GeoJournal manuscript No.
(will be inserted by the editor)

Spatial Big Data: Platforms, Analytics, and Science

Daniel Cintra Cugler · Dev Oliver · Michael R. Evans · Shashi Shekhar · Claudia Bauzer Medeiros

Received: date / Accepted: date

Abstract Emerging non-traditional spatial datasets from geo-social media, sensor networks, and volunteers are important due to societal applications such as situation assessment after natural disasters, monitoring urban traffic, etc. However, such datasets, called spatial big data, often exceed the capacity of commonly used spatial computing platforms. Spatial big data presents new challenges for their capture, curation, analysis, exploration, and sharing. This paper provides an overview of the emerging ideas and research needs across different platforms, analytics, and science methodologies for spatial big data.

1 Introduction

Increasingly, the volume, velocity, and variety of spatial datasets exceed the capacity of commonly used spatial computing and spatial database technologies to learn, manage, and process the data with reasonable effort. We refer to these datasets as Spatial Big Data (SBD). Examples of emerging SBD datasets include GPS trace data from cell-phones, geo-social media (e.g., tweets), temporally detailed (TD) roadmaps that provide speeds every minute for every road-segment, and engine measurements of fuel consumption, greenhouse gas (GHG) emissions, etc.

A 2011 McKinsey Global Institute report estimates savings of “about $600 billion annually by 2020” in terms of fuel and time saved [60] by helping vehicles avoid congestion and reduce idling at red lights or left turns. Preliminary evidence for the transformative potential includes the experience of UPS, which saves millions of gallons of fuel by simply avoiding left turns and associated engine-idling when selecting routes [54]. Immense savings in fuel-cost and GHG emission are possible in the future if other fleet owners and consumers avoided left-turns and other hot spots of idling, low fuel-efficiency, and congestion.

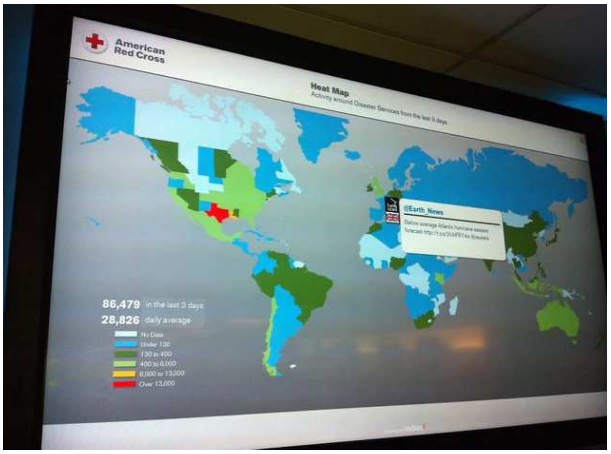


Fig. 1 Red Cross’ new social media monitoring center leveraging social media for disaster monitoring [38]

Disaster monitoring is also leveraging geo-social media, e.g., tweets as illustrated in Figure 1. Even before cable news outlets began reporting the tornadoes that ripped through Texas on Tuesday, a map of the state

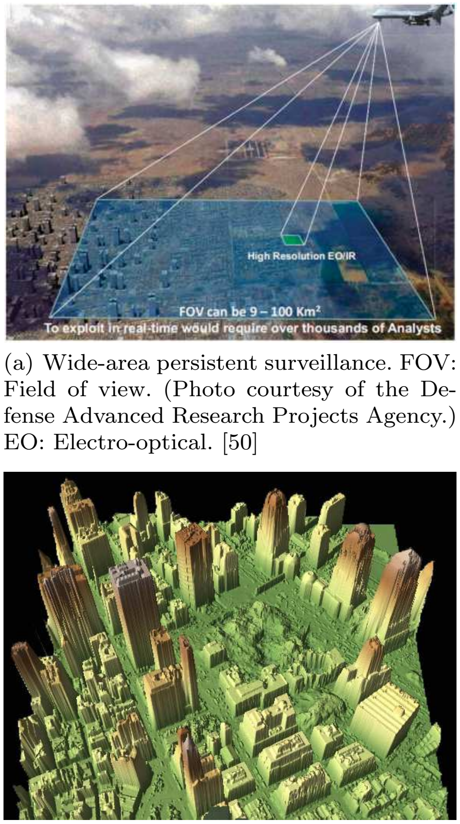
D. C. Cugler and C.B. Medeiros
Institute of Computing
University of Campinas
13.083-970 - Campinas - SP - Brazil
E-mail: {danielcugler,cmbm}@ic.unicamp.br

D. Oliver, M. R. Evans and S. Shekhar
Department of Computer Science
University of Minnesota
55455 - Minneapolis - MN - USA
E-mail: {oliver,mevans,shekhar}@cs.umn.edu

began blinking red on a screen in the Red Cross' new social media monitoring center [38], alerting weather watchers that something was happening in the hard-hit area.

1.1 Types of Spatial Big Data

Spatial datasets are discrete representations of typically continuous phenomena. Discretization of continuous space is necessitated by the nature of digital representation. There are three basic models to represent spatial data, namely, raster (grid), vector, and graph. Satellite images are good examples of raster data. Vector data consists of points, lines, polygons and their aggregate (or multi-) counterparts. Graphs consisting of spatial networks are the third important data type. Spatial Big Data can also be represented via these basic models, only the datasets have become much bigger and richer.


High Resolution EO/IR
FOV can be 9 – 100 Km²
To exploit in real-time would require over thousands of Analysts

(a) Wide-area persistent surveillance. FOV: Field of view. (Photo courtesy of the Defense Advanced Research Projects Agency.) EO: Electro-optical. [50]

(b) LIDAR images of ground zero rendered Sept. 27, 2001 by the U.S. Army Joint Precision Strike Demonstration from data collected by NOAA flights. Thanks to NOAA/U.S. Army JPSD.

Fig. 2 Raster SBD Examples (Best viewed in color)

Raster data, such as geo-images (Google Earth), are frequently used for remote sensing and land classification. New Spatial Big Raster datasets are emerging from a number of sources.

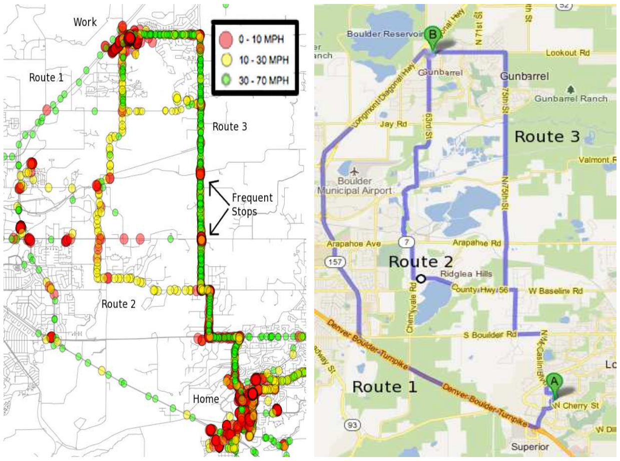
WAMI Data: Wide area motion imagery (WAMI) sensors are increasingly being used for persistent surveillance of large areas, including densely populated urban areas. The wide-area video cover-age and 24/7 persistence of these sensor systems allow for new and interesting patterns to be found via temporal aggregation of information. However, there are several challenges associated with using unmanned aerial vehicles (UAVs) in gathering and managing raster datasets. First, a UAV has a small footprint due to the relatively low flying height; therefore, it has to capture a large amount of images in a very short period of time to achieve the spatial coverage for many applications. Image processing is another challenge because it would be too time consuming and costly to rectify and mosaic the UAV photography for large areas. The large quantity of data far exceeds the capacity of the available pool of human analysts [71]. It is essential to develop automated, efficient, and accurate techniques to handle these spatial big data.

LiDAR: Lidar (Light Detection and Ranging or Laser Imaging Detection and Ranging) data is generated by timing laser pulses from an aerial position (plane or satellite) over a selected area to produce a surface mapping [73]. Lidar data are very rich to analyze surfaces or extract features. These large volumes of data pose a big challenge on management, analysis, and timely accessibility. Particularly, Lidar points and their attributes have tremendous sizes, making it difficult to categorize these datasets for end-users. Therefore, Spatial Big Data is an essential issue for Lidar remote sensing.

Vector data over space is a framework to formalize specific relationships among a set of objects. Traditional vector data consists of points, lines and polygons; and with the rise of Spatial Big Data, corresponding datasets have arisen from a variety of sources.

VGI Data: Volunteered geographic information (VGI) brings a new notion of infrastructure to collect, synthesize, verify, and redistribute geographic data through geo-location technology, mobile devices, and geo-databases. These geographic data are provided, modified, and shared through user interactive online services (e.g., OpenStreetMap, Wikimapia, GoogleMap, GoogleEarth, Microsofts Virtual Earth, Flickr, etc.). In recent years, VGI has lead to an explosive growth in the availability of user-generated geographic information and requires bigger storage models to handle large scale spatial datasets. The challenge for VGI is to enhance data service quality regard to accuracy, credibility, reliability, and overall value [38].

GPS Trace Data: GPS trajectories are quickly becoming available due to the rapid proliferation of cellphones, in-vehicle navigation devices, and other GPS data-logging devices [26] such as those distributed by insurance companies [94]. GPS traces make it possible to provide personalized route suggestions to users to reduce fuel consumption and GHG emissions. For example, Figure 3 shows 3 months of GPS trace data from a commuter with each point representing a GPS record taken at 1 minute intervals, 24 hours a day, 7 days a week. As can be seen, 3 alternative commute routes were identified between home and work from this dataset. These routes may be compared for engine idling which are represented by darker (red) circles. Assuming the availability of a model to estimate fuel consumption from speed profiles, one may even rank alternative routes for fuel efficiency. In recent years, consumer GPS products [26,91] are evaluating the potential of this approach. Again, a key hurdle is the dataset size, which can reach 10^{13} items per year given constant minute-resolution measurements for all 100 million US vehicles.



(a) GPS trace data. Color indicates speed.

(b) Routes 1, 2, & 3 [32].

Fig. 3 A commuter's GPS tracks over three months reveal preferred routes. (Best viewed in color)

Graph data is commonly used in spatial computing to represent road maps for routing queries. While the network structure of the graph may not be changing, the amount of information about the network is rising dramatically. New temporally-detailed road maps give minute by minute speed information, along with elevation and engine measurements to allow for more sophisticated querying of road networks.

Spatio-Temporal Engine Measurement Data: Many modern fleet vehicles include rich instrumentation such as GPS receivers and sensors to periodically measure sub-system properties [41,42,57,61,89,90], as well as auxiliary computing, storage and communication devices to log and transfer accumulated datasets. Engine measurement datasets may be used to study the impacts of the environment (e.g., elevation changes, weather), vehicles (e.g., weight, engine size, energy-source), traffic management systems (e.g., traffic light timing policies), and driver behaviors (e.g., gentle acceleration or braking) on fuel savings and GHG emissions. These datasets may include a time-series of attributes such as vehicle location, fuel levels, vehicle speed, odometer values, engine speed in revolutions per minute (RPM), engine load, emissions of greenhouse gases (e.g., CO2 and NOX), etc. Fuel efficiency can be estimated from fuel levels and distance traveled as well as engine idling from engine RPM. These attributes may be compared with geographic contexts such as elevation changes and traffic signal patterns to improve understanding of fuel efficiency and GHG emission. For example, Figure 4 shows heavy truck fuel consumption as a function of elevation from a recent study at Oak Ridge National Laboratory [15]. Notice how fuel consumption changes drastically with elevation slope changes. Fleet owners have studied such datasets to fine-tune routes to reduce unnecessary idling [6,5]. It is tantalizing to explore the potential of this dataset to help consumers gain similar fuel savings and GHG emission reduction. However, these datasets can grow big. For example, measurements of 10 engine variables, once a minute, over the 100 million US vehicles in existence [86,25], may have 10^{14} data-items per year.

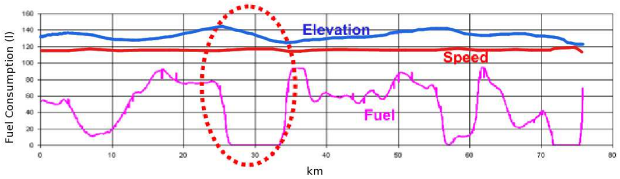
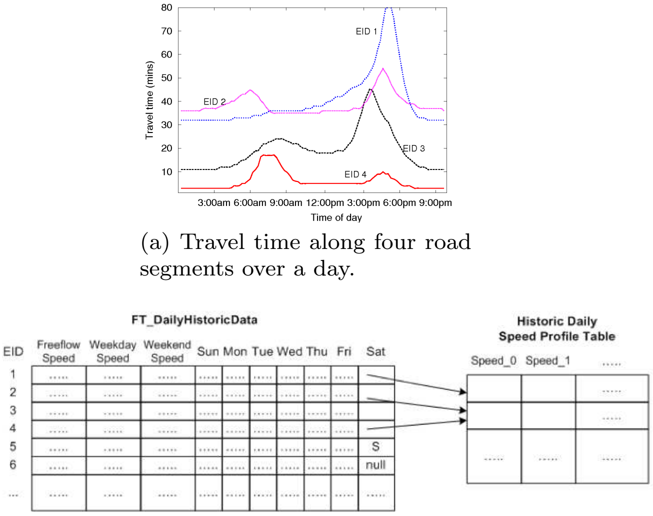


Fig. 4 Engine measurement data improve understanding of fuel consumption [15]. (Best in color)

Temporally-Detailed (TD) Roadmaps: Traditionally, digital road maps consisted of center lines and topologies of the road networks [28,93]. These maps are used by navigation devices and web applications such as Google Maps [32] to suggest routes to users. New datasets from companies such as NAVTEQ [67], use probe vehicles and highway sensors (e.g., loop detectors) to compile travel time information across road segments for all times of the day and week at fine temporal resolutions (seconds or minutes). This data is applied to a profile model, and patterns in the road speeds are identified throughout the day. The profiles have data for every five minutes, which can then be applied to the road segment, building up an accurate picture of speeds based on historical data. Such TD roadmaps contain much more speed information than traditional digital roadmaps. While traditional roadmaps have only one scalar value of speed for a given road segment (e.g., EID 1), TD roadmaps may potentially list speed/travel time for a road segment (e.g., EID 1) for thousands of time points (Figure 5(a)) in a typical week. This allows a commuter to compare alternate start-times in addition to alternative routes. It may even allow comparison of (start-time, route) combinations to select distinct preferred routes and distinct start-times. For example, route ranking may differ across rush hour and non-rush hour and in general across different start times. However, TD roadmaps are big and their size may exceed 10^{13} items per year for the 100 million road-segments in the US when associated with per-minute values for speed or travel-time.



(a) Travel time along four road segments over a day.

(b) Schema for Daily Historic Speed Data.

Fig. 5 Spatial Big Data on Historical Speed Profiles. (Best viewed in color)

2 Disruption caused by SBD

SBD poses challenges for which ordinary computing technologies and methodologies may not suffice. SBD has changed the way scientists do research, posing new opportunities in SBD platforms, SBD analytics and SBD science methodologies.

2.1 Platform

The management of large volumes of spatial data is a big challenge for several applications, such as satellite monitoring systems, intelligent transportation systems, and disaster management systems. These applications need to provide a solution to the increasing data demands and offer a shared, distributed computing infrastructure as well as a reliable system. The complexity and nature of spatial datasets makes them ideal for applying parallel processing. Recently, the concept of a cloud environment has been introduced to provide a solution for these requirements. Existing approaches and solutions provide a general framework for distributed file systems (e.g., Google file [29] system and HDFS [12]) and process these data sets based on replicas of data blocks (e.g., map-reduce [18] and Hadoop [12]). Column-oriented database systems have also been introduced to support OLAP or join processing (e.g., MongoDB and HBase).

However, generalizing the infrastructure to handle spatial problems is challenging. The reason is that many spatial problems require multiple iterations, which challenges frameworks such as MapReduce. Examples of spatial problems requiring multiple iterations are hotspot queries, co-location mining, spatial autoregression (SAR), etc.

2.2 Analytics

Spatial Big Data provides the opportunity to solve some of the long-standing challenges in spatial data analytics which stemmed from lack of data, such as estimating spatial neighborhood relationships, supporting place-based ensemble models, simplifying spatial models, and on-line spatio-temporal data analytics.

SDB creates an opportunity to simplify spatial models in traditional spatial data mining (SDM). It may be the case that some of the complexity from SDM is due to the paucity of data at individual places which in turn forces one to leverage data at nearby places via spatial autocorrelation and spatial joins.

SDB may be large enough to provide a reliable estimate of neighborhood relationships. This may ultimately relieve user burden and may improve model accuracy. Traditional assumptions might not have to be made such as limited interaction length (e.g., the Markov assumption), spatially invariant neighbor relationships (e.g., the eight-neighborhood contiguity matrix), and tele-connections derived from short-distance relationships.

SBD may support a Place-based ensemble of models beyond *geographically weighted regression* (GWR). Examples include place-based ensembles of decision trees for land-cover classification and place-based ensembles of spatial auto-regression models. The computational challenge stems from the fact that naive approaches may run a learning algorithm for each place. Reducing the computation cost by exploiting spatial autocorrelation is an interesting possibility that will need to be explored further.

2.3 Science

The classical scientific methodologies, e.g., theories and experiments [27] have been used for centuries. The traditional experimental methodology comprises the following activities: a) Formulation of a question; b) Creation of hypotheses; c) Execution of inferences/prediction; d) Execution of tests/controlled experiments and e) Execution of statistical tests to support the hypothesis. However, traditional methods may be time consuming due to the manual creation of hypotheses and manual experiments. SBD may speed-up hypothesis generation by helping to generate hypotheses, thereby complementing manual hypothesis generation. For example, big data can be used to detect unreported side effects of combination of drugs. The common way physicians know about this kind of side effects is based on previous physicians' reports. However, this technique is limited, since a new side effect is generated only when a physician notices something and reports it. In order to overcome such issues, scientists are getting from big data the opportunity of combining data-mining techniques with millions/billions of users' queries from web search engines in order to provide evidence of unreported drug side effects. For example, a search containing the words *paroxetine, pravastatin* and *blurry vision* might suggest that the combination of such two drugs can cause high blood sugar. This kind of technique poses several science challenges, such as integrating new sources of data while protecting user privacy [72].

3 SBD Platform

3.1 Background

Recently, the concept of a cloud computing [97] environment has been introduced to provide a solution for these requirements. One of the most popular free and open source solutions is the Apache Hadoop [1] framework that provides an environment in which programmers do not have to deal with issues of parallelization, remote execution, data distribution, load balancing, or fault tolerance. Hadoop libraries have been created for several programming languages, such as Java, Perl, Python and PHP. Apache Hadoop also provides the Hadoop Distributed File System (HDFS) that enables high aggregate bandwidth across the cluster as well as reliable storage of very large files across machines in a large cluster. Figure 6 shows the Intel distribution for Apache Hadoop software components [39]. The figure shows many components running on top of the HDFS for distributed processing (MapReduce), workflow (Oozie), scripting (Pig), machine learning (Mahout), sql queries (Hive), and column store (HBase).

MapReduce is one of the most popular big data platforms. It is a programming model for processing and generating large datasets. MapReduce was created at Google in 2004 and it was inspired by the map and reduce primitives present in Lisp and many other functional programming languages. Software developed to run in MapReduce platforms divide a problem into many small parts, each of which may be executed in any node of a cluster. The core of MapReduce are the map and reduce functions, that are responsible to parallelize computation across large-scale clusters of machines, handles machine failures, and schedules inter-machine communication to make efficient use of the network and disks. These features allow programmers without low or none experience with parallel and distributed systems to utilize the resources of a large distributed system and process huge amounts of data, e.g., terabytes and petabytes [19]. Both Hadoop MapReduce and HDFS are designed so that node failures are automatically handled by the framework.

3.2 Limitations

A key problem with current big data platforms such as MapReduce is the difficulty of efficiently parallelizing existing iterative algorithms. Many spatial data mining algorithms (e.g., colocation mining, parameter estimation for spatial autoregression) and spatial graph algorithms (e.g., breadth-first search and shortest path) use previous information for the next iteration. Although processing one iteration is parallelizable, the synchronization overhead for a MapReduce environment is too enormous to sustain parallelism across multiple iterations. In addition, HDFS is designed for large sequential access, i.e., scanning of large files. These are not effective for queries retrieving a small portion of a data file. Future work should include non-iterative algorithms or different parallel programming models.

3.3 Future Directions

Three directions emerge when trying to address the limitations of current systems. First, research on alterna-

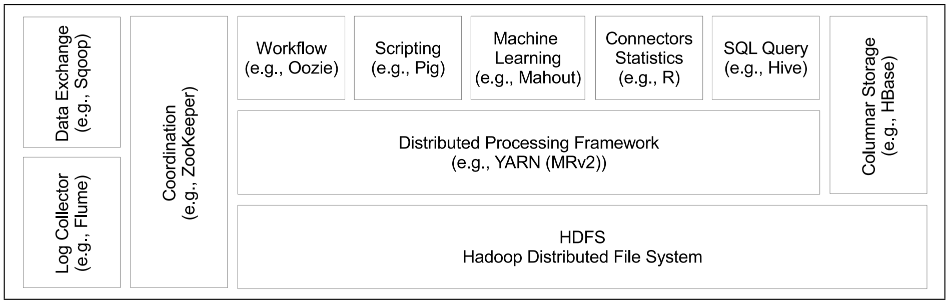


Fig. 6 Intel Distribution for Apache Hadoop software components [39]

tives to MapReduce is needed to address some of the emerging challenges that spatial big data raises (e.g., the need to iterate multiple times). Initial efforts in this vein include Pregel [59], GraphLab [55], PowerGraph [31], HaLoop [14], PrIter [98], and CIEL [65], which focus on large-scale, fault-tolerant graph or iterative computing. Second, provision of spatial indexes (e.g., R-trees, distributed partitioned R-trees) helps improve the I/O cost of queries retrieving a small part of the data file. Representative efforts include SpatialHadoop [22], which is a MapReduce extension to Apache Hadoop designed specially to work with spatial data by providing specialized spatial data types, spatial indexes, and spatial operations. Third, the design of non-iterative algorithms for spatial data mining and spatial statistical parameter estimation may make an interesting alternative research direction. Which spatial problems can have non-iterative solutions even if those are solved using iterative algorithms today? What are effective ways to reformulate inherently iterative approaches such as colocation mining or paramter estimation for spatial autoregression so that multiple iterations are not needed, while maintaining correctness and completeness?

4 SBD Analytics

Traditional space-time approaches have been useful for predicting location, time, and paths for different use cases such as predicting bird nest sites, minerals, hurricanes, tornadoes, etc. [79,4,53,40,85,87,52,83,7,13]. Spatial data does not follow traditional assumptions of independent and identical distribution due to properties such as spatial autocorrelation (Tobler's Law, Markov Model), heterogeneity (Geographically Weighted Regression) [79,13], and edge-effects [13].

Space-Time models are often computationally more expensive than traditional models. For example, spatial auto-regression requires more computing power due to the fact that W is quadratic in the number of locations. Geographically weighted regression has the same limitation as opposed to classical linear regression, also due to the inclusion of the W matrix. Colocation pattern mining, which finds the subsets of features frequently located together is more computationally expensive that traditional association rule mining ([3]) and confidence estimation adds more costs (e.g., M.C.M.C. simulations).

4.1 Computational View

The Holy Grail of SBD analytics from a computational standpoint is to scale up traditional methods by several orders of magnitude. One way to achieve this is to port traditional machine learning and data mining techniques to Hadoop/MapReduce (or alternatives discussed in Section 3). Examples include Apache Mahout [9], that provides core algorithms for clustering, classfication and batch based collaborative filtering, as illustrated in Table 1 (left column). Spatial statistics techniques (e.g., hotspot detection and others listed in Table 1 (right column)) may also be scaled up in a similar way (e.g., ESRI's GIS on Hadoop [23]). One has to take into account computational considerations such as space partitioning, multidimensional data structures, static and dynamic load balancing, etc. SBD analytics algorithms usually require multiple iterations, which challenges MapReduce. Computational issues may also arise due to high dimensionality of the spatial data set, spatial join process required in co-location mining and spatial outlier detection, estimation of SAR model parameters in the presence of large neighborhood matrix W, etc.

4.2 Statistical View

Since the early 1900s, society has depended on using representative samples [63] when faced with large numbers. The idea was to extrapolate unbiased information from a small representative sample about the general population and make appropriate inferences about a few statistics (e.g., mean) on the general population.

Table 1 Representative Analytics Techniques in the Hadoop Environment

MAHOUT [9]	SBD ANALYTICS ON HADOOP
Collaborative Filtering	Spatial Join and Index (supporting statistics and data mining)
User and Item based recommenders	Raster data support - fast zonal statistics
K-Means, Fuzzy K-Means clustering	Kernel density estimations
Mean Shift clustering	Kriging and Bayesian Kriging
Dirichlet process clustering	Neighborhood relationship support
Latent Dirichlet Allocation	Topological functions (overlap, union, intersect)
Singular value decomposition	High-volume distance calculations
Parallel Frequent Pattern mining	Hotspot queries
Complementary Naive Bayes classifier	Spatial regression (SAR)
Random forest decision tree based classifier	

When the assumptions (e.g., representative unbiased samples, central limit theorem) hold, statistical sampling methods provide a cost-effective method to estimate overall population behavior (e.g., mean, standard deviation) while controlling previously identified bias. However, if further questions arise on subgroups (e.g., defined by gender, age, race, etc.), one needs to design new surveys, draw new samples, and analyze new datasets, which often takes a long time.

Random sampling does not scale easily to include subcategories and increases the possibility of erroneous predictions. In contrast, SBD (e.g., census) enables drill down and roll up to facilitate analysis of all facets of the data. In addition, edge effects may be pronounced in random samples. For example, the convex hull of a random sample will not equal the true convex hull of a population (e.g., street address of patients). SBD may complement representative samples to improve the estimate of a convex hull. Spatial big data may also reduce Simpson's Paradox effects in spatio-temporal data where trends that appear in different groups of data disappears or reverses when these groups are combined.

4.3 Limitations

The volume, variety, and velocity of data, the complexity of spatial big data types and relationships, and the need to identify spatial autocorrelation poses numerous computational challenges in SBD analytics.

First, SBD may expose weaknesses in complex models. *"There are always implicit assumptions behind a model and its solution method. But human beings have limited foresight and great imagination, so that, inevitably, a model will be used in ways its creator never intended. This is especially true in trading environments... but its also a matter of principle: you just cannot foresee everything. So, even a "correct" model, "correctly" solved, can lead to problems. The more complex the model, the greater this possibility"*. (Emanuel Derman 1996) [20].

A fundamental limitation of traditional spatial data mining is off-line batch processing where spatial models are usually not learned in real time (e.g., spatial autoregression, colocation pattern mining, and hotspot detection). However, SBD includes streaming data such as event reports and sensor measurements. Furthermore, the use cases for SBD include monitoring and surveillance which requires on-line algorithms. Examples of such applications include 1) the timely detection of outbreak of disease, crime, unrest and adverse events, 2) the displacement or spread of a hotspot to neighboring geographies, and 3) abrupt or rapid change detection in land cover, forest-fire, etc. for quick response. Models that are purely local may leverage time-series data analytics models but regional and global models are more challenging. For spatial interactions (e.g., colocations and tele-connections) with time-lags, SBD may provide opportunities for precisely computing them in an on-line manner. If precise on-line computation is not possible, SBD might be useful in providing on-line approximations.

4.4 Future Directions

New research directions in SBD analytics include simpler models, online SBD analytics, and new SBD pattern families.

4.4.1 Bigger the SBD, Simpler the Model

The availability of spatial big data provides an opportunity to forego complicated models in favor of simpler ones, i.e., the bigger the SBD, the simpler the model. An example of a simple model is the nearest neighbor technique where a data element is classified by a majority vote of its neighbors, with the element being assigned to the class most common amongst its nearest neighbors. This has wide application in classification and recommender systems (e.g., Facebook, Amazon).

4.4.2 On-line SBD Analytics

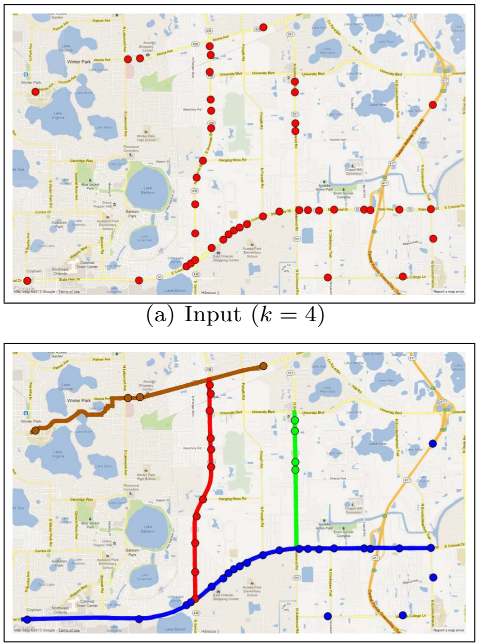
A fundamental limitation of traditional space-time data mining is off-line batch processing where space-time models are usually not learned in real time (e.g., spatial auto-regression, colocation pattern mining, and hotspot detection). However, SBD includes streaming data such as event reports and sensor measurements. Furthermore, the use cases for SBD include monitoring and surveillance which requires on-line algorithms. Examples of such applications include 1) the timely detection of outbreak of disease, crime, unrest and adverse events, 2) the displacement or spread of a hotspot to neighboring geographies, and 3) abrupt or rapid change detection in land cover, forest-fire, etc. for quick response.

Models that are purely local may leverage time-series data analytics models but regional and global models are more challenging. For spatial interactions (e.g., colocations and tele-connections) with time-lags, SBD may provide opportunities for precisely computing them in an on-line manner. If precise on-line computation is not possible, SBD might be useful in providing on-line approximations.

4.4.3 Additional Opportunities

Spatial big data provides new opportunities (previously not possible due to low data resolution) for estimating spatial neighbor relationships and supporting place-based ensemble models.

Additionally, emerging, novel patterns in spatial big data include summarization, obfuscated outliers, rare associations, and obfuscated event / process prediction. For example, summarization (i.e., finding a compact description or representation of a dataset) may be an important first step for understanding SBD (e.g., by providing novel GPS track and route-based summaries). Figure 7(a) shows a summarization example where the input consists of pedestrian fatalities in Orlando, Florida and the output (shown in Figure 7(b)) uses paths and network distance to summarize or group activities on the spatial network. An example of obfuscated outliers is shown in Figure 8. Obfuscated outliers reveal exceptions to patterns of life and SBD raises questions such as why is the highlighted block avoided by cyclists? Table 2 summarizes example pattern families from traditional data mining and spatial data mining, and contrasts them with newly emerging patterns in spatial big data analytics.


(a) Input ($k = 4$)

(b) KMR

Fig. 7 Summarizing on pedestrian fatality data from Orlando, FL ([24]).

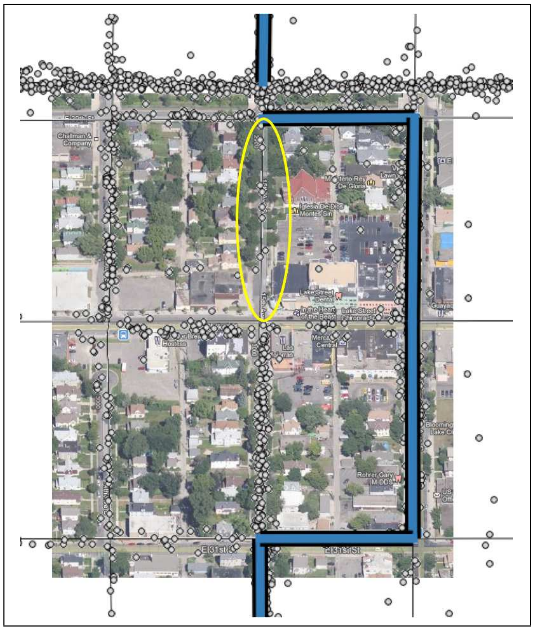


Fig. 8 Obfuscated outliers reveal exceptions to patterns of life. SBD raises questions such as why is the highlighted block avoided by cyclists? (Best in color)

5 Science Methodologies leveraging SBD

“Sudden influxes of data have transformed researchers’ understanding of nature before even back in the days when ’computer’ was still a job description. Unfortunately, the institutions and culture of science remain rooted in that pre-electronic era. Taking full advantage

Table 2 New Pattern Families emerging from Spatial Big Data Analytics

Data Mining	**Spatial Data Mining**	**Spatial Big Data Analytics**
Clustering [16,43,36,74,2,49,58,68,77,76]	Hotspots, Emerging/Spreading Hotspots [21,56,51,47,69,48,70,44]	Summarization
Outlier Detection [92,33,10,34]	Spatial Outlier, Abrupt Change [56,78,88,96,81,17,8,82,95]	Obfuscated Outliers
Association Rule Mining [3]	Colocation, Teleconnection/Co-occurrence [46,64,37,80,62]	Rare Associations
Prediction/Classification [79,4,53,40,85,87,52,83,7,13]	Location Prediction, Event/Process Prediction [79,13]	Obfuscated Event / Process Prediction

of electronic data will require a great deal of additional infrastructure, both technical and cultural" [66]

5.1 Background

Science methodologies have traditionally included manual models such as theory (e.g., differential equations) or experiments (e.g., hypothesis testing, random sampling, correlation, and regression). Computer-assisted forward models later emerged which included computational simulations using differential equations, agent-based models, etc. The 2013 Chemistry Nobel Prize recognized work on computational modeling of atoms and molecules using high performance computing [75]. Recently, scientific methodology experienced the computer-assisted backward model with ideas such as the fourth paradigm [35] and A-B testing. The fourth paradigm suggests that big data may help generate hypothesis (to complement manual hypothesis generation) via spatial statistics, machine learning, spatial data mining, and geo-visual analytics.

Table 3 Scientific Methodologies

Models	**Traditional (Manual)**	**Computer Assisted**
Forward	Theory	Simulations
Backward	Experiments	1. Data-intensive Hypothesis Generation (4th Paradigm) 2. A-B Experiments on the web

Science methodologies have seen hypothesis generation moving from manual to data driven. An example of data driven hypothesis generation can be seen in the London Cholera outbreak, where the prevailing hypothesis of spread of cholera via air was changed to a new hypothesis of spread of cholera via water [84], which led to germ theory a few decades later. Current examples of big-data driven hypothesis generation can be seen at Microsoft, Stanford and Columbia University were scientists were able to detect evidence of unreported prescription drug side effects before they were found by the Food and Drug Administrations warning system by examining millions of web search queries [72]. Google has also been able to estimate flu activity through the use of certain search terms [30].

A-B testing [45] of candidate hypothesis about world wide web phenomena (e.g., advertising) is gaining popularity in industry. A-B testing is a methodology of using randomized experiments with two variants A and B, which are the control and treatment in the controlled experiment. A-B testing is important in domains such as advertising where the impact on desired key performance indicators (KPIs) or outcomes is measured based on the control and treatment (e.g., click through rate, estimated revenue). Table 3 summarizes the various scientific methodologies, including the fourth paradigm.

5.2 Limitations and Future Directions

Science limitations stem from three major sources. First, balancing the tradeoff between timeliness and bias in spatial big data is essential. SBD may be biased because only information about certain locations, potentially gleaned from the world wide web (WWW) or volunteered geographic information (VGI), may be available. Information about certain areas may be readily available, which may provide a wealth of information for scientific analysis whereas information about other areas may be non-existent thereby potentially creating bias in the overall analysis.

Second, reproducibility of spatial big data is another limitation due to private ownership of many SBDs related to web searches and social media (e.g., Facebook, tweets). Data sources that contain such kind of private data is known as the deep web [11]. Important questions arise such as private ownership of big data (e.g., from a Google search). Spatial big data also raises challenges related to provenance and data quality for citizen science.

Third, establishing institutional review boards for VGI is a major challenge. What would be the best way to perform a risk-benefit analysis for potentially millions of volunteers in a non-controlled setting to determine whether or not research should be done?

6 Conclusion

Spatial big data is emerging from numerous sources (e.g., GPS data from cellphones, UAV data, etc.) as a valuable resource for many domains including eco-routing, public safety, and climate science. However, its size, variety, and update rate exceeds the capabilities of existing computing systems to learn, manage, and process the data with reasonable effort. New opportunities are emerging from the platforms, analytics, and science perspectives. However, SBD research is still largely unexplored territory with alternatives to MapReduce for SBD, the design of non-iterative algorithms for SBD, simpler models, online SBD analytics, new SBD pattern families, and data driven hypothesis generation emerging as new frontiers for the future.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant No. 1029711, IIS-1320580, 0940818 and IIS-1218168 as well as the USDOD under Grant No. HM1582-08-1-0017, and HM0210-13-1-0005. Work was also partially financed by FAPESP under Grant No. 2011/19284-3 and 2012/11395-3, the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), and INCT in Web Science. We would like to thank Kim Koffolt and the members of the University of Minnesota Spatial Computing Research Group for their comments.

References

1. Apache Hadoop. URL \url{http://hadoop.apache.org/}
2. 0010, W.W., Yang, J., Muntz, R.R.: Sting: A statistical information grid approach to spatial data mining. In: M. Jarke, M.J. Carey, K.R. Dittrich, F.H. Lochovsky, P. Loucopoulos, M.A. Jeusfeld (eds.) VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece, pp. 186–195. Morgan Kaufmann (1997)
3. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB, vol. 1215, pp. 487–499 (1994)
4. de Almeida, C.M., Souza, I.M., Alves, C.D., Pinho, C.M.D., Pereira, M.N., Feitosa, R.Q.: Multilevel object-oriented classification of quickbird images for urban population estimates. In: GIS, p. 12 (2007)
5. American Transportation Research Institute (ATRI): Atri and fhwa release bottleneck analysis of 100 freight significant highway locations (2010). URL \url{http://goo.gl/CONuD}
6. American Transportation Research Institute (ATRI): Fpm congestion monitoring at 250 freight significant highway location: Final results of the 2010 performance assessment (2010). URL \url{http://goo.gl/3cAjr}
7. Anselin, L.: Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dorddrecht (1988)
8. Anselin, L., Getis, A.: Spatial statistical analysis and geographic information systems. In: Perspectives on Spatial Data Analysis, pp. 35–47. Springer (2010)
9. Apache: Mahout (October 9, 2013). URL \url{http://mahout.apache.org/}
10. Barnett, V., Lewis, T.: Outliers in statistical data. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, Chichester: Wiley, 1984, 2nd ed. **1** (1984)
11. Bergman, M.K.: The deep web: Surfacing hidden value. Journal of electronic publishing **7**(1), 07–01 (2001)
12. Borthakur, D.: The hadoop distributed file system: Architecture and design. Hadoop Project Website **11**, 21 (2007)
13. Brunsdon, C., Fotheringham, S., Charlton, M.: Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician) **47**(3), 431–443 (1998)
14. Bu, Y., Howe, B., Balazinska, M., Ernst, M.D.: Haloop: Efficient iterative data processing on large clusters. Proceedings of the VLDB Endowment **3**(1-2), 285–296 (2010)
15. Capps, G., Franzese, O., Knee, B., Lascurain, M., Otaduy, P.: Class-8 heavy truck duty cycle project final report. ORNL/TM-2008/122 (2008)
16. Cervone, G., Franzese, P., Ezber, Y., Boybeyi, Z.: Risk assessment of atmospheric hazard releases using k-means clustering. In: ICDM Workshops, pp. 342–348 (2008)
17. Cressie, N.A.C.: Statistics for Spatial Data. Wiley (1993)
18. Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. Communications of the ACM **51**(1), 107–113 (2008)
19. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Communications of the ACM **51**(1), 107–113 (2008)
20. Derman, E.: Model risk: What are the assumptions made in using models to value securities and what are the consequent risks? RISK-LONDON-RISK MAGAZINE LIMITED- **9**, 34–38 (1996)
21. van Eck, N.J., Waltman, L.: Bibliometric mapping of the computational intelligence field. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **15**(5), 625–645 (2007)
22. Eldawy, A., Mokbel, M.F.: A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data. In: 39th Proceedings of the International Conference on Very Large Databases (VLDB) (2013)
23. ESRI: Breathe Life into Big Data: ArcGIS Tools and Hadoop Analyze Large Data Stores. URL \url{http://www.esri.com/esri-news/arcnews/summer13articles/breathe-life-into-big-data}
24. Fatality Analysis Reporting System (FARS): (May 7, 2013). URL \url{http://www.nhtsa.gov/FARS}. National Highway Traffic Safety Administration (NHTSA)
25. Federal Highway Administration: Highway Statistics. HM-63, HM-64 (2008)
26. Garmin: www.garmin.com/us/ (May 7, 2013). URL \url{http://www.garmin.com/us/}
27. Gauch, H.G.: Scientific method in practice. Cambridge University Press (2003)
28. George, B., Shekhar, S.: Road maps, digital. In: Encyclopedia of GIS, pp. 967–972. Springer (2008)
29. Ghemawat, S., Gobioff, H., Leung, S.: The google file system. In: ACM SIGOPS Operating Systems Review, vol. 37, pp. 29–43. ACM (2003)

30. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature **457**(7232), 1012–1014 (2008)
31. Gonzalez, J.E., Low, Y., Gu, H., Bickson, D., Guestrin, C.: Powergraph: Distributed graph-parallel computation on natural graphs. In: Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI), pp. 17–30 (2012)
32. Google Maps: maps.google.com (May 7, 2013). URL `\url{http://maps.google.com}`
33. Hawkins, D.: Identification of Outliers. Chapman and Hall (1980)
34. Hawkins, D.M.: Identification of outliers, vol. 11. Chapman and Hall London (1980)
35. Hey, A., Tansley, S., Tolle, K.: The fourth paradigm: data-intensive scientific discovery. Microsoft research Redmond, WA (2009)
36. Hu, T., Xiong, H., Gong, X., Sung, S.Y.: Anemi: An adaptive neighborhood expectation-maximization algorithm with spatial augmented initialization. In: PAKDD, pp. 160–171 (2008)
37. Huang, Y., Shekhar, S., Xiong, H.: Discovering colocation patterns from spatial data sets: a general approach. Knowledge and Data Engineering, IEEE Transactions on **16**(12), 1472–1485 (2004)
38. InformationWeek: Red Cross Unveils Social Media Monitoring Operation (May 7, 2013). URL `\url{http://www.informationweek.com/government/information-management/red-cross-unveils-social-media-monitorin/232602219}`
39. Intel: Intel Distribution for Apache Hadoop Software (October 9, 2013). URL `\url{http://hadoop.intel.com/pdfs/IntelDistributionProductBrief.pdf}`
40. Jhung, Y., Swain, P.H.: Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. IEEE Transaction on Pattern Analysis and Machine Intelligence **34**(1), 67–75 (1996)
41. Kargupta, H., Gama, J., Fan, W.: The next generation of transportation systems, greenhouse emissions, and data mining. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1209–1212. ACM (2010)
42. Kargupta, H., Puttagunta, V., Klein, M., Sarkar, K.: Onboard vehicle data stream monitoring using minefleet and fast resource constrained monitoring of correlation matrices. New Generation Computing **25**(1), 5–32 (2006). Springer
43. Karypis, G., Han, E.H., Kumar, V.: Chameleon: Hierarchical clustering using dynamic modeling. IEEE Computer **32**(8), 68–75 (1999)
44. Kisilevich, S., Mansmann, F., Nanni, M., Rinzivillo, S.: Spatio-temporal clustering. In: Data Mining and Knowledge Discovery Handbook, pp. 855–874 (2010)
45. Kohavi, R., Longbotham, R., Sommerfield, D., Henne, R.M.: Controlled experiments on the web: survey and practical guide. Data Mining and Knowledge Discovery **18**(1), 140–181 (2009)
46. Koperski, K., Han, J.: Discovery of Spatial Association Rules in Geographic Information Databases. In: Proc. Fourth International Symposium on Large Spatial Databases, Maine. 47-66 (1995)
47. Kulldorff, M.: A spatial scan statistic. Communications in Statistics-Theory and methods **26**(6), 1481–1496 (1997)
48. Kulldorff, M., Athas, W., Feurer, E., Miller, B., Key, C.: Evaluating cluster alarms: a space-time scan statistic and brain cancer in los alamos, new mexico. American journal of public health **88**(9), 1377–1380 (1998)
49. Lai, C., Nguyen, N.T.: Predicting density-based spatial clusters over time. In: ICDM, pp. 443–446 (2004)
50. Levchuk, G., Bobick, A., Jones, E.: Activity and function recognition for moving and static objects in urban environments from wide-area persistent surveillance inputs. In: Proceedings of SPIE, vol. 7704, p. 77040P (2010)
51. Levine, N.: CrimeStat 3.0: A Spatial Statistics Program for the Analysis of Crime Incident Locations. Ned Levine & Associatiates: Houston, TX / National Institute of Justice: Washington, DC (2004)
52. Li, S.Z.: Markov random field modeling in image analysis. Springer (2009)
53. Little, B., Schucking, M., Gartrell, B., Chen, B., Ross, K., McKellip, R.: High granularity remote sensing and crop production over space and time: Ndvi over the growing season and prediction of cotton yields at the farm field level in texas. In: ICDM Workshops, pp. 426–435 (2008)
54. Lovell, J.: Left-hand-turn elimination (December 9, 2007). URL `\url{http://goo.gl/3bkPb}`. New York Times
55. Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., Hellerstein, J.M.: Graphlab: A new framework for parallel machine learning. arXiv preprint arXiv:1006.4990 (2010)
56. Luc, A.: Local Indicators of Spatial Association: LISA. Geographical Analysis **27**(2), 93–115 (1995)
57. Lynx GIS: www.lynxgis.com/ (May 7, 2013). URL `\url{http://www.lynxgis.com/}`
58. Ma, D., Zhang, A.: An adaptive density-based clustering algorithm for spatial database with noise. In: ICDM, pp. 467–470 (2004)
59. Malewicz, G., Austern, M., Bik, A., Dehnert, J., Horn, I., Leiser, N., Czajkowski, G.: Pregel: a system for large-scale graph processing. In: Proceedings of the 2010 international conference on Management of data, pp. 135–146. ACM (2010)
60. Manyika, J., et al.: Big data: The next frontier for innovation, competition and productivity. McKinsey Global Institute, May (2011)
61. MasterNaut: Green Solutions (May 7, 2013). URL `\url{http://www.masternaut.co.uk/carbon-calculator/}`
62. Mohan, P., Shekhar, S., Shine, J.A., Rogers, J.P.: Cascading spatio-temporal pattern discovery. IEEE Trans. Knowl. Data Eng. **24**(11), 1977–1992 (2012)
63. Moore, D.S., Neal, D.K.: Introduction to the Practice of Statistics TI-83 Graphing Calculator Manual. Macmillan (2005)
64. Morimoto, Y.: Mining frequent neighboring class sets in spatial databases. In: ACM SIGKDD (2001)
65. Murray, D.G., Schwarzkopf, M., Smowton, C., Smith, S., Madhavapeddy, A., Hand, S.: Ciel: a universal execution engine for distributed data-flow computing. In: Proceedings of the 8th USENIX conference on Networked systems design and implementation, p. 9 (2011)
66. Nature: Big Data: Science in the Petabye Era (4 September 2008). Volume 455 Number 7209 pp1-136
67. NAVTEQ Maps: www.navteq.com (May 7, 2013). URL `\url{www.navteq.com}`
68. Neill, D., Moore, A.: Rapid detection of significant spatial clusters. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 256–265. ACM New York, NY, USA (2004)

69. Neill, D.B., Moore, A.W., Cooper, G.F.: A bayesian spatial scan statistic. Advances in neural information processing systems **18**, 1003 (2006)
70. Neill, D.B., Moore, A.W., Pereira, F., Mitchell, T.M.: Detecting significant multidimensional spatial clusters. In: Advances in Neural Information Processing Systems, pp. 969–976 (2004)
71. New York Times: Military Is Awash in Data From Drones (January 10, 2010). URL \url{http://www.nytimes.com/2010/01/11/business/11drone.html?pagewanted=all}
72. New York Times: Unreported Side Effects of Drugs Are Found Using Internet Search Data, Study Finds (March 6, 2013). URL \url{http://goo.gl/h6VvYP}
73. New York Times: Mapping Ancient Civilization, in a Matter of Days (May 10, 2010). URL \url{http://www.nytimes.com/2010/05/11/science/11maya.html}
74. Ng, R.T., Han, J.: Clarans: A method for clustering objects for spatial data mining. IEEE Trans. Knowl. Data Eng. **14**(5), 1003–1016 (2002)
75. Nobelprize.org: The Nobel Prize in Chemistry 2013. URL \url{http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/}
76. Pang-Ning, T., Steinbach, M., Kumar, V.: Introduction to data mining. Pearson (2006)
77. Pei, T., Jasra, A., Hand, D.J., Zhu, A.X., Zhou, C.: Decode: a new method for discovering clusters of different densities in spatial data. Data Min. Knowl. Discov. **18**(3), 337–369 (2009)
78. Pei, Y., Zaïane, O.R., Gao, Y.: An efficient reference-based approach to outlier detection in large datasets. In: ICDM, pp. 478–487 (2006)
79. Shekhar, S., Evans, M.R., Kang, J.M., Mohan, P.: Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1**(3), 193–214 (2011)
80. Shekhar, S., Huang, Y., Xiong, H.: Discovering spatial co-location patterns: A summary of results. In: 7th International Symp. on Spatial and Temporal Databases (SSTD). L.A., CA (2001)
81. Shekhar, S., Lu, C., Zhang, P.: Detecting graph-based spatial outliers: Algorithms and applications. Proc of the ACM SIGKDD (2001)
82. Shekhar, S., Lu, C.T., Zhang, P.: Detecting graph-based spatial outliers. Intelligent Data Analysis **6**(5), 451–468 (2002)
83. Shekhar, S., Schrater, P.R., Vatsavai, R.R., Wu, W., Chawla, S.: Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. IEEE Transactions on Multimedia **4**(2) (2002)
84. Snow, J.: On the mode of communication of cholera. John Churchill (1855)
85. Solberg, A.H., Taxt, T., Jain, A.K.: A Markov Random Field Model for Classification of Multisource Satellite Imagery. IEEE Transaction on Geoscience and Remote Sensing **34**(1), 100–113 (1996)
86. Sperling, D., Gordon, D.: Two billion cars. Oxford University Press (2009)
87. Strahler, A.: The use of prior probabilities in maximum likelihood classificaiton of remote sensing data. Remote Sensing of Environment **10**, 135–163 (1980)
88. Sun, P., Chawla, S.: On local spatial outliers. In: ICDM, pp. 209–216 (2004)
89. TeleNav: www.telenav.com/ (May 7, 2013). URL \url{http://www.telenav.com/}
90. TeloGIS: www.telogis.com/ (May 7, 2013). URL \url{http://www.telogis.com/}
91. TomTom: TomTom GPS Navigation (May 7, 2013). URL \url{http://www.tomtom.com/}
92. Varnett, V., Lewis, T.: Outliers in Statistical Data. John Wiley (1994)
93. Wah, B.W. (ed.): Wiley Encyclopedia of Computer Science and Engineering. John Wiley & Sons, Inc. (2008)
94. Wikipedia: Usage-based insurance — wikipedia, the free encyclopedia (May 7, 2013). URL \url{http://en.wikipedia.org/wiki/Usage-based_insurance}
95. Wu, E., Liu, W., Chawla, S.: Spatio-temporal outlier detection in precipitation data. In: Knowledge discovery from sensor data, pp. 115–133. Springer (2010)
96. Wu, W., Cheng, X., Ding, M., Xing, K., Liu, F., Deng, P.: Localized outlying and boundary data detection in sensor networks. IEEE Trans. Knowl. Data Eng. **19**(8), 1145–1157 (2007)
97. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications **1**(1), 7–18 (2010)
98. Zhang, Y., Gao, Q., Gao, L., Wang, C.: Priter: a distributed framework for prioritized iterative computations. In: Proceedings of the 2nd ACM Symposium on Cloud Computing, p. 13. ACM (2011)