RESEARCH ARTICLE

Space-Time Big Data: An Analytics Perspective

Dev Oliver∗, Michael R. Evans, Xun Zhou, Zhe Jiang, Shashi Shekhar

Department of Computer Science and Engineering, University of Minnesota, USA;
(Received 00 Month 200x; final version received 00 Month 200x)

Recent years have seen the proliferation of space-time datasets whose size, variety, and update rate exceed the capacity of commonly used spatio-temporal computing and database technologies to learn, manage, and process the data with reasonable effort. Examples of such datasets, which we refer to as space-time big data (STBD), include GPS trajectory data from cell-phones, UAV Data, roadmaps that provide traffic speed values every minute for every road segment in a city, climate simulation data, and engine measurement data consisting of fuel consumption, greenhouse gas emissions, etc. STBD is of critical societal importance; a recent McKinsey Global Institute estimates that “services enabled by personal-location data can allow consumers to capture $600 billion in economic surplus”. However, STBD is challenging traditional data analytics, requiring new and interesting conceptual models, interest measures, and algorithms to be developed. In this paper, we explore the emerging field of space-time big data analytics, providing a broad overview of novel STBD pattern families including summarization, obfuscated outliers, rare associations, and obfuscated event/process prediction. We conclude with a look at future research directions.

Keywords: space-time big data; big data analytics

1. Introduction

Space-Time Big Data refers to datasets whose size, variety, and update rate exceed the capacity of commonly used spatio-temporal computing and database technologies to learn, manage, and process the data with reasonable effort. Examples include temporally detailed (TD) roadmaps that provide speeds every minute for every road-segment, GPS track data from cell-phones, climate simulation data, engine measurements of fuel consumption, greenhouse gas (GHG), emissions, etc.

∗Corresponding author. Email: oliver@cs.umn.edu
Space-Time Big Data is important for critical societal applications such as eco-routing, public safety and security, and understanding climate change. Eco-routing may leverage various forms of STBD to compare routes by fuel consumption or greenhouse gas (GHG) emissions rather than total distance or travel-time. For example, UPS has saved millions of gallons of fuel by simply avoiding left turns and associated engine-idling when selecting routes [Lovell (2007)]. Immense savings in fuel-cost and GHG emission are possible if other fleet owners and consumers avoided hot spots of idling, low fuel-efficiency, and congestion. Public safety professionals may use STBD analysis to identify crime hotspots to select police patrol routes, social interventions, etc. A region with an increase in number of crime reports may indicate the need for more attention. Climate scientists are interested in a number of change related questions such as how to verify and quantify global climate change. Specifically, they may ask questions such as: when did the precipitation in the Sahel region in Africa change from normal to very limited? Is the intensity, frequency, or duration of extreme climate events (e.g., flood) changing?

STBD raises new challenges for the state of the art in data analytics. First, it requires a change in frame of reference, from a global snapshot perspective to the perspective of the individual object traveling through a road network, river network, etc. Second, the sheer volume, variety, velocity and velocity of space-time big data exceeds the capability of current systems to process and manage the data with reasonable effort. Third, the growing diversity of STBD sources makes it less likely that single algorithms, working on specific spatio-temporal datasets, will be sufficient to discover answers appropriate for all situations.

In this paper, we provide a structured and a broad overview of the research on space-time big data including a taxonomy of emerging space-time big data techniques. The rest of this paper is organized as follows: Space-Time Big Data Inputs begins with a description of the data input characteristics of several tasks in STBD analytics. Space time big data pattern families explains in detail four main output patterns and methods of STBD related to summarization, obfuscated outliers, rare associations, and obfuscated event/process prediction. Future Directions and Research Needs concludes this paper with an examination of research needs and future directions.

2. Space-Time Big Data Inputs

The data inputs of STBD have both spatial and temporal attributes. Spatial attributes are used to define the spatial location and extent of space-time objects. The spatial attributes of a space-time object most often includes information related to spatial locations, for example, longitude, latitude, and elevation, defined in a spatial reference frame, as well as shape. Temporal attributes, on the other hand, are used to define the temporal extent and granularity of space time objects (e.g., snapshot, time-series). Examples of space-time big data include GPS Trace Data, unmanned aerial vehicle (UAV) data, LiDAR, VGI Data, Spatio-Temporal Engine Measurement Data, and Historical Speed Profile Data.

GPS Trace Data: An example of emerging Space-Time Big Data, GPS trajectories, are becoming available for a larger collection of vehicles due to the rapid proliferation of cell-phones, in-vehicle navigation devices, and other GPS data-logging devices [Garmin (2013)] such as those distributed by insurance companies [Wikipedia (2011)]. Such GPS traces allow indirect estimation of fuel efficiency and GHG emissions via estimation of
vehicle-speed, idling, and congestion. They also make it possible to provide personalized route suggestions to users to reduce fuel consumption and GHG emissions. For example, Figure 10 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 are 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 (Garmin (2013), TomTom (2011)) 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.

![Fig 1](image1.png)

(a) GPS Trace Data. Color indicates speed.

![Fig 2](image2.png)

(b) Routes 1, 2, & 3 (Google Maps 2013).

Figure 1. A commuter’s GPS tracks over three months reveal preferred routes. (Best viewed in color)

**UAV 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 coverage 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 UAVs in gathering and managing raster datasets. First, UAVs have a small footprint due to the relatively low flying height, therefore, they capture a large amount of images in a very short period of time to achieve the spatial coverage for many applications. This poses a significant challenge to store rapidly increasing digital images. 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 (New York Times 2010b). It is essential to develop automated, efficient, and accurate technique to handle such space-time 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 (New York Times 2010a). Lidar data are very rich to analyze surface or extract features. However, these data sets contain irrelevant
data for space-time analysis and sometimes miss critical information. These large volumes of data from multiple sources poses 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. Data integration from multiple spatial sources is another challenge due to the massive amounts of Lidar datasets. Therefore, dealing with space-time big data is an essential issue for Lidar remote sensing.

**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 based on user interactive online services (e.g., OpenStreetMap, Wikimapia, GoogleMap, GoogleEarth, Microsoft’s Virtual Earth, Flickr, etc). In recent years, VGI has lead an explosive growth in the availability of user-generated geographic information and has required bigger storage models to handle large scale space-time datasets. The challenge for VGI is to enhance data service quality with regard to accuracy, credibility, reliability, and overall value (InformationWeek (2012)).

**Spatio-Temporal Engine Measurement Data:** Many modern fleet vehicles include rich instrumentation such as GPS receivers, sensors to periodically measure sub-system properties (Kargupta et al. (2010, 2006), Lynx GIS (2013), MasterNaut (2013), TeleNav (2013), TeloGIS (2013)), and 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 3 shows heavy truck fuel consumption as a function of elevation from a recent study at Oak Ridge National Laboratory (Capps et al. (2008)). Notice how fuel consumption changes drastically with elevation slope changes. Fleet owners have studied such datasets to fine-tune routes to reduce unnecessary idling [ATRI] (2010b, a). 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 very large. For example, measurements of 10 engine variables, once a minute, over the 100 million US vehicles in existence [Sperling and Gordon (2009), Federal Highway Administration (2008)], may have $10^{14}$ data-items per year.

![Figure 4. Space-Time Big Data on Historical Speed Profiles. (Best viewed in color)](image)

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

(b) Schema for Daily Historic Speed Data.

**Historical Speed Profiles:** Traditionally, digital road maps consisted of center lines and topologies of the road networks (George and Shekhar (2008), Shekhar and Xiong (2007)). These maps were used by navigation devices and web applications such as Google Maps (Google Maps (2013)) to suggest routes to users. New datasets from companies such as NAVTEQ (NAVTEQ (2013)), 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 temporally-detailed (TD) roadmaps contain much more speed information than traditional 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 4(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 large 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. Thus, industry is using speed-profiles, a lossy compression based on the idea of a typical day of a week, as illustrated in Figure 4(b), where each (road-segment, day of the week) pair is associated with a time-series of speed values for each hour of the day.

The advent of these new space-time big data sources has created room for novel STBD patterns, which we discuss next.

3. **Space-Time Big Data Pattern Families**

Emerging patterns in Space-Time Big Data include summarization, obfuscated outliers, rare associations, and obfuscated event/process prediction. This section elaborates on these techniques by briefly describing their applications and relating them to traditional techniques. Table 1 summarizes example pattern families from traditional data mining, spatial data mining, and space-time data mining, and contrasts them with newly emerging patterns in space-time big data.

<table>
<thead>
<tr>
<th>Traditional Data Mining</th>
<th>Spatial Data Mining</th>
<th>Space-Time Data Mining</th>
<th>Space-Time Big Data Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>Hotspots</td>
<td>Emerging/Spreading Hotspots</td>
<td>Summarization</td>
</tr>
<tr>
<td>Outlier Detection</td>
<td>Spatial Outlier</td>
<td>Abrupt Change</td>
<td>Obfuscated Outliers</td>
</tr>
<tr>
<td>Association Rule Mining</td>
<td>Colocation</td>
<td>Teleconnection/Co-occurrence</td>
<td>Rare Associations</td>
</tr>
<tr>
<td>Prediction/Classification</td>
<td>Location Prediction</td>
<td>Event/Process Prediction</td>
<td>Obfuscated Event/Process Prediction</td>
</tr>
</tbody>
</table>

### 3.1. **Summarization**

Data mining has traditionally employed clustering techniques for dividing data into groups, which is an important task in many domains (e.g., biology, information retrieval, climate, etc.) Pang-Ning et al. (2006). In traditional clustering, the attribute values within a cluster could be independent and identically distributed (i.i.d), which is an assumption that does not hold in space-time data. Spatial data mining hotspots took this a step further by accounting for the fact that the spatial correlation of the attribute values within a hotspot could be high and possibly drops dramatically at the boundary Shekhar et al. (2011). Space-time data mining techniques group objects based on both their spatial and temporal similarity Kisilevich et al. (2010).

However, several computational issues arise with space-time big data. The volume of data, the complexity of space-time data types and relationships, and the need to identify autocorrelation pose numerous computational challenges to STBD analytics. When designing algorithms, one has to take into account considerations such as space-time partitioning, predicate approximation, multidimensional data structures, etc. Many data analysis techniques have a time or space complexity of $O(M^2)$ or higher (where $m$ is the number of objects), and thus, are not practical for space-time big data sets.

Summarization (i.e., finding a compact description or representation of a dataset) may be an important first step for reducing the semantic gap in space-time big data by providing novel GPS track and route-based summaries. The process typically involves defining a set of groups, finding a representative for each group, and reporting a statistic for each group (e.g., sum, mean, standard deviation). These notions differ depending on
the genre of the data being summarized. Next, we present two examples of summarization techniques for GPS trajectories and spatial networks.

3.1.1. GPS Trajectory Summarization Example: Identifying \( k \)-Primary Bike Corridors

Given a set of trajectories on a road network, the goal of the \( k \)-Primary Corridors problem is to summarize trajectories into \( k \) groups, each represented by its most central trajectory. Figure 5(a) shows a real-world GPS dataset of a number of trips taken by bicyclists in Minneapolis, MN. The darkness indicates the usage levels of each road segment. The computational problem is summarizing this set of trajectories into a set of \( k \)-Primary Corridors. One potential solution is shown in Figure 5(b) for \( k = 8 \). Each identified primary corridor represents a subset of bike tracks from the original dataset. Note that the output of the \( k \)-PC problem is distinct from that of hot or frequent routes, as it is a summary of all the given trajectories partitioned into \( k \) primary corridors.

The \( k \)-Primary Corridors problem is important due to a number of societal applications, such as city-wide bus route modification or bicycle corridor selection, among other urban development applications. Let us consider the problem of determining primary bicycle corridors through a city to facilitate safe and efficient bicycle travel. By selecting representative trajectories for a given group of commuters, the overall alteration to commuters routes is minimized, encouraging use. Facilitating commuter bicycle traffic has shown in the past to have numerous societal benefits, such as reduced greenhouse gas emissions and healthcare costs [Marcotty 2012].

This problem is challenging due to the computational cost of computing pairwise graph-based minimum-node-distance similarity metrics between trajectories in large GPS datasets as shown in our previous work [Evans et al. 2012]. We proposed a baseline algorithm using a graph-based approach to compute a single element of the trajectory similarity matrix, requiring multiple invocations of common shortest-path algorithms (e.g., Dijkstra [Cormen 2001]). For example, given two trajectories consisting of 100 nodes each, a baseline approach to calculate network distance would need to compute the shortest distance between all pairs of nodes \( (10^3) \), which over a large trajectory dataset (e.g., 10,000 trajectories) would require \( 10^{12} \) shortest path distance computations. This quickly becomes computationally prohibitive without faster algorithms.

Figure 5. Example Input and Output of the \( k \)-Primary Corridor problem.
3.1.2. Spatial Network Summarization Example: Discovering Accident Prone Roads

Given a spatial network, a collection of activities and their locations (e.g., placed on a node or an edge), and a desired number of paths $k$, the spatial network activity summarization problem finds a set of $k$ shortest paths that maximizes the sum of activities on the paths (counting activities that are on overlapping paths only once) and a partitioning of activities across the paths. An activity may be the location of a pedestrian fatality or a crime (e.g., theft), depending on the domain. Figure 6(a) shows an example of the input consisting of 43 pedestrian Fatalities (represented as dots) in Orlando, Florida occurring between 2000 and 2009. The output (shown in in Figure 6(b)) uses paths and network distance to group activities on a spatial network. SNAS assumes that each path is a shortest path because in applications such as transportation planning, many people aim to arrive at their destination as fast as possible.

Spatial network activity summarization (SNAS) is important in several application domains including transportation planning and crime analysis where observations occur on paths in the spatial (e.g., road) network. For example, transportation planners and engineers may need to identify frequently used, accident prone road segments/stretches requiring redesign, crime analysts may look for concentrations of crimes along certain streets to guide law enforcement, and hydrologists and environmental engineers may try to summarize environmental change on water resources to understand the behavior of river networks and lakes.

Finding a set of $k$ shortest paths that maximizes the number of activities on selected paths is computationally challenging. This is due to the fact that if $k$ shortest paths are selected from all shortest paths in a spatial network, there are a large number of possibilities for large $k$. This is because different subsets of $k$ shortest paths could be overlapping or have the same shortest paths. For disjoint paths, the problem would be relatively less computationally challenging. However, due to overlapping paths, the general problem of SNAS is NP-complete.

3.2. Obfuscated Outliers

Outlier detection has progressed from traditional data mining, spatial data mining, and space-time data mining to space-time big data analytics. Outliers have traditionally been defined as observations in a data set that appear to be inconsistent with the remainder of that set of data Barnett and Lewis (1984), or which deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism Hawkins.
The identification of global outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as detection of credit card fraud and voting irregularities. Spatial outliers extended this idea by identifying spatially referenced objects whose non-spatial attribute values differ significantly from those of other spatially referenced objects in their spatial neighborhoods Shekhar et al. (2002). An example is a new house in an old neighborhood of a growing metropolitan area based on the non-spatial attribute house age. Spatio-temporal outliers are spatio-temporal objects whose thematic (non-spatial and non-temporal) attributes are significantly different from those of other objects in their spatial and temporal neighborhoods Wu et al. (2010).

The University of Minnesota and the Track Management Center (TMC) Freeway Operations group created a database to archive sensor network measurements from the freeway system in the Twin Cities Shekhar et al. (2002). The sensor network includes about nine hundred stations, each of which contains one to four loop detectors, depending on the number of lanes. Sensors embedded in the freeways and interstate monitor the occupancy and volume of track on the road. At regular intervals, this information is sent to the track Management Center for operational purposes, e.g., ramp meter control, as well as research on track modeling and experiments. Figure 7(a) shows a map of the stations on the highways within the Twin-Cities metropolitan area, where each polygon represents one station. The interstate freeways include I-35W, I35E, I-94, I-394, I-494, and I-694. The state trunk highways include TH-100, TH-169, TH-212, TH-252, TH-5, TH-55, TH-62, TH-65, and TH-77. I-494 and I-694 together form a ring around the Twin-Cities. I-94 passes from East to North-West, while I-35W and I-35E run in a South-North direction. Downtown Minneapolis is located at the intersection of I-94, I-394, and I-35W, and downtown Saint Paul is located at the intersection of I-35E and I-94. For each station, there is one detector installed in each lane. The track flow information measured by each detector can then be aggregated to the station level. The system records all the volume and occupancy information within each 5-minute time slot at each particular station. This is an example of space-time big data.

In outlier detection, we may be interested in discovering the location of stations whose measurements are inconsistent with those of their graph-based spatial neighbors and time periods when those abnormalities arise. We use three neighborhood definitions in this application as shown in Figure 7(b). First, we define a neighborhood based on the spatial graph connectivity as a spatial graph neighborhood. In Figure 7(b) (s1; t2) and (s3; t2) are the spatial neighbors of (s2; t2) if s1 and s3 are connected to s2 in a spatial graph. Second, we define a neighborhood based on time series as a temporal neighborhood. In Figure 7(b) (s2; t1) and (s2; t3) are the temporal neighbors of (s2; t2) if t1, t2, and t3 are
consecutive time slots. In addition, we define a neighborhood based on both space and time series as a spatial-temporal neighborhood. In Figure 7(b), \((s1; t1), (s1; t2), (s1; t3), (s2; t1), (s2; t3), (s3; t1), (s3; t2),\) and \((s3; t3)\) are the spatial-temporal neighbors of \((s2; t2)\) if \(s1\) and \(s3\) are connected to \(s2\) in a spatial graph, and \(t1, t2,\) and \(t3\) are consecutive time slots. The test for detecting an outlier may be described as follows: \(| \frac{S(x) - \mu_s}{\sigma_s} | > \theta \)

For each data object \(x\) with an attribute value \(f(x)\), the \(S(x)\) is the difference of the attribute value of data object \(x\) and the average attribute value of its neighbors. \(\mu_s\) is the mean value of all \(S(x)\), and \(\sigma_s\) is the standard deviation of all \(S(x)\). Choice of \(\theta\) depends on a specified confidence interval. For example, a confidence interval of 95 percent will lead to \(\theta \approx 2\).

![Figure 8. Obfuscated outliers reveal exceptions to patterns of life. STBD raises questions such as why is the highlighted block avoided by cyclists? (Best in color)](image)

Space-time big data has provided new opportunities for detecting obfuscated outliers such as exceptions to patterns of life, which has application in transportation planning, accident analysis and prevention, etc. Figure 8 shows an example of real GPS data indicating various bicyclist commute routes in a part of the road network in the twin cities Evans et al. (2012). As can be seen, the highlighted cycling route is not straight but seems to detour when traveling north to south. With STBD, one may ask questions such as “why is a block avoided by cyclists?” Such anomalies become more apparent with the presence of more data. Another example may be seen in analysis done by Xoom, a company that specializes in international money transfers. Xoom raised an alarm in 2011 after noticing an increase in Discover Card transactions from New Jersey, it was later discovered that these transactions came from a criminal group Mayer-Schonberger and Cukier (2013).

### 3.3. Rare Associations

Traditional data mining has benefited from association rule mining to discover items that were bought together in transaction data Agrawal et al. (1994). However, be-
cause transactions were not natural for continuous space, colocation patterns were proposed [Huang et al. (2004)]. Co-location patterns represent subsets of boolean spatial features whose instances are often located in close geographic proximity. Examples include symbiotic species and crime attractors (e.g., bars, misdemeanors, etc.). Boolean spatial features describe the presence or absence of geographic object types at different locations in a two-dimensional or three dimensional metric space, for example, the surface of the Earth. Examples of boolean spatial features include plant species and crime. Space-time data mining has taken the colocation idea a step further to find subsets of event-types whose instances are located together and occur in stages (e.g., cascading spatio-temporal pattern discovery [Mohan et al. (2012)]). For example, analysis of crime data sets may reveal frequent occurrence of misdemeanors and drunk driving after and near bar closings on weekends, as well as after and near large gatherings such as football games [Mohan et al. (2012)].

Space-time big data will undoubtedly lead to further progress, facilitating new analytics questions in domains such as climate science. Teleconnections (e.g., El Nino/La Nia events), an important exception to Toblers first law, play a crucial role in climate science and must be accounted for in next generation STBD systems. An example includes questions such as: What is the impact of El Nino of a city? What are relevant spatio-temporal co-occurrence patterns? Next we discuss a relevant emerging technique in this space.

3.3.1. Interesting Subpath Discovery in Climate Data

Sub-paths (i.e., intervals) in spatio-temporal (ST) datasets can be defined as contiguous subsets of locations. Given a ST dataset and a path in its embedding ST framework, the goal of the interesting ST sub-path discovery problem is to identify all the dominant (i.e., not a subset of any other) interesting sub-paths along the path defined by an interest measure. The ability to discover interesting sub-path is important to many societal applications. For example, coastal area authorities may be interested in intervals of coastal lines which are prone to rapid environmental change due to rising ocean and melting polar icecap. Water quality monitors may be interested in river segments where water quality changes abruptly.

![Smoothed Africa vegetation dataset (measured in NDVI) in August, 1981.](image1)

![Longitudinal intervals of abrupt vegetation change in August, 1981.](image2)

**Figure 9.** An application example of the interesting interval discovery problem (best viewed in color).

An extended example from eco-climate science illustrates the interesting sub-path discovery problem in detail. This example comes from our collaboration with scientists
studying the response of ecosystems to climate change by observing changes in vegetation cover across ecological zones. Sub-paths of abrupt vegetation cover change may serve to outline the spatial footprint of ecotones, the transitional areas between these zones [Noble (STOR)]. Due to their vulnerability to climate changes, finding and tracking ecotones gives us important information about how the ecosystem responds to climate changes. Figure 9 illustrates the application of interesting sub-path discovery on the Africa vegetation cover in normalized difference vegetation index (NDVI) data, August, 1981. Figure 9(a) shows a map of vegetation cover in Africa ?. Each longitudinal path is taken as an input of the problem. The output of this problem, as shown in Figure 9(b), is a map of longitudinal sub-paths with abrupt vegetation cover changes, where red and blue represent sub-paths of abrupt vegetation cover decrease and increase northward respectively. As indicated by the the two colors as in Figure 9 footprints of several ecotones in Africa are discovered. One of them is the Sahel region(in the middle in red), where vegetation cover exhibits an abrupt decreasing trend from south to north.

Discovering interesting sub-paths is challenging due to the following reasons: First, the length of the sub-paths of interest may vary, without a pre-defined maximum length. For example, the length of flood-prone interval in long rivers (e.g., the Gange, Mississippi, etc) may extend hundreds or thousands of miles. Second, the interestingness in a sub-path may not exhibit monotonicity, i.e., uninteresting intervals may be included in an interesting sub-path. Third, the data volume is potentially large. For example, consider the problem of finding all the interesting longitude sub-paths exhibiting abrupt change in an eco-climate dataset with attributes such as vegetation, temperature, precipitation, etc., over hundreds of years from different global climate models and sensor networks. The volume of such datasets will range from terabytes to petabytes.

3.4. **Obfuscated Event/Process Prediction**

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. Space-time data does not follow independent identical distribution due to properties such as spatial autocorrelation (Tobler’s Law, Markov Model) and heterogeneity (Geographically Weighted Regression) [Shekhar et al (2011)]. Space-time big data provides new opportunities for estimating spatial neighbor relationships and supporting place-based ensemble models.

3.4.1. **Estimating Spatial Neighbor Relationships**

Spatial data inputs are complex because they include extended objects such as points, lines, and polygons in vector representation and field data in regular or irregular tessellation such as raster data [Bolstad (2005)]. During data input, relationships among spatial objects are often implicit (e.g., overlap, intersect, etc.) and are often captured by models or techniques that incorporate spatial information into the STBD process. One such technique is to model the spatial relationship among locations in a spatial framework via a contiguity matrix which may represent a neighborhood relationship defined using adjacency or Euclidean distances. These neighborhood or W matrices are used in many STBD tasks such as spatial outlier detection, co-location pattern discovery, spatial classification and regression modeling, spatial clustering, and spatial hotspot analysis [Shekhar et al (2011)].

The W matrix poses a significant burden to end users due to the fact that W is quadratic in the number of locations and reliable estimation of W needs a very large
number of data samples. In spatial classification and regression modeling, for example, the logistic spatial autoregressive model (SAR) includes the neighborhood relationship contiguity matrix. Table 2 shows a comparison of the classical linear regression model and the spatial auto-regression model where the spatial dependencies of the error term, or the dependent variable, are directly modeled in the regression equation.

STBD may be large enough to provide a reliable estimate of W. 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.

3.4.2. Supporting Place-based Ensemble Models

Spatial heterogeneity (or nonstationarity) is an important concept in STBD that is rarely modeled. An important feature of spatial data sets is the variability of observed processes over space. Spatial heterogeneity refers to the inherent variation in measurements of relationships over space. In other words, no two places on Earth are identical. The influence of spatial context on spatial relationships can be seen in the variation of human behavior over space (e.g., differing cultures). Different jurisdictions tend to produce different laws (e.g., speed limit differences between Minnesota and Wisconsin). The term spatial heterogeneity is most often used interchangeably with spatial nonstationarity, which is defined as the change in the parameters of a statistical model or change in the ranking of candidate models over space [Bailey and Gatrell (1995)].

Traditional Astro-Physics-based models have been place-independent for the most part, with the notable exception of geographically weighted regression (GWR) [Brunsdon et al. (1996), Fotheringham et al. (2002)]. The regression equation for GWR, shown by Eq. 1, has the same structure as standard linear regression, with the exception that the parameters are spatially varying, where \( \beta(s) \) and \( \epsilon(s) \) represent the spatially varying parameters and the errors, respectively. GWR provides an ensemble of linear regression models, one per place of interest.

\[
y = X \beta(s) + \epsilon(s)
\]

STBD may support a Place-based ensemble of models beyond 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 auto-correlation is an interesting possibility that will need to be explored further.
4. Future Directions

This section presents future directions and research needs in STBD analytics. There are several new areas of research, but the four we will focus on are simplifying space-time models, online STBD analytics, statistical basis for STBD, and medical STBD analytics.

4.1. Simplifying Space-Time Models

Space-Time models are usually 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 (Table 2). Geographically weighted regression has the same limitation as opposed to classical linear regression, also due to the inclusion of the $W$ matrix (Eq. 1). Colocation pattern mining, which finds the subsets of features frequently located together is more computationally expensive that traditional association rule mining [Agrawal et al. (1994)] and confidence estimation adds more costs (e.g., M.C.M.C. simulations).

STBD creates an opportunity to simplify space-time models in traditional spatio-temporal data mining. It may be the case that some of the complexity from spatio-temporal data mining 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. STBD may provide a large volume of data at each place which may allow algorithmic techniques such as place-based divide and conquer. Consequently, it may only be necessary to build one model per place using local data and simpler models. There are, however, a few challenges that must be considered when comparing place-based ensembles of simpler models with current spatial models. First, it is unclear when bigger data leads to simpler models. Second, there is little consensus on the definition of STBD from an analytics perspective (e.g., ratio of samples to number of parameters).

4.2. On-line Spatio-Temporal Big Data 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 autoregression, colocation pattern mining, and hotspot detection). However, STBD includes streaming data such as event reports and sensor measurements. Furthermore, the use cases for STBD 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, STBD may provide opportunities for precisely computing them in an on-line manner. If precise on-line computation is not possible, STBD might be useful in providing on-line approximations.
4.3. **Statistical Basis for Space-Time Big Data**

Traditionally, society has depended on using samples when faced with large numbers. The idea was to extrapolate useful information from a small sample about the general population and make appropriate inferences about the general population based on the sample. This has largely been an artifact of information scarcity. For example, census were so complex, costly, and time consuming that they were conducted only rarely and samples were used to facilitate acquiring the information. However, the accuracy of samples depends on randomness when collecting the sample data which may be non-trivial since semantic biases in the way data is collected can lead to the extrapolated results being wrong [Mayer-Schonberger and Cukier (2013)].

Space-time big data facilitates access to subgroups that samples can’t access. Random sampling does not scale easily to include subcategories and increases the possibility of erroneous predictions. In contrast STBD enables drill down and roll up to facilitate analysis of all facets of the data. Space-time big data may reduce the edge effect due to ignorance of interdependences that occur outside the bounded region. Edge effects may be lost in random samples. For example, the convex hull of a random sample will not equal the convex hull of big data, even though big data is large. This violates traditional assumptions of random samples wherein the mean of \( k \) samples is expected to approach the mean of the population as the sample sizes increase. Space-time big data may also reduce Simpson’s Paradox in spatio-temporal data where trends that appear in different groups of data disappears when these groups are combined.

4.4. **Medical Space-Time Big Data Analytics**

Medical Space-Time Big Data is available in many forms such as spatial networks formed by bodily systems and 3-D medical images [Oliver and Steinberger (2011)]. Spatial networks in the body include the circulatory system or blood vessels, the network of nerves, the network of bronchi and bronchioles in the lungs, and the skeletal system. An example source of a spatial network in this context is an angiogram showing blood vessels with blockage. 3-D images, on the other hand, include computer aided tomography (CT), ultrasound, etc.; they allow visualization of important structures in great detail and are therefore an important tool for the diagnosis and surgical treatment of many pathologies. A patient's spatial network or 3-D image data taken over time can be used in new ways such as long term study in which crucial monitoring, predictive, and routing questions may be answered algorithmically. Examples of monitoring questions include discovering how an anomalous growth (e.g., cancer) is changing over time or detecting the narrowing of blood vessels. Predictive questions involve determining therapy effect on tumors across a population as a guide for future therapies. Routing questions are asked to find a route through the body for minimally invasive surgery to remove a tumor.

A medical space-time big data framework for answering long term questions plays a critical role in improving health care quality by providing answers regarding the progression of disease and the comparative effectiveness of interventions. Such a framework (in conjunction with other information technologies) may also assist individuals to stay healthier by helping those with chronic or acute conditions to manage their disease outside of acute care settings. Providers are empowered with a means of simplifying the tracking of multi-focal disease based on 3D images or spatial networks taken over time and this might go a long way in reducing expenses such as the $2 trillion a year that the United States spends on health care.
However, one of the main challenges with medical STBD is data volume where there is a need to scale up to potentially petabyte and exabyte-sized data sets. Large amounts of data are produced from medical imaging techniques and replicating this data across different snapshots makes long term analysis prohibitive. Compression techniques (lossy and lossless) have been used to enable fast retrieval of static 3-D data (i.e., a single snapshot) but they are not adequate for dynamic 3-D data with features like interactive zoom in and out across the time dimension. For example, each snapshot of a large image might be approximately 8 - 16 gigabytes [Peng et al. (2010)]; when this is multiplied by number of visits, number of images/visit and number of patients, scale increases to exabytes.

5. Summary

Space-time big data is emerging from several 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 STBD pattern families are becoming apparent including summarization, obfuscated outliers, rare associations, and obfuscated event/process prediction. However, STBD research is still largely unexplored territory with simplified space-time models, online STBD analytics, STBD statistics, and medical STBD emerging as new themes in this space.

References


REFERENCES


Garmin, Accessed: May 7, 2013. [online] [????].


InformationWeek, 2012. Red Cross Unveils Social Media Monitoring Operation. [online] [????].


Lynx GIS, Accessed: May 7, 2013. [online] [????].

REFERENCES

NAVTEQ, Accessed: May 7, 2013. , , [online] [????].
TeleNav, Accessed: May 7, 2013. , , [online] [????].
TeloGIS, Accessed: May 7, 2013. , , [online] [????].