to represent spatial data. *NATO ASI Series*, vol. F18, 229-247.

Sander J, Ester M, Kriegel HPP, Xu X (1998) Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. Data Mining and Knowledge *Discovery* 2.2:169–194. doi: 10.1023/A:1009745219419

Sardadi MM, Shafry M, Jupri Z, Daman D (2008) Choosing R-tree or Quadtree Spatial Data Indexing in One Oracle Spatial Database System to Make Faster Showing Geographical Map in Mobile Geographical Information System Technology, World Academy of Science, Engineering and Technology, 249-257.

Sellis TK, Roussopoulos N, Faloutsos C (1987) The R+-Tree: A Dynamic Index for Multi-Dimensional Objects. Vldb 507–518.

Shekhar S, Jiang Z, Ali RY, et al (2015) Spatiotemporal data mining: A computational perspective. ISPRS International Journal of Geo-Information 4:2306–2338. doi: 10.3390/ijgi4042306

Shekhar S, Varsavai RR, Celik M (2008) Spatial and spatiotemporal data mining: Recent advances. Data Mining: Next Generation Challenges and Future Directions

Spaccapietra S, Parent C, Damiani ML, De Macedo JA, Porto F, Vangenot C (2008) A Conceptual View on Trajectories. Data and Knowledge Engineering vol 65:126–146.

SRTM Water Body Dataset | The Long Term Archive https://lta.cr.usgs.gov/srtm_water_body_dataset . Accessed 14 March 2017

The R Project for Statistical Computing. https://www.r-project.org. Accessed 15 Feb 2017

Tork HF (2012) Spatio-Temporal Clustering Methods Classification. Doctoral Symposium on Informatics Engineering (DSIE’2012) 1–12. doi: 10.13140/RG.2.1.3812.7204

UNISYS, Atlantic Tropical Storm Tracking by Year. http://weather.unisys.com/hurricane/atlantic/. Accessed 11 March 2017

Upton G, Fingleton B (1985) Spatial data analysis by example volume1:Point pattern and quantitative data, Wiley, New York.

Vieira MR, Bakalov P, Tsotras VJ (2009) On-line discovery of flock patterns in spatio-temporal data. Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS ’09 286. doi: 10.1145/1653771.1653812

Wang M., Wang A, Li A (2006, August) Mining spatial-temporal clusters from geo-databases. In *International Conference on Advanced Data Mining and Applications* (pp. 263-270). Springer Berlin Heidelberg.

Wang X, Wang J (2010). Using Clustering Methods in Geospatial Information Systems. GEOMATICA 64: 347- 361.

Weingessel A, Dimitriadou E, Dolničar S (1999) An examination of indexes for determining the number of clusters in binary data sets. Psychometrika 67:1–21. doi: 10.1007/BF02294713

Yao X (2003). Research issues in spatio-temporal data mining. A white paper submitted to the University Consortium for Geographic Information Science (UCGIS) workshop on Geospatial Visualization and Knowledge Discovery, Lansdowne, Virginia, Nov. 18-20.

Y - U.S. Forest Service https://www.fs.fed.us/pnw/starkey/ . Accessed 7 Dec 2016

Zaghlool E, ElKaffas S, Saad, A (2015) A Density-Based Clustering of Spatio-Temporal Data. *New Contributions in Information Systems and Technologies*, 41-50. doi: 10.1007/978-3-319-16528-8_5

Zhang D, Lee K, Lee I (2018) Hierarchical trajectory clustering for spatio-temporal periodic pattern mining. Expert Systems With Applications 92:1–11.

Zhang P, Huang Y, Shekhar S, Kumar V (2003) Correlation analysis of spatial time series datasets: A filter-and-refine approach. Advances in Knowledge Discovery and Data Mining 532–544. doi: 10.1007/3-540-36175-8_53

Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: An Efficient Data Clustering Databases Method for Very Large. ACM SIGMOD International Conference on Management of Data 1:103–114. doi: 10.1145/233269.233324

Zhao Q, Xu M, Fränti P (2009) Sum-of-square based cluster validity index and significance analysis. Proc. of the 17th Int. Conf. on Adaptive and Natural Computing Algorithms :313–322.

Zheng Y, Zhang L, Xie X, Ma W-Y (2009) Mining interesting locations and travel sequences from GPS trajectories. Proceedings of the 18th international conference on World wide web - WWW ’09 791. doi: 10.1145/1526709.1526816

Zhou P, Salzberg B (2008) The hB-pi* Tree: An Optimized Comprehensive Access Method for Frequent-Update Multi-dimensional Point Data. In: Ludäscher B., Mamoulis N. (eds) Scientific and Statistical Database Management. SSDBM 2008. Lecture Notes in Computer Science, vol 5069. Springer, Berlin, Heidelberg