Zhe Jiang · Shashi Shekhar

Spatial Big Data Science

Classification Techniques for Earth Observation Imagery

Springer

Chapter 2
Spatial and Spatiotemporal Big Data Science

Abstract This chapter provides an overview of spatial and spatiotemporal big data science. This chapter starts with the unique characteristics of spatial and spatiotemporal data, and their statistical properties. Then, this chapter reviews recent computational techniques and tools in spatial and spatiotemporal data science, focusing on several major pattern families, including spatial and spatiotemporal outliers, spatial and spatiotemporal association and tele-connection, spatial and spatiotemporal prediction, partitioning and summarization, as well as hotspot and change detection.

This chapter overviews the state-of-the-art data mining and data science methods [1] for spatial and spatiotemporal big data. Existing overview tutorials and surveys in spatial and spatiotemporal big data science can be categorized into two groups: early papers in the 1990s without a focus on spatial and spatiotemopral statistical foundations, and recent papers with a focus on statistical foundation. Two early survey papers [2, 3] review spatial data mining from a database approach. Recent papers include brief tutorials on current spatial [4] and spatiotemporal data mining [1] techniques. There are also other relevant book chapters [5–7], as well as survey papers on specific spatial or spatiotemporal data mining tasks such as spatiotemporal clustering [8], spatial outlier detection [9], and spatial and spatiotemporal change footprint detection [10, 11].

This chapter makes the following contributions: (1) We provide a categorization of input spatial and spatiotemporal data types; (2) we provide a summary of spatial and spatiotemporal statistical foundations categorized by different data types; (3) we create a taxonomy of six major output pattern families, including spatial and spatiotemporal outliers, associations and tele-connections, predictive models, partitioning (clustering) and summarization, hotspots, and changes. Within each pattern family, common computational approaches are categorized by the input data types; and (4) we analyze the research trends and future research needs.

Organization of the chapter: This chapter starts with a summary of input spatial and spatiotemporal data (Sect. 2.1) and an overview of statistical foundation (Sect. 2.2). It then describes in detail six main output pattern families including spatial and spatiotemporal outliers, associations and tele-connections, predictive models, partitioning (clustering) and summarization, hotspots, and changes (Sect. 2.3). An

examination of research trend and future research needs is in Sect. 2.4. Section 2.5 summarizes the chapter.

2.1 Input: Spatial and Spatiotemporal Data

2.1.1 *Types of Spatial and Spatiotemporal Data*

The data inputs of spatial and spatiotemporal big data science tasks are more complex than the inputs of classical big data science tasks because they include discrete representations of continuous space and time. Table 2.1 gives a taxonomy of different spatial and spatiotemporal data types (or models). Spatial data can be categorized into three models, i.e., the object model, the field model, and the spatial network model [12, 13]. Spatiotemporal data, based on how temporal information is additionally modeled, can be categorized into three types, i.e., temporal snapshot model, temporal change model, and event or process model [14–16]. In the temporal snapshot model, spatial layers of the same theme are time-stamped. For instance, if the spatial layers are points or multi-points, their temporal snapshots are trajectories of points or spatial time series (i.e., variables observed at different times on fixed locations). Similarly, snapshots can represent trajectories of lines and polygons, raster time series, and spatiotemporal networks such as time-expanded graphs (TEGs) and time-aggregated graphs (TEGs) [17, 18]. The temporal change model represents spatiotemporal data with a spatial layer at a given start time together with incremental changes occurring afterward. For instance, it can represent motion (e.g., Brownian motion, random walk [19]) as well as speed and acceleration on spatial points, as well as rotation and deformation on lines and polygons. Event and process models represent temporal information in terms of *events* or *processes*. One way to distinguish events from processes is that events are entities whose properties are possessed timelessly and therefore are not subject to change over time, whereas processes are

Table 2.1 Taxonomy of spatial and spatiotemporal data models

Spatial data	Temporal snapshots (Time series)	Temporal change (Delta/Derivative)	Events/processes
Object model	Trajectories, Spatial time series	Motion, speed, acceleration, split or merge	Spatial or spatiotemporal point process
Field model	Raster time series	Change across raster snapshots	Cellular automation
Spatial network	Spatiotemporal network	Addition or removal of nodes, edges	

entities that are subject to change over time (e.g., a process may be said to be accelerating or slowing down) [20].

2.1.2 *Data Attributes and Relationships*

There are three distinct types of data attributes for spatiotemporal data, including non-spatiotemporal attributes, spatial attributes, and temporal attributes. Non-spatiotemporal attributes are used to characterize non-contextual features of objects, such as name, population, and unemployment rate for a city. They are the same as the attributes used in the data inputs of classical big data science [21]. Spatial attributes are used to define the spatial location (e.g., longitude and latitude), spatial extent (e.g., area, perimeter) [22, 23], shape, as well as elevation defined in a spatial reference frame. Temporal attributes include the time stamp of a spatial object, a raster layer, or a spatial network snapshot, as well as the duration of a process. Relationships on non-spatial attributes are often explicit, including arithmetic, ordering, and subclass. Relationships on spatial attributes, in contrast, are often implicit, including those in topological space (e.g., meet, within, overlap), set space (e.g., union, intersection), metric space (e.g., distance), and directions. Relationships on spatiotemporal attributes are more sophisticated, as summarized in Table 2.2.

One way to deal with implicit spatiotemporal relationships is to materialize the relationships into traditional data input columns and then apply classical big data science techniques [37–41]. However, the materialization can result in loss of information [7]. The spatial and temporal vagueness which naturally exists in data and relationships usually creates further modeling and processing difficulty in spatial and spatiotemporal big data science. A more preferable way to capture implicit spatial and spatiotemporal relationships is to develop statistics and techniques to incorporate spatial and temporal information into the data science process. These statistics and techniques are the main focus of the survey.

Table 2.2 Relationships on spatiotemporal data

Spatial data	Temporal snapshots (Time series)	Change (Delta/Derivative)	Event/Process
Object model	Spatiotemporal predicates [24], Trajectory distance [25, 26], spatial time series correlation [27], tele-connection [28]	Motion, speed, acceleration, attraction or repulsion, split/merge	Spatiotemporal covariance [19], spatiotemporal coupling for point events, or extended spatial objects [29–34]
Field model	Cubic map algebra [35], temporal correlation, tele-connection	Local, focal, zonal change across snapshots [10]	Cellular automation [36]

2.2 Statistical Foundations

2.2.1 Spatial Statistics for Different Types of Spatial Data

Spatial statistics [19, 42–44] is a branch of statistics concerned with the analysis and modeling of spatial data. The main difference between spatial statistics and classical statistics is that spatial data often fails to meet the assumption of an identical and independent distribution (i.i.d.). As summarized in Table 2.3, spatial statistics can be categorized according to their underlying spatial data type: Geostatistics for point referenced data, lattice statistics for areal data, and spatial point process for spatial point patterns.

Table 2.3 Taxonomy of spatial and spatiotemporal statistics

Spatial model	Spatial statistics	Spatiotemporal statistics
Object model	Geostatistics: • Stationarity, isotropy, variograms, Kriging Spatial point processes: • Poisson point process, spatial scan statistics, Ripley's K-function	Statistics for spatial time series: • Spatiotemporal stationarity, variograms, covariance, Kriging; • Temporal autocorrelation, tele-coupling. Spatiotemporal point processes: • Spatiotemporal Poission point process; Spatiotemporal scan statistics; Spatiotemporal K-function.
Field model	Lattice statistics (areal data model): • W-matrix, spatial autocorrelation, local indicators of spatial association (LISA); • MRF, SAR, CAR, Bayesian hierarchical model	Statistics for raster time series: • EOF analysis, CCA analysis; • Spatiotemporal autoregressive model (STAR), Bayesian hierarchical model, dynamic spatiotemporal model (Kalman filter), data assimilation
Spatial network	Spatial network autocorrelation, Network K-function, Network Kriging	

Geostatistics: Geostatistics [44] deal with the analysis of the properties of point reference data, including spatial continuity (i.e., dependence across locations), weak stationarity (i.e., first and second moments do not vary with respect to locations), and isotropy (i.e., uniformity in all directions). For example, under the assumption of weak stationarity (or more specifically intrinsic stationarity), variance of the difference of non-spatial attribute values at two point locations is a function of point location difference regardless of specific point locations. This function is called a variogram [45]. If the variogram only depends on distance between two locations (not varying with respect to directions), it is further called isotropic. Under the assumptions of these properties, Geostatistics also provides a set of statistical tools such as Kriging [45], which can be used to interpolate non-spatial attribute values at unsampled locations. Finally, real-world spatial data may not always satisfy the stationarity assumption. For example, different jurisdictions tend to produce different laws (e.g., speed limit differences between Minnesota and Wisconsin). This effect is called spatial heterogeneity or non-stationarity. Special models (e.g., geographically weighted regression, or GWR [46]) can be further used to model the varying coefficients at different locations.

Lattice statistics: Lattice statistics studies statistics for spatial data in the field (or areal) model. Here a lattice refers to a countable collection of regular or irregular cells in a spatial framework. The range of spatial dependency among cells is reflected by a neighborhood relationship, which can be represented by a contiguity matrix called a W-matrix. A spatial neighborhood relationship can be defined based on spatial adjacency (e.g., rook or queen neighborhoods) or Euclidean distance or, in more general models, cliques and hypergraphs [47]. Based on a W-matrix, spatial autocorrelation statistics can be defined to measure the correlation of a non-spatial attribute across neighboring locations. Common spatial autocorrelation statistics include Moran's I, Getis-Ord $Gi*$, Geary's C, Gamma index Γ [48], as well as their local versions called local indicators of spatial association (LISA) [49]. Several spatial statistical models, including the spatial autoregressive model (SAR), conditional autoregressive model (CAR), Markov random field (MRF), as well as other Bayesian hierarchical models [42], can be used to model lattice data. Another important issue is the modifiable areal unit problem (MAUP) (also called the multi-scale effect) [50], an effect in spatial analysis that results for the same analysis method will change on different aggregation scales. For example, analysis using data aggregated by states will differ from analysis using data at individual family level.

Spatial point processes: A spatial point process is a model for the spatial distribution of the points in a point pattern. It differs from point reference data in that the random variables are locations. Examples include positions of trees in a forest and locations of bird habitats in a wetland. One basic type of point process is a homogeneous spatial Poisson point process (also called complete spatial randomness, or CSR) [19], where point locations are mutually independent with the same intensity over space. However, real-world spatial point processes often show either spatial aggregation (clustering) or spatial inhibition instead of complete spatial independence as in CSR. Spatial statistics such as Ripley's K-function [51, 52], i.e., the average number of points within a certain distance of a given point normalized by

the average intensity, can be used to test spatial aggregation of a point pattern against CSR. Moreover, real-world spatial point processes such as crime events often contain hotspot areas instead of following homogeneous intensity across space. A spatial scan statistic [53] can be used to detect these hotspot patterns. It tests whether the intensity of points inside a scanning window is significantly higher (or lower) than outside. Though both the K-function and spatial scan statistics have the same null hypothesis of CSR, their alternative hypotheses are quite different: The K-function tests whether points exhibit spatial aggregation or inhibition instead of independence, while spatial scan statistics assume that points are independent and test whether a local hotspot with much higher intensity than outside exists. Finally, there are other spatial point processes such as the Cox process, in which the intensity function itself is a random function over space, as well as a cluster process, which extends a basic point process with a small cluster centered on each original point [19]. For extended spatial objects such as lines and polygons, spatial point processes can be generalized to line processes and flat processes in stochastic geometry [54].

Spatial network statistics: Most spatial statistics research focuses on the Euclidean space. Spatial statistics on the network space are much less studied. Spatial network space, e.g., river networks and street networks, is important in applications of environmental science and public safety analysis. However, it poses unique challenges including directionality and anisotropy of spatial dependency, connectivity, as well as high computational cost. Statistical properties of random fields on a network are summarized in [55]. Recently, several spatial statistics, such as spatial autocorrelation, K-function, and Kriging, have been generalized to spatial networks [56–58]. Little research has been done on spatiotemporal statistics on the network space.

2.2.2 *Spatiotemporal Statistics*

Spatiotemporal statistics [19, 59] combine spatial statistics with temporal statistics (time series analysis [60], dynamic models [59]). Table 2.3 summarizes common statistics for different spatiotemporal data types, including spatial time series, spatiotemporal point process, and time series of lattice (areal) data.

Spatial time series: Spatial statistics for point reference data have been generalized for spatiotemporal data [61]. Examples include spatiotemporal stationarity, spatiotemporal covariance, spatiotemporal variograms, and spatiotemporal Kriging [19, 59]. There is also temporal autocorrelation and tele-coupling (high correlation across spatial time series at a long distance). Methods to model spatiotemporal process include physics inspired models (e.g., stochastically differential equations) [19] and hierarchical dynamic spatiotemporal models (e.g., Kalman filtering) for data assimilation [19].

Spatiotemporal point process: A spatiotemporal point process generalizes the spatial point process by incorporating the factor of time. As with spatial point processes, there are spatiotemporal Poisson process, Cox process, and cluster process. There

are also corresponding statistical tests including a spatiotemporal K-function and spatiotemporal scan statistics [19].

Time series of lattice (areal) data: Similar to lattice statistics, there are spatial and temporal autocorrelation, SpatioTemporal Autoregressive Regression (STAR) model [62], and Bayesian hierarchical models [42]. Other spatiotemporal statistics include empirical orthogonal function (EOF) analysis (principle component analysis in geophysics), canonical correlation analysis (CCA), and dynamic spatiotemporal models (Kalman filter) for data assimilation [59].

2.3 Output Pattern Families

2.3.1 *Spatial and Spatiotemporal Outlier Detection*

This section reviews techniques for spatial and spatiotemporal outlier detection. The section begins with a definition of spatial or spatiotemporal outliers by comparison with global outliers. Spatial and spatiotemporal outlier detection techniques are summarized according to their input data types.

Problem definition: To understand the meaning of spatial and spatiotemporal outliers, it is useful first to consider global outliers. Global outliers [63, 64] have been informally defined as observations in a dataset which appear to be inconsistent with the remainder of that set of data, or which deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism. In contrast, a spatial outlier [65] is a spatially referenced object whose non-spatial attribute values differ significantly from those of other spatially referenced objects in its spatial neighborhood. Informally, a spatial outlier is a local *instability* or *discontinuity*. For example, a new house in an old neighborhood of a growing metropolitan area is a spatial outlier based on the non-spatial attribute house age. Similarly, a spatiotemporal outlier generalizes spatial outliers with a spatiotemporal neighborhood instead of a spatial neighborhood.

Statistical foundation: The spatial statistics for spatial outlier detection are also applicable to spatiotemporal outliers as long as spatiotemporal neighborhoods are well-defined. The literature provides two kinds of bipartite multi-dimensional tests: graphical tests, including variogram clouds [66] and Moran scatterplots [44, 49], and quantitative tests, including scatterplot [67] and neighborhood spatial statistics [65].

2.3.1.1 Spatial Outlier Detection

The *visualization approach* plots spatial locations on a graph to identify spatial outliers. The common methods are variogram clouds and Moran scatterplot as introduced earlier.

The *neighborhood approach* defines a spatial neighborhood, and a spatial statistic is computed as the difference between the non-spatial attribute of the current location and that of the neighborhood aggregate [65]. Spatial neighborhoods can be identified by distance on spatial attributes (e.g., K-nearest neighbors), or by graph connectivity (e.g., locations on road networks). This work has been extended in a number of ways to allow for multiple non-spatial attributes [68], average and median attribute value [69], weighted spatial outliers [70], categorical spatial outlier [71], local spatial outliers [72], and fast detection algorithms [73], and parallel algorithms on GPU for big spatial event data [74].

2.3.1.2 Spatiotemporal Outlier Detection

The intuition behind spatiotemporal outlier detection is that they reflect "discontinuity" on non-spatiotemporal attributes within a spatiotemporal neighborhood. Approaches can be summarized according to the input data types.

Outliers in spatial time series: For spatial time series (on point reference data, raster data, as well as graph data), basic spatial outlier detection methods, such as visualization-based approaches and neighborhood-based approaches, can be generalized with a definition of spatiotemporal neighborhoods.

Flow Anomalies: Given a set of observations across multiple spatial locations on a spatial network flow, flow anomaly discovery aims to identify dominant time intervals where the fraction of time instants of significantly mismatched sensor readings exceeds the given percentage-threshold. Flow anomaly discovery can be considered as detecting *discontinuities* or *inconsistencies* of a non-spatiotemporal attribute within a neighborhood defined by the flow between nodes, and such discontinuities are persistent over a period of time. A time-scalable technique called SWEET (Smart Window Enumeration and Evaluation of persistent-Thresholds) was proposed [75] that utilizes several algebraic properties in the flow anomaly problem to discover these patterns efficiently.

2.3.2 *Spatial and Spatiotemporal Associations, Tele-Connections*

This section reviews techniques for identifying spatial and spatiotemporal association as well as tele-connections. The section starts with the basic spatial association (or colocation) pattern and moves on to spatiotemporal association (i.e., spatiotemporal co-occurrence, cascade, and sequential patterns) as well as spatiotemporal tele-connection.

Pattern definition: Spatial association, also known as spatial colocation patterns [76], represents subsets of spatial event types whose instances are often located in close geographic proximity. Real-world examples include symbiotic species, e.g., the Nile Crocodile and Egyptian Plover in ecology. Similarly, spatiotemporal

association patterns represent spatiotemporal object types whose instances often occur in close geographic and temporal proximity. Spatiotemporal coupling patterns can be categorized according to whether there exists temporal ordering of object types: spatiotemporal (mixed drove) co-occurrences [77] are used for unordered patterns, spatiotemporal cascades [31] for partially ordered patterns, and spatiotemporal sequential patterns [33] for totally ordered patterns. Spatiotemporal tele-connection [27] represents patterns of significantly positive or negative temporal correlation between a pair of spatial time series.

Challenges: Mining patterns of spatial and spatiotemporal association are challenging due to the following reasons: First, there is no explicit transaction in continuous space and time; second, there is potential for over-counting; and third, the number of candidate patterns is exponential, and a trade-off between statistical rigor of output patterns and computational efficiency has to be made.

Statistical foundation: The underlying statistic for spatiotemporal coupling patterns is the cross-K-function, which generalizes the basic Ripley's K-function (introduced in Sect. 2.2) for multiple event types.

Common approaches: The following subsections categorize common computational approaches for discovering spatial and spatiotemporal couplings by different input data types.

Spatial colocation: Mining colocation patterns can be done via statistical approaches including cross-K-function with Monte Carlo simulation [44], mean nearest neighbor distance, and spatial regression model [78], but these methods are often computationally very expensive due to the exponential number of candidate patterns. In contrast, data mining approaches aim to identify colocation patterns like association rule mining. Within this category, there are transaction-based approaches and distance-based approaches. A transaction-based approach defines transactions over space (e.g., around instances of a reference feature) and then uses an Apriori-like algorithm [79]. A distance-based approach defines a distance-based pattern called k-neighboring class sets [80] or using an event centric model [76] based on a definition of *participation index*, which is an upper bound of cross-K-function statistic and has an anti-monotone property. Recently, approaches have been proposed to identify colocations for extended spatial objects [81] or rare events [82], regional colocation patterns [83–85] (i.e., pattern is significant only in a subregion), statistically significant colocation [86], as well as design fast algorithms [87].

Spatiotemporal event associations represent subsets of two or more event types whose instances are often located in close spatial and temporal proximity. Spatiotempral event associations can be categorized into *spatiotemporal co-occurrences*, *spatiotemporal cascades*, and *spatiotemporal sequential patterns* for temporally unordered events, partially ordered events, and totally ordered events, respectively. To discover spatiotemporal co-occurrences, a monotonic composite interest measure and novel mining algorithms are presented in [77]. A filter-and-refine approach has also been proposed to identify spatiotemporal co-occurrences on extended spatial objects [30]. A spatiotemporal sequential pattern represents a "chain reaction" from different event types. A measure of *sequence index*, which can be interpreted by K-function statistic, was proposed in [33], together with computationally efficient

algorithms. For spatiotemporal cascade patterns, a statistically meaningful metric was proposed to quantify interestingness and pruning strategies were proposed to improve computational efficiency [31].

Spatiotemporal association from moving objects trajectories: Mining spatiotemporal association from trajectory data is more challenging than from spatiotemporal event data due to the existence of temporal duration, different moving directions, and imprecise locations. There are a variety of ways to define spatiotemporal association patterns from moving object trajectories. One way is to generalize the definition from spatiotemporal event data. For example, a pattern called spatiotemporal colocation episodes is defined to identify frequent sequences of colocation patterns that share a common event (object) type [88]. As another example, a spatiotemporal sequential pattern is defined based on decomposition of trajectories into line segments and identification of frequent region sequences around the segments [89]. Another way is to define spatiotemporal association as group of objects that frequently move together, either focusing on the footprints of subpaths (region sequences) that are commonly traversed [90] or subsets of objects that frequently move together (also called *travel companion*) [91].

Spatial time series oscillation and tele-connection: Given a collection of spatial time series at different locations, tele-connection discovery aims to identify pairs of spatial time series whose correlation is above a given threshold. Tele-connection patterns are important in understanding oscillations in climate science. Computational challenges arise from the large number of candidate pairs and the length of time series. An efficient index structure, called a cone-tree, as well as a filter-and-refine approach [27], has been proposed which utilizes spatial autocorrelation of nearby spatial time series to filter out redundant pairwise correlation computation. Another challenge is the existence of spurious "high correlation" patterns that happen by chance. Recently, statistical significant tests have been proposed to identify statistically significant tele-connection patterns called dipoles from climate data [28]. The approach uses a "wild bootstrap" to capture the spatiotemporal dependencies and takes account of the spatial autocorrelation, the seasonality, and the trend in the time series over a period of time.

2.3.3 *Spatial and Spatiotemporal Prediction*

Problem definition: Given training samples with features and a target variable as well as a spatial neighborhood relationship among samples, the problem of *spatial prediction* aims to learn a model that can predict the target variable based on features. What distinguishes spatial prediction from traditional prediction problem in data mining is that data items are embedded in space and often violate the common assumption of an identical and independent distribution (i.i.d.). Spatial prediction problems can be further categorized into *spatial classification* for nominal (i.e., categorical) target variables and *spatial regression* for numeric target variables.

Challenges: The unique challenges of spatial and spatiotemporal prediction come from the special characteristics of spatial and spatiotemporal data, which include spatial and temporal autocorrelation, spatial heterogeneity, and temporal non-stationarity, as well as the multi-scale effect. These unique characteristics violate the common assumption in many traditional prediction techniques that samples follow an identical and independent distribution (i.i.d.). Simply applying traditional prediction techniques without incorporating these unique characteristics may produce hypotheses or models that are inaccurate or inconsistent with the dataset.

Statistical foundations: Spatial and spatiotemporal prediction techniques are developed based on spatial and spatiotemporal statistics, including spatial and temporal autocorrelation, spatial heterogeneity, temporal non-stationarity, and multiple areal unit problem (MAUP) (see Sect. 2.2).

Computational approaches: The following subsections summarize common spatial and spatiotemporal prediction approaches for different data types. We further categorize these approaches according to the challenges that they address, including spatial and spatiotemporal autocorrelation, spatial heterogeneity, spatial multi-scale effect, and temporal non-stationarity, and introduce each category separately below.

2.3.3.1 Spatial Autocorrelation or Dependency

According to Tobler's first law of geography [92], "everything is related to everything else, but near things are more related than distant things." The spatial autocorrelation effect tells us that spatial samples are not statistically independent, and nearby samples tend to resemble each other. There are different ways to incorporate the effect of spatial autocorrelation or dependency into predictive models, including spatial feature creation, explicit model structure modification, and spatial regularization in objective functions.

Spatial feature creation: The main idea is to create new features that incorporate spatial contextual (neighborhood) information. Spatial features can be generated directly from spatial aggregation [93] and indirectly from multi-relationship (or spatial association) rules between spatial entities [94–96] or from spatial transformation of raw features [97]. After spatial features are generated, they can be fed into a general prediction model. One advantage of this approach is that it could utilize many existing predictive models without significant modification. However, spatial feature creation in preprocessing phase is often application specific and time-consuming.

Spatial interpolation: Given observations of a variable at a set of locations (point reference data), spatial interpolation aims to measure the variable value at an unsampled location [98]. These techniques are broadly classified into three categories: geostatistical, non-geostatistical, and some combined approaches. Among the non-geostatistical approaches, the nearest neighbors, inverse distance weighting, etc., are the mostly used techniques in the literature. *Kriging* is the most widely used geostatistical interpolation technique, which represents a family of generalized least-squares regression-based interpolation techniques [99]. *Kriging* can be broadly classified into two categories: univariate (only variable to be predicted) and multivariate (there are

some *covariates*, also called explanatory variables). Unlike the non-geostatistical or traditional interpolation techniques, this estimator considers both the distance and the degree of variation between the sampled and unsampled locations for the random variable estimation. Among the univariate kriging methods, the *simple kriging* and *ordinary kriging*, and in multivariate scenario, the *ordinary cokriging*, *universal kriging* and *kriging with external drift* are the most popular and widely used technique in the study of spatial interpolation [98, 100]. However, the *kriging* suffers from some acute shortcomings of assuming the isotopic nature of the random variables.

Markov random field (MRF): MRF [45] is a widely used model in image classification problems. It assumes that the class label of one pixel only depends on the class labels of its predefined neighbors (also called Markov property). In spatial classification problem, MRF is often integrated with other non-spatial classifiers to incorporate the spatial autocorrelation effect. For example, MRF has been integrated with maximum likelihood classifiers (MLC) to create Markov random field (MRF)-based Bayesian classifiers [101], in order to avoid salt-and-pepper noise in prediction [102]. Another example is the model of Support Vector Random Fields [103].

Spatial Autoregressive Model (SAR): In the spatial autoregression model, the spatial dependencies of the error term, or the dependent variable, are directly modeled in the regression equation [104]. If the dependent values y_i are related to each other, then the regression equation can be modified as $y = \rho Wy + X\beta + \epsilon$, where W is the neighborhood relationship contiguity matrix and ρ is a parameter that reflects the strength of the spatial dependencies between the elements of the dependent variable. For spatial classification problems, logistic transformation can be applied to SAR model for binary classes.

Conditional autoregressive model (CAR): In the conditional autoregressive model [45], the spatial autocorrelation effect is explicitly modeled by the conditional probability of the observation of a location given observations of neighbors. CAR is essentially a Markov random field. It is often used as a spatial term in Bayesian hierarchical models.

Spatial accuracy objective function: In traditional classification problems, the objective function (or loss function) often measures the zero-one loss on each sample, no matter how far the predicted class is from the location of the actuals. For example, in bird nest location prediction problem on a rasterized spatial field, a cell's predicted class (e.g., bird nest) is either correct or incorrect. However, if a cell mistakenly predicted as the bird nest class is very close to an actual bird nest cell, the prediction accuracy should not be considered as zero. Thus, spatial accuracy [105, 106] has been proposed to measure not only how accurate each cell is predicted itself but also how far it is from an actual class locations. A case study has shown that learning models based on proposed objective function produce better accuracy in bird nest location prediction problem. Spatial objective function has also been proposed in active learning [107], in which the cost of additional label not only considers accuracy but also travel cost between locations to be labeled.

2.3.3.2 Spatial Heterogeneity

Spatial heterogeneity describes the fact that samples often do not follow an identical distribution in the entire space due to varying geographic features. Thus, a global model for the entire space fails to capture the varying relationships between features and the target variable in different subregion. The problem is essentially the multi-task learning problem, but a key challenge is how to identify different tasks (or regional or local models). Several approaches have been proposed to learn local or regional models. Some approaches first partition the space into homogeneous regions and learn a local model in each region. Others learn local models at each location but add spatial constraint that nearby models have similar parameters.

Geographically Weighted Regression (GWR): One limitation of the spatial autoregressive model (SAR) is that it does not account for the underlying spatial heterogeneity that is natural in the geographic space. Thus, in a SAR model, coefficients β of covariates and the error term ϵ are assumed to be uniform throughout the entire geographic space. One proposed method to account for spatial variation in model parameters and errors is Geographically Weighted Regression [46]. The regression equation of GWR is $y = X\beta(s) + \epsilon(s)$, where $\beta(s)$ and $\epsilon(s)$ represent the spatially parameters and the errors, respectively. GWR has the same structure as standard linear regression, with the exception that the parameters are spatially varying. It also assumes that samples at nearby locations have higher influence on the parameter estimation of a current location. Recently, a multi-model regression approach is proposed to learn a regression model at each location but regularize the parameters to maintain spatial smoothness of parameters at neighboring locations [108].

2.3.3.3 Multi-scale Effect

One main challenge in spatial prediction is the Multiple Area Unit Problem (MAUP), which means that analysis results will vary with different choices of spatial scales. For example, a predictive model that is effective at the county level may perform poorly at states level. Recently, a computation technique has been proposed to learn a predict models from different candidate spatial scales or granularity [94].

2.3.3.4 Spatiotemporal Autocorrelation

Approaches that address the spatiotemporal autocorrelation are often extensions of previously introduced models that address spatial autocorrelation effect by further considering the time dimension. For example, SpatioTemporal Autoregressive Regression (STAR) model [44] extends SAR by further modeling temporal or spatiotemporal dependency across variables at different locations. Spatiotemporal Kriging [59] generalizes spatial kriging with a spatiotemporal covariance matrix and variograms. It can be used to make predictions from incomplete and noisy spatiotemporal data. *Spatiotemporal relational probability trees and forests* [109]

extend decision tree classifiers with tree node tests on spatial properties on objects and random field as well as temporal changes. To model spatiotemporal events such as disease counts in different states at a sequence of times, *Bayesian hierarchical models* are often used, which incorporate the spatial and temporal autocorrelation effects in explicit terms.

2.3.3.5 Temporal Non-stationarity

Hierarchical dynamic spatiotemporal models (DSTMs) [59], as the name suggests, aim to model spatiotemporal processes dynamically with a Bayesian hierarchical framework. There are three levels of models in the hierarchy: a data model on the top, a process model in the middle, and a parameter model at the bottom. A data model represents the conditional dependency of (actual or potential) observations on the underlying hidden process with latent variables. A process model captures the spatiotemporal dependency within the process model. A parameter model characterizes the prior distributions of model parameters. DSTMs have been widely used in climate science and environment science, e.g., for simulating population growth or atmospheric and oceanic processes. For model inference, Kalman filter can be used under the assumption of linear and Gaussian models.

2.3.3.6 Prediction for Moving Objects

Mining moving object data such as GPS trajectories and check-in histories has become increasingly important. Due to space limit, we briefly discuss some representative techniques for three main problems: trajectory classification, location prediction, and location recommendation.

Trajectory classification: This problem aims to predict the class of trajectories. Unlike spatial classification problems for spatial point locations, trajectory classification can utilize the order of locations visited by moving objects. An approach has been proposed that uses frequent sequential patterns within trajectories for classification [110].

Location prediction: Given historical locations of a moving object (e.g., GPS trajectories, check-in histories), the location prediction problem aims to forecast the next place that the object will visit. Various approaches have been proposed [111–113]. The main idea is to identify the frequent location sequences visited by moving objects, and then, next location can be predicted by matching the current sequence with historical sequences. Social, temporal, and semantic information can also be incorporated to improve prediction accuracy. Some other approaches use hidden Markov model to capture the transition between different locations. Supervised approaches have also been used.

Location recommendation: Location recommendation [114–118] aims to suggest potentially interesting locations to visitors. Sometimes, it is considered as a special location prediction problem which also utilizes location histories of other

moving objects. Several factors are often considered for ranking candidate locations, such as local popularity and user interests. Different factors can be simultaneously incorporated via generative models such as latent Dirichlet allocation (LDA) and probabilistic matrix factorization techniques.

2.3.4 *Spatial and Spatiotemporal Partitioning (Clustering) and Summarization*

Problem definition: Spatial partitioning aims to divide spatial items (e.g., vector objects, lattice cells) into groups such that items within the same group have high proximity. Spatial partitioning is often called *spatial clustering*. We use the name "spatial partitioning" due to the unique nature of spatial data, i.e., grouping spatial items also mean partitioning the underlying space. Similarly, *spatiotemporal partitioning*, or *spatiotemporal clustering*, aims to group similar spatiotemporal data items and thus partition the underlying space and time. After spatial or spatiotemporal partitioning, one often needs to find a compact representation of items in each partition, e.g., aggregated statistics or representative objects. This process is further called *spatial or spatiotemporal summarization*.

Challenges: The challenges of spatial and spatiotemporal partitioning come from three aspects. First, patterns of spatial partitions in real-world datasets can be of various shapes and sizes and are often mixed with noise and outliers. Second, relationships between spatial and spatiotemporal data items (e.g., polygons, trajectories) are more complicated than traditional non-spatial data. Third, there is a trade-off between quality of partitions and computational efficiency, especially for large datasets.

Computational approaches: Common spatial and spatiotemporal partitioning approaches are summarized in below according to the input data types.

2.3.4.1 Spatial Partitioning (Clustering)

Spatial and spatiotemporal partitioning approaches can be categorized by input data types, including spatial points, spatial time series, trajectories, spatial polygons, raster images, raster time series, spatial networks, and spatiotemporal points.

Spatial point partitioning (clustering): The goal is to partition two-dimensional points into clusters in Euclidean space. Approaches can be categorized into global methods, hierarchical methods, and density-based methods according to the underlying assumptions on the characteristics of clusters [119]. Global methods assume clusters to have "compact" or globular shapes and thus minimize the total distance from points to their cluster centers. These methods include K-means, K-medoids, EM algorithm, CLIQUE, BIRCH, and CLARANS [21]. Hierarchical methods [21] form clusters hierarchically in a top-down or bottom-up manner and are robust to outliers since outliers are often easily separated out. Chameleon [120] is a graph-based

hierarchical clustering method that first creates a sparse k-nearest neighbor graph, then partitions the graph into small clusters, and hierarchically merges small clusters whose properties stay mostly unchanged after merging. Density-based methods such as DB-Scan [121] assume clusters to contain dense points and can have arbitrary shapes. When the density of points varies across space, the similarity measure of *shared nearest neighbors* [122] can be used. Voronoi diagram [123] is another space partitioning technique that is widely used in applications of location-based service. Given a set of spatial points in Euclidean space, a Voronoi diagram partitions the space into cells according to the nearest spatial points.

Spatial polygon clustering: Spatial polygon clustering is more challenging than point clustering due to the complexity of distance measures between polygons. Distance measures on polygons can be defined based on dissimilarities on spatial attribute (e.g., Hausdorff distance, ratio of overlap, extent, direction, and topology) as well as non-spatial attributes [124, 125]. Based on these distance measures, traditional point clustering algorithms such as K-means, CLARANS, and shared nearest neighbor algorithm can be applied.

Spatial areal data partitioning: Spatial areal data partitioning has been extensively studied for image segmentation tasks. The goal is to partition areal data (e.g., images) into regions that are homogeneous in non-spatial attributes (e.g., color or gray tone and texture) while maintaining spatial continuity (without small holes). Similar to spatial point clustering, there is no uniform solution. Common approaches can be categorized into non-spatial attribute-guided spatial clustering, single, centroid, or hybrid linkage region growing schemes, and split-and-merge scheme. More details can be found in a survey on image segmentation [126].

Spatial network partitioning: Spatial network partitioning (clustering) is important in many applications such as transportation and VLSI design. Network Voronoi diagram is a simple method to partition spatial network based on common closest interesting nodes (e.g., service centers). Recently, a connectivity constraint network Voronoi diagram (CCNVD) has been proposed to add capacity constraint to each partition while maintaining spatial continuity [127]. METIS [128] provides a set of scalable graph partitioning algorithms, which have shown high partition quality and computational efficiency.

2.3.4.2 Spatiotemporal Partitioning (clustering)

Spatiotemporal event partitioning (clustering): Most methods for 2-D spatial point clustering [119] can be easily generalized to 3-D spatiotemporal event data [129]. For example, ST-DBSCAN [130] is a spatiotemporal extension of the density-based spatial clustering method DBSCAN. ST-GRID [131] is another example that extends grid-based spatial clustering methods into 3-D grids.

Spatial time series partitioning (clustering): Spatial time series clustering aims to divide the space into regions such that the similarity between time series within the same region is maximized. Global partitioning methods such as K-means, K-medoids, and EM, as well as the hierarchical methods, can be applied.

Common (dis)similarity measures include Euclidean distance, Pearson's correlation, and dynamic time warping (DTW) distance. More details can be found in a recent survey [132]. However, due to the high dimensionality of spatial time series, density-based approaches and graph-based approaches are often not used. When computing similarities between spatial time series, a filter-and-refine approach [27] can be used to avoid redundant computation.

Trajectory partitioning: Trajectory partitioning approaches can be categorized by their objectives, namely trajectory grouping, flock pattern detection, and trajectory segmentation. Trajectory grouping aims to partition trajectories into groups according to their similarity. There are mainly two types of approaches, i.e., distance-based and frequency-based. The *density-based approaches* [133–135] first break trajectories into small segments and apply distance-based clustering algorithms similar to K-means or DBSCAN to connect dense areas of segments. The *frequency-based approach* [136] uses association rule mining [40] algorithms to identify subsections of trajectories which have high frequencies (also called high "support").

2.3.4.3 Spatial and Spatiotemporal Summarization

Data summarization aims to find compact representation of a dataset [137]. It is important for data compression as well as for making pattern analysis more convenient. Summarization can be done on classical data, spatial data, as well as spatiotemporal data.

Classical data summarization: Classical data can be summarized with aggregation statistics such as count, mean, and median. Many modern database systems provide query support for this operation, e.g., "Group by" operator in SQL.

Spatial data summarization: Spatial data summarization is more difficult than classical data summarization due to its non-numeric nature. For Euclidean space, the task can be done by first conducting spatial partitioning and then identifying representative spatial objects. For example, spatial data can be summarized with the centroids or medoids computed from K-means or K-medoids algorithms. For network space, especially for spatial network activities, summarization can be done by identifying several primary routes that cover those activities as much as possible. A K-Main Routes (KMR) algorithm [138] has been proposed to efficiently compute such routes to summarize spatial network activities. To reduce the computational cost, the KMR algorithm uses network Voronoi diagrams, divide and conquer, and pruning techniques.

Spatiotemporal data summarization: For spatial time series data, summarization can be done by removing spatial and temporal redundancy due to the effect of autocorrelation. A family of such algorithms has been used to summarize traffic data streams [139]. Similarly, the centroids from K-means can also be used to summarize spatial time series. For trajectory data, especially spatial network trajectories, summarization is more challenging due to the huge cost of similarity computation. A recent approach summarizes network trajectories into k-primary corridors

[140, 141]. The work proposes efficient algorithms to reduce the huge cost for network trajectory distance computation.

2.3.5 *Spatial and Spatiotemporal Hotspot Detection*

Problem definition: Given a set of spatial objects (e.g., points) in a study area, the problem of *spatial hotspot detection* aims to find regions where the number of objects is unexpectedly or anomalously high. Spatial hotspot detection is different from spatial partitioning or clustering, since spatial hotspots are a special kind of clusters whose intensity is "significantly" higher than the outside. *Spatiotemporal hotspots* can be seen as a generalization of spatial hotspots with a specified time window.

Challenges: Spatial and spatiotemporal hotspot detection is a challenging task since the location, size, and shape of a hotspot are unknown beforehand. In addition, the number of hotspots in a study area is often not known either. Moreover, "false" hotspots that aggregate events only by chance should often be avoided, since these false hotspots impede proper response by authorities (e.g., wasting police resources). Thus, it is often important to test the statistical significance of candidate spatial or spatiotemporal hotspots.

Statistical foundation: Spatial (or spatiotemporal) scan statistics [53, 142] (also discussed in Sect. 3.1) are used to detect statistically significant hotspots from spatial (or spatiotemporal) datasets. It uses a window (or cylinder) to scan the space (or space–time) for candidate hotspots and performs hypothesis testing. The null hypothesis states that the spatial (or spatiotemporal) points are completely spatially random (a homogeneous Poisson point process). The alternative hypothesis states that the points inside of the window (or cylinder) have higher intensity of points than outside. A test statistic called log likelihood ratio is computed for each candidate hotspot, and the candidate with the highest likelihood ratio can be further evaluated by its significance value (i.e., p-value).

Computational approaches: The following subsections summarize common spatial and spatiotemporal hotspot detection approaches by different input data types.

2.3.5.1 Spatial Hotspot from Spatial Point Pattern

Spatial partitioning approaches: Spatial point partitioning or clustering methods (Sect. 4.4.1) can be used to identify candidate hotspot patterns. After this, statistical tools may be used to evaluate the statistical significance of candidate patterns. Many of these methods have been implemented in CrimeStat, a software package for crime hotspot analysis [143].

Spatial scan statistics based approaches: These approaches use a window with varying sizes to scan the 2-D plane and identifies the candidate window with the highest likelihood ratio. Statistical significance (p-value) is computed for this candidate based on Monte Carlo simulation. Scanning windows with different shapes,

including circular, elliptical, as well as ring-shaped, have been proposed together with efficient computational pruning strategies [142, 144–146]. SaTScan [142] is a popular spatial scan statistics tool in epidemiology to analyze circular or elliptical hotspots.

Kernel Density Estimation: Kernel density estimation (KDE) [147] identifies spatial hotspots via a density map of point events. It first creates a grid over the study area and uses a kernel function with a user-defined radius (bandwidth) on each point to estimate the density of points on centers of grid cells. A subset of grid cells with high density are returned as spatial hotspots.

2.3.5.2 Spatial Hotspot from Areal Model

Local Indicators of Spatial Association: Local indicators of spatial association (LISA) [148, 149] is a set of local spatial autocorrelation statistics, including local Moran's I, Geary's C, or Ord Gi and Gi* functions. It differs from global spatial autocorrelation in that the statistics are computed within the neighborhood of a location. For example, a high local Moran's I indicates that values of the current location as well as its neighbors are both extremely high (or low) compared to values at other locations, and thus, the neighborhood is a spatial hotspot (or "cold spot").

2.3.5.3 Spatiotemporal Hotspot Detection

Hot routes from spatial network trajectories: Hot routes detection from spatial network trajectories aims to detect network paths with high density [133] or frequency of trajectories [136]. Other approaches include organizing police patrol routes [150], main streets [151], and clumping [152].

Spatiotemporal Scan Statistics based approaches: Two types of spatiotemporal hotspots can be detected by spatiotemporal scan statistics: "persistent" spatiotemporal hotspots and "emerging" spatiotemporal hotspots. A "persistent" spatiotemporal hotspot is a region where the rate of increase in observations is a high and almost constant value over time. Thus, approaches to detect a persistent spatiotemporal hotspot involves counting observations in each time interval [142]. An "emerging" spatiotemporal hotspot is a region where the rate of observations monotonically increases over time [145, 153]. This kind of spatiotemporal hotspot occurs when an outbreak emerges causing a sudden increase in the number observations. Tools for the detection of emerging spatiotemporal hotspots use spatial scan statistics to identify changes in expectation over time [154].

2.3.6 *Spatiotemporal Change*

2.3.6.1 What Are Spatiotemporal Changes and Change Footprints

Although the single term "change" is used to name the spatiotemporal change footprint patterns in different applications, the underlying phenomena may differ significantly. This section briefly summarizes the main ways a change may be defined and detected in spatiotemporal data [10].

Change in Statistical Parameter: In this case, the data is assumed to follow a certain distribution and the change is defined as a shift in this statistical distribution. For example, in statistical quality control, a change in the mean or variance of the sensor readings is used to detect a fault.

Change in Actual Value: Here, change is modeled as the difference between a data value and its spatial or temporal neighborhood. For example, in a one-dimensional continuous function, the magnitude of change can be characterized by the derivative function, while on a two-dimensional surface, it can be characterized by the gradient magnitude.

Change in Models Fitted to Data: This type of change is identified when a number of models are fitted to the data and one or more of the models exhibits a change (e.g., a discontinuity between consecutive linear functions) [155].

2.3.6.2 Common Approaches

This section follows the taxonomy of spatiotemporal change footprint patterns as proposed in [10]. In this taxonomy, spatiotemporal change footprints are classified along two dimensions: temporal and spatial. Temporal footprints are classified into four categories: single snapshot, set of snapshots, point in a long series, and interval in a long series. Single snapshot refers to a purely spatial change that does not have a temporal context. A set of snapshots indicate a change between two or more snapshots of the same spatial field, e.g., satellite images of the same region.

Spatial footprints can be classified as raster footprints or vector footprints. Vector footprints are further classified into four categories: point(s), line(s), polygon(s), and network footprint patterns. Raster footprints are classified based on the scale of the pattern, namely local, focal, or zonal patterns. This classification describes the scale of the change operation of a given phenomenon in the spatial raster field [156]. Local patterns are patterns in which change at a given location depends only on attributes at this location. Focal patterns are patterns in which change in a location depends on attributes in that location and its assumed neighborhood. Zonal patterns define change using an aggregation of location values in a region.

Spatiotemporal Change Patterns with Raster-based Spatial Footprint: This includes patterns of *spatial changes between snapshots*. In remote sensing, detecting changes between satellite images can help identify land cover change due to human activity, natural disasters, or climate change [157–159]. Given two geographically aligned raster images, this problem aims to find a collection of pixels that have significant changes between the two images [160]. This pattern is classified as a local change between snapshots since the change at a given pixel is assumed to be independent of changes at other pixels. Alternative definitions have assumed that a change at a pixel also depends on its neighborhoods [161]. For example, the pixel values in each block may be assumed to follow a Gaussian distribution [162]. We refer to this type of change footprint pattern as a focal spatial change between snapshots. Researchers in remote sensing and image processing have also tried to apply image change detection to objects instead of pixels [163–165], yielding zonal spatial change patterns between snapshots.

A well-known technique for detecting a local change footprint is simple differencing. The technique starts by calculating the differences between the corresponding pixels intensities in the two images. A change at a pixel is flagged if the difference at the pixel exceeds a certain threshold. Alternative approaches have also been proposed to discover focal change footprints between images. For example, the block-based density ratio test detects change based on a group of pixels, known as a block [166, 167]. Object-based approaches in remote sensing [165, 168, 169] employ image segmentation techniques to partition temporal snapshots of images into homogeneous objects [170] and then classify object pairs in the two temporal snapshots of images into no change or change classes.

Spatiotemporal Change Patterns with Vector-based Spatial Footprint: This includes the *Spatiotemporal Volume Change Footprint* pattern. This pattern represents a change process occurring in a spatial region (a polygon) during a time interval. For example, an outbreak event of a disease can be defined as an increase in disease reports in a certain region during a certain time window up to the current time. Change patterns known to have an spatiotemporal volume footprint include the spatiotemporal scan statistics [171, 172], a generalization of the spatial scan statistic, and emerging spatiotemporal clusters defined by [154].

2.4 Research Trend and Future Research Needs

Most current research in spatial and spatiotemporal data science uses Euclidean space, which often assumes isotropic property and symmetric neighborhoods. However, in many real-world applications, the underlying space is network space, such as river networks and road networks [138, 173, 174]. One of the main challenges in spatial and spatiotemporal network data science is to account for the network structure in the dataset. For example, in anomaly detection, spatial techniques do not consider the spatial network structure of the dataset; that is, they may not be able to model graph properties such as one ways, connectivity, and left-turns. The network structure often violates the isotropic property and symmetry of neighborhoods and instead requires asymmetric neighborhood and directionality of neighborhood relationship (e.g., network flow direction).

Recently, some cutting edge research has been conducted in the spatial network statistics and data science [57]. For example, several spatial network statistical methods have been developed, e.g., network K-function and network spatial autocorrelation. Several spatial analysis methods have also been generalized to the network space, such as network point cluster analysis and clumping method, network point density estimation, network spatial interpolation (Kriging), as well as network Huff model. Due to the nature of spatial network space as distinct from Euclidean space, these statistics and analysis often rely on advanced spatial network computational techniques [57].

We believe more spatial and spatiotemporal big data science research is still needed in the network space. First, though several spatial statistics and big data science techniques have been generalized to the network space, few spatiotemporal network statistics and big data science have been developed, and the vast majority of research is still in the Euclidean space. Future research is needed to develop more spatial network statistics, such as spatial network scan statistics, spatial network random field model, as well as spatiotemporal autoregressive models for networks. Furthermore, phenomena observed on spatiotemporal networks need to be interpreted in an appropriate frame of reference to prevent a mismatch between the nature of the observed phenomena and the mining algorithm. For instance, moving objects on a spatiotemporal network need to be studied from a traveler's perspective, i.e., the Lagrangian frame of reference [175–178] instead of a snapshot view. This is because a traveler moving along a chosen path in a spatiotemporal network would experience a road segment (and its properties such as fuel efficiency and travel time) for the time at which he/she arrives at that segment, which may be distinct from the original departure time at the start of the journey. These unique requirements (non-isotropy and Lagrangian reference frame) call for novel spatiotemporal statistical foundations [173] as well as new computational approaches for spatiotemporal network big data science.

Another future research need is to develop spatiotemporal graph big data platforms, motivated by the upcoming rich spatiotemporal network data collected from vehicles. Modern vehicles have rich instrumentation to measure hundreds of attributes at high frequency and are generating big data (Exabyte [179]). This vehicle measurement big data consists of a collection of trips on a transportation graph such as a road map annotated with several measurements of engine subsystems. Collecting and analyzing such big data during real-world driving conditions can aid in understanding the underlying factors which govern real-world fuel inefficiencies or high greenhouse gas emissions [180]. Current relevant big data platforms for spatial and spatiotemporal big data science include ESRI GIS Tools for Hadoop [181, 182] and Hadoop GIS [183]. These provide distributed systems for geometric data (e.g., lines, points, and polygons) including geometric indexing and partitioning methods such as R-tree, R+-tree, or Quad tree. Recently, SpatialHadoop has been developed [184]. SpatialHadoop embeds geometric notions in language, visualization, storage, MapReduce, and operations layers. However, spatiotemporal graphs violate the core assumptions of current spatial big data platforms that the geometric concepts are adequate for conveniently representing spatiotemporal

graph analytics operations and for partition data for load-balancing. Spatiotemporal graphs also violate core assumptions underlying graph analytics software (e.g., Giraph [185], GraphLab [186], and Pregel [187]) that traditional location-unaware graphs are adequate for conveniently representing STG analytics operations and for partition data for load-balancing. Therefore, novel spatiotemporal graph big data platforms is needed. Several challenges should be addressed; e.g., spatiotemporal graph big data requires novel distributed file system (DFS) to partition the graph, and a novel programming model is still needed to support abstract data types and fundamental spatiotemporal graphs operations, etc.

2.5 Summary

This chapter provides an overview of current research in the field of spatial and spatiotemporal (SST) big data science from a computational perspective. SST big data science has broad application domains including ecology and environmental management, public safety, transportation, earth science, epidemiology, and climatology. However, the complexity of SST data and intrinsic SST relationships limits the usefulness of conventional big data science techniques. We provide a taxonomy of different SST data types and underlying statistics. We also review common SST big data science techniques organized by major output pattern families: SST outlier, coupling and tele-coupling, prediction, partitioning and summarization, hotspots, and change patterns. Finally, we discuss the recent research trends and future research needs.

References

1. S. Shekhar, Z. Jiang, R.Y. Ali, E. Eftelioglu, X. Tang, V.M.V. Gunturi, X. Zhou, Spatiotemporal data mining: a computational perspective. ISPRS Int. J. Geo-Inf. **4**(4), 2306 (2015)
2. K. Koperski, J. Adhikary, J. Han, Spatial data mining: progress and challenges survey paper, in *Proceedings of ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada* (Citeseer, 1996), pp. 1–10
3. M. Ester, H.-P. Kriegel, J. Sander, Spatial data mining: a database approach, in *Proceedings of Fifth Symposium on Rules in Geographic Information Databases* (1997)
4. S. Shekhar, M.R. Evans, J.M. Kang, P. Mohan, Identifying patterns in spatial information: a survey of methods. Wiley Interdis. Rev. Data Min. Knowl. Disc. **1**(3), 193–214 (2011)
5. H.J. Miller, J. Han, *Geographic Data Mining and Knowledge Discovery* (Taylor & Francis Inc., Bristol, 2001)
6. H.J. Miller, J. Han, in *Geographic Data Mining and Knowledge Discovery* (CRC Press, 2009)
7. S. Shekhar, P. Zhang, Y. Huang, R.R. Vatsavai, Trends in spatial data mining, in *Data Mining: Next Generation Challenges and Future Directions* (2003), pp. 357–380
8. S. Kisilevich, F. Mansmann, M. Nanni, S. Rinzivillo, in *Spatio-Temporal Clustering* (Springer, Berlin, 2010)
9. C.C. Aggarwal, in *Outlier Analysis* (Springer Science & Business Media, 2013)
10. X. Zhou, S. Shekhar, R.Y. Ali, Spatiotemporal change footprint pattern discovery: an inter-disciplinary survey. Wiley Interdis. Rev. Data Min. Knowl. Disc. **4**(1), 1–23 (2014)