Spatio-Temporal Big Data Analytics for Environmental Health

Prof. Shashi Shekhar

November 9, 2013

1 Specific Aims

2 Architecture

The framework for our proposed big data analytics platform is shown in Figure 1. Two complimentary systems support the wide variety of spatial analytics algorithms and techniques we are providing. On the left half of Figure 1, the more-traditional unix filesystem supports high-throughput computation (e.g., MPI [Snir et al., 1995], OpenMP [Dagum and Menon, 1998], GPGPU/CUDA Luebke et al. [2006]) typically used for spatial statistics and matrix computations. It also supports spatial data mining algorithms requiring massive iterations in computing. The right portion of the figure represents the distributed filesystem and distributed processing portion of the framework, represented at the lowest layer by HDFS (Hadoop Distributed File System). In this proposal, we focus on the Apache Hadoop system for the distributed portion of the framework, but will evaluate that choice as the distributed processing community evolves.

The proposed system is presented to the user via the Big Map Visualization & User Interface layer of the architecture. This layer will consist of a user-friendly application to import and export data, select desired analytical algorithms or queries to be run, and a map-based visualization component. This interface will have functions from each of the three components directly underneath it in Figure 1. We will examine each component and its underlying infrastructure in the next three sections.

Big Bayesian Spatial Statistics: This component provides functionality such as kernel density estimation, data kriging, etc. These advanced spatial statistics require massive low-latency parallel computing. The Big Matrix Algorithms component provides advanced matrix algorithms (e.g., matrix inversion) on parallel computing infrastructure as shown in the MPI, OpenMP, GPGPU component. This infrastructure combines traditional parallel system computing (MPI) with recent developments in GPU programming to enable low-latency parallel computing. These architectures typically utilize the Unix Filesystem, underlying this half of our Spatio-Temporal Big Data Analytics Architecture.

Spatial Data Mining: Crossing the two primary infrastructures, spatial data mining loosens the statistical rigor of traditional spatial statistics to provide computationally efficient algorithms to find patterns in spatial data. More traditional techniques, such as the Co-Location algorithm [Shekhar and Huang, 2001] we discuss in the next section, run on the Unix Filesystem with performance enhancements from MPI/OpenMP. We are proposing to develop Spatial Mahout, a spatial extension of the popular Mahout [Owen et al., 2011] distributed processing data mining package (incorporating a spatial join index, one of the fundamental requirements for spatial big analytics). Spatial Mahout will run on current distributed processing systems, such as Apache Hadoop (YARN, MRv2) [White, 2012], and shown in the Distributed Processing component. Another key cross-component interaction of Spatial Data Mining and Spatial Mahout is its use of Spatial Hadoop for efficient data access.

Spatial Hadoop: The Spatial Hadoop [Eldawy and Mokbel, 2013] component spans multiple layers of the architecture as it has multiple subcomponents for querying and processing spatial data. Higher-level functions of Spatial Hadoop include a spatial query language extending PIG [Olston et al., 2008] to efficiently query spatial data. Spatial Hadoop also provides a spatial index that catalogs the underlying Hadoop Distributed File System [Borthakur, 2008] to provide smart retrieval of spatial data.

3 Background and Significance

3.1 Use Cases - Geographic Surveillance of Cancers, Environmental Burdens, etc.

Maps have long been used in surveillance of disease and in epidemiological investigations. Geographic surveillance is the process of detecting, characterizing, tracking, and responding to disease outbreaks and other relevant health patterns. Left undiscovered, outbreaks can cause major hardship to communities, especially when the disease is a result of some air, water, or soil contaminant that could be controlled or removed. In 2003, a New Jersey state report summarized the analytic results confirming a statistically significant increase in childhood cancers for the period 1979 to 1995 in Dover Township [Blumenstock et al., 2000]. This increase was due primarily to excess leukemia, brain, and central nervous system cancers in females less
than 5 years old. A case-control analysis revealed increased association for leukemia in young girls and high exposure to well water in proximity to an industrial site. This cancer cluster was not formally investigated as an outbreak until the 1990’s, after residents demanded an investigation of this potential hotspot of cancer activity. Automated geographic surveillance has the potential to find these events before they persist and do significant damage to a community.

Complex non-traditional datasets, specifically Spatio-Temporal Big Data, can provide new insights and increase the ability to provide situation awareness. Non-traditional clustering patterns, such as linear patterns or network patterns, can readily be found in public health literature (e.g., proximity to highway increasing cancer [McEntee and Ogneva-Himmelberger, 2008]) but are difficult to detect using traditional surveillance techniques. Geographic features are frequently a source of concern for cancer outbreaks, as shown in the recent news about a creek in St. Louis County, MO in Figure 2. Geographic features, such as creeks, rivers, valleys, mountains, etc. do not follow any predetermined shape or size. New algorithms and techniques are needed to find these linear and networked clusters, among other spatial patterns beyond ellipsoidal.

![Figure 1: Spatio-Temporal Big Data Analytics Architecture for Environmental Health](image)

(a) Traditional county-level outbreak detection and (b) Geographic features play an important role in outbreak situation awareness [St. Louis Post-Dispatch, 2013]. Detection, as cancer hotspots may be clustered in non-traditional ways (e.g., along creeks, rivers, valleys).

Figure 2: Coldwater Creek is adjacent to several sites in the St. Louis, MO area which were involved in the processing and recovery of uranium during World War II. These sites were contaminated with radioactive waste as a result of the processing and recovery activities. While most of the radioactive wastes have been cleaned up, citizens have expressed concern that exposure to the wastes has increased the number of cancer cases in the area [Yun et al., 2013].
Electronic medical records provide an incredibly rich source of massive spatiotemporal data that can be harnessed to analyze a wide variety of health-related research questions. EMRs provide patient addresses, which can be linked to specific locations on Earth or aggregated to known geographic entities like cities, counties, or ZIP codes. EMRs also provide time stamps on tests, procedures, and diagnoses given to patients. We can then exploit the spatial and temporal aspects of the data to analyze regional variations in health outcomes and diagnoses, to identify similar patterns of disease in disparate regions of the country and examine potential causes, to analyze geographic patterns in health care spending for new populations (e.g., not just Medicare patients), and to monitor populations for emerging communicable and chronic diseases. We need scalable, easy-to-use and understand tools in order to handle, process, and analyze the spatiotemporal data volumes generated by EMRs.

Exploratory analysis of Electronic Medical Records can be difficult due to the sheer number of variables (diseases) to examine for covariance. Co-location detection algorithms can highlight spatially-correlated events from a huge number of disease types, including temporal aspects to examine regional variations. These techniques can indicate when two specific disease types are spatially-clustered over a region. Dataset visualization techniques, such as KMR, can highlight geographic-feature patterns of interest in the data, moving beyond the standard elliptical summarization techniques commonly used in the area.

The types and volume of spatially explicit health surveillance data are exploding. Air and water quality sensor networks continuously report levels of health-adverse pollutants to a variety of groups, including government agencies, health care providers, and the general public. These data can be used to prepare hospitals for increases in patients suffering from respiratory illness or to warn patients who are susceptible to respiratory illness to avoid going outside. Social media data such from Twitter and other sources can be used to detect possible outbreaks of communicable diseases and allow agencies to precisely target resources to contain the outbreak.

However, these new spatially-explicit health datasets are exceeding the capacity of traditional spatial computing systems. Distributed filesystems and efficient processing, storing, and indexing methods are required before any analysis of the data can be done. We have shown preliminary work in this field with Spatial Hadoop, which provides spatial indexing over Hadoop / HDFS, significantly decreasing the access time of spatial queries. Fast and targeted access to these spatially explicit health datasets will provide functionality for a number of other analytics in this proposal, such as rapid geographic-feature summarization of diseases.

3.2 Why Spatio-Temporal is Key

Spatio-Temporal Big Data Analytics has transformative opportunities for environmental health. Spatial aspects, e.g., neighborhood context [Thomas et al., 2009], are critical in understanding many contributors to disease process including environmental toxicant exposure as well as human behavior and lifestyle choices. Public health data, such as the SEER database or the US Census, contain vast amounts of information with spatial components that can be used and analyzed in conduction with new datasets coming online (e.g., cell phones that record air quality). This exposome, a characterization of a person’s lifetime exposures, is becoming an increasingly popular subject of research for public health [Lioy and Rappaport, 2011, Rappaport, 2010]. For example, epidemiologists use spatial analysis techniques [AH et al., 2012] to identify cancer clusters [Pickle et al., 2006] (i.e., locations with unusually high densities) and track infectious disease such as SARS and bird flu.

3.3 Qualifications

Shashi Shekhar, a Distinguished University Professor, is a leading researcher in the area of geographic information systems (GIS), spatial databases, and spatial data mining. For outstanding contributions to these areas, he received the IEEE-CS Technical Achievement Award (2006) and was elected an IEEE Fellow (2003) as well as an AAAS Fellow (2008). He was also named a key difference-maker for the field of GIS by the most popular GIS textbook. He has a distinguished academic record that includes 250+ refereed papers, a popular textbook on Spatial Databases (Prentice Hall, 2003) and an authoritative Encyclopedia of GIS (Springer, 2008). He pioneered the research area of spatial data mining via pattern families (e.g. collocation, mixed-drove co-occurrence, and cascade), keynote speeches, survey papers and workshop organization. Furthermore, he contributed significantly to design of the UMN map server software for publishing geographic data on the Internet and the Crime-Stat 3.0 software to identify spatial patterns in geographic datasets. His NIH connections include consulting on a number of NIH grants (MPC), invited speaker to the NIH-AAG Symposium on Geospatial Frontiers in Health and Social Environments, organizer of the CCC visioning workshop on Spatial Computing 2020 with participation of researchers from NIH, NIEHS, and NCI.

4 Innovation

In general, geographic surveillance tasks go beyond outbreak detection to include situation awareness, trend detection, and case adjustment. Current methods in geographic surveillance, such as SatScan, GeoSurveillance, and related methods (e.g., CUSUM, scan statistics), have the ability to identify circular disease outbreaks in traditional public health data (e.g., counts of disease per county, time-series information, ellipsoidal clustering of disease reports). These methods have proven effective
but are limited in two significant dimensions when considering the rise of new non-traditional datasets: (1) limited ability to handle large volumes of non-traditional datasets (e.g., GPS trajectories of ozone readings, volunteered-geographic information, internet search queries, sensor networks), and (2) limited ability to find non-traditional outbreak patterns (e.g., clustering on spatial networks, co-location between diseases event types).

Hotspot detection, or disease outbreak detection from point data, is a popular technique in geographic surveillance. However, these techniques all focus on traditional hotspots, e.g., circle or ellipsoidal patterns. Few techniques are capable of detecting outbreaks along non-traditional features, such as roads, rivers, valleys, etc. We have published preliminary work in this area, finding dense clusters of point data along road networks. We propose to investigate geographic hot region detection as shown in Figure 3(b), finding outbreaks along geographic features as compared to administrative boundaries (e.g., counties).

Spatial associations between entity types (e.g., types of cancers that co-cluster) can currently be found in micro-data for point-based events. However, associations between events and geographic features (such as cancer reports with rivers) have not been explored in the literature. We propose to investigate point, line, and polygon spatial associations in micro-datasets along with exploring aggregate-datasets for potential spatial associations as shown in Figure 3(a).

---

5 Approach - Spatiotemporal Analytics and Computing

The specialty of spatial data mining originates from three central concepts in spatial statistics namely, spatial autocorrelation, spatial non-stationarity and spatial simpson’s paradox [Cressie, 1993, ?]. Spatial statistics is a branch of statistics that is concerned with the analysis and modeling spatial data [?]. Spatial statistics classifies spatial data into three basic types for ease of interpretation namely, (a) point referenced data, which is modeled a fixed collection of spatial locations, S in a two dimensional framework D (e.g. set of police stations in a metropolitan city); (b) areal data, which modeled as a finite set of irregular shaped polygons in a two dimensional framework D (e.g. set of police districts in a metropolitan city ) and ; (c) point process data which is modeled as a random collection of spatial events collectively referred to as the spatial point pattern over a two dimensional framework D (e.g. home locations of patients infected by a disease). Without loss of generality spatial statistics refers to any instance of spatial data as a spatial unit.

The explosive growth of spatial data and widespread use of spatial databases [Guting, 1994, Shekhar and Chawla, 2002, Shekhar et al., 1999, Worboys, 1995] have heightened the need for the automated discovery of spatial knowledge. Spatial data mining [Shekhar and Chawla, 2002, Stolorz et al., 1995] is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The size of some spatial data limits the usefulness of conventional statistical techniques for extracting spatial patterns. Efficient tools for extracting information from geo-spatial data are crucial to organizations which make decisions based on large spatial datasets.

Increasingly, the size, variety, and update rate 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, processed via spatial computing. Spatial Big Data provides the opportunity to solve some of the long-standing challenges in spatial computing which stemmed from lack of data. Here we describe four novel opportunities: estimating spatial neighbor relationships, supporting place-based ensemble models, simplifying spatial models, and on-line spatio-temporal data analytics.

These three fields all attempt to find new, interesting, and non-trivial patterns in spatial datasets, but specialize in different properties as shown in Figure ???. Spatial statistics provides traditional statistical interpretability of its findings (e.g., confidence levels), but has difficulty scaling to large datasets. This resulted in the field of spatial data mining, which relaxes some statistical properties (e.g., statistical significance) to increase ability to handle large datasets. Recently, big data has emerged which challenges even traditional spatial data mining techniques, where datasets require distributed computing for even simple analysis. Spatial Big Data Analytics is a novel field that finds novel algorithms and techniques to find interesting patterns in very large spatial datasets.
5.1 Spatial Mahout

We will develop Spatial Mahout, a robust, distributed computing spatial data mining software package to be used with Hadoop-like systems. Spatial Mahout’s goal is to build scalable spatial data analytics libraries for spatial big data, which 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), and edge-effects. As such, space-time models are often computationally more expensive than traditional models. Spatial Mahout offers core algorithms for spatial data analytics that will be initially implemented on top of Apache Hadoop using the map/reduce paradigm (future implementations will include non-Hadoop clusters). Examples of the spatial data analytics techniques provided by Spatial Mahout, include topological functions (overlap, union, intersect), neighborhood relationship support, raster data support (e.g., fast zonal statistics), hotspot detection, high-volume distance calculations, spatial Join and index (supporting statistics and data mining), etc.

Spatial Mahout addresses computational considerations such as space partitioning, multidimensional data structures, static and dynamic load balancing, etc. Spatial big data analytics algorithms usually require multiple iterations, which is handled by the proposed framework. Spatial Mahout also tackles computational issues that 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. Researchers can add this component to their distributed platform systems to provide scalable spatial data analytics on a variety of data types for use cases in surveillance, EMR and hypothesis generation, and disease mapping. This package will provide a number of useful analytical functions as shown in Table 1. These essential functions will enable development of advanced spatial analytics algorithms, two of which we detail below.

5.1.1 Spatio-Temporal Associations = Correlates

Traditional data mining has benefited from association rule mining to discover items that were bought together in transaction data ([Agrawal et al., 1994]). However, because 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. 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 pollutants and cancer types. Spatio-temporal 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 cancer datasets may highlight co-occurrence of different cancer types that indicate specific environmental toxins [Blumenstock et al., 2000].

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

Spatial co-location rule mining approaches can be grouped into two broad categories: approaches that use spatial statistics and algorithms that use association rule mining kind of primitives. Spatial statistics based approaches utilize statistical measures such as cross-K function, mean nearest-neighbor distance, and spatial autocorrelation. However, these approaches are computationally expensive. Association rule-based approaches focus on the creation of transactions over space so that an apriori like algorithm [Agrawal and Srikant, 1994] can be used. Transactions in space can use a reference-feature centric [Koperski and Han, 1995] approach or a data-partition [Morimoto, 2001] approach.

Spatio-Temporal Big Data Associations on Distributed Processing Systems: The computational challenges of the spatio-temporal co-location problem prompts us to explore distributed computing as a candidate solution paradigm. We now describe at a high level key steps in solving the co-location problem that exhibit parallel formulations. Figure ?? illustrates a simple spatial dataset exhibiting co-location patterns. The key operation in association mining is counting relationships between entities, which can be partitioned across distributed systems. We show in Figure ?? the process of parallelizing the Co-Location algorithm via data partitioning, accounting for edge cases when neighborhood relationships are lost. In Figure 6, we illustrate how the Co-Location algorithm itself can be parallelized, distributing the necessary item set counting across nodes and reporting those counts through the standard MapReduce protocol.

5.1.2 Spatio-Temporal Clustering = Hotspots

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 Spatio-temporal 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]). Spatio-temporal data mining techniques group objects based on both their spatial and temporal similarity ([Kisilevich et al., 2010]).

In network-based summarization, spatial objects are grouped using network (e.g., road) distance. Existing methods of network-based summarization such as Mean Streets [Celik et al., 2007], Maximal Subgraph Finding (MSGF) [Buchin et al.],
K-Main Routes (KMR) aims to maximize the number of activities covered on each k route. KMR employs an inactive node pruning algorithm where instead of calculating the shortest paths between all pairs of nodes, only shortest paths between active nodes and all other nodes in the spatial network are calculated. This results in computational savings (without affecting the resulting summary paths) that are reported in the experimental evaluation. The inputs of KMR include the following: 1) an undirected spatial network $G = (N, E)$, 2) a set of activities $A$ and 3) a number of routes, $k$, where $k \geq 1$. The output of KMR is a set of $k$ routes where the objective is to maximize the activity coverage of each $k$ route. Each $k$ route is a shortest path between its end-nodes and each activity $a_i \in A$ is associated with only one edge $e_i \in E$.

In this section, we discuss preliminary work of our group, a K-Main Routes (KMR) approach that finds a set of $k$ routes to summarize activities. KMR employs an inactive node pruning algorithm where instead of calculating the shortest paths between all pairs of nodes, only shortest paths between active nodes and all other nodes in the spatial network are calculated. This results in computational savings (without affecting the resulting summary paths) that are reported in the experimental evaluation. The inputs of KMR include the following: 1) an undirected spatial network $G = (N, E)$, 2) a set of activities $A$ and 3) a number of routes, $k$, where $k \geq 1$. The output of KMR is a set of $k$ routes where the objective is to maximize the activity coverage of each $k$ route. Each $k$ route is a shortest path between its end-nodes and each activity $a_i \in A$ is associated with only one edge $e_i \in E$.

Figure 7: Comparing KMR and Crimestat K-means output for $k = 4$ on pedestrian fatality data from Orlando, FL [Fatality Analysis Reporting System (FARS), May 7, 2013].
then proceeds in two main phases. First, it forms k groups by assigning each activity to its closest summary path. Then, it updates the summary path of each group by calculating the shortest path that maximizes activity coverage. Assigning and updating repeat until the summary paths no longer change and the final summary paths and groups are returned. Due to its iterative nature, KMR challenges current big data platforms such as MapReduce in terms of efficiently parallelizing computations. Because KMR’s assigning and updating steps repeat until the summary paths no longer change, the algorithm needs to use information from previous iterations 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. Previous approaches have explored parallelizing K-Means (e.g., Apache Mahout [2]). However, the update step in KMR is significantly more computationally intensive than that of K-Means. The reason is that in KMR all shortest paths in the spatial network may have to be evaluated to determine the new representative for each group. In K-Means, the update step only involves computing a new mean or center.

Research Task T1: Implement a control parallelism-based approach: An initial approach to parallelize KMR lies in distributing the tasks of each phase of the main loop of the algorithm across different computing nodes, i.e., a task or control parallelism approach. In this paradigm, certain tasks within the algorithm are parallelized with the intent of speeding up the overall (sequential) algorithm. Many copies of KMR begin executing on a cluster of machines and one of these act as the master, responsible for coordinating worker activity. In the first phase in which groups are formed, the master tasks each compute node with determining the closest summary path to each activity. Each worker has a local copy of the road network and is supplied with a summary path set and an activity that must be grouped with the closest summary path. Each worker reports the closest summary path for its activity. In this case, parallelism on the number of activities may be achieved. The master collects the group information for each activity from each worker and proceeds to the second phase. In the second phase in which representatives are updated, the master tasks each worker with calculating the representative for each group. Each worker is supplied with one group of activities and returns the shortest path which maximizes the sum of activities on the path. In this phase, parallelism on the order of number of the number of groups may be achieved. The master collects the new representative for each group from each worker and avails this information for the next iteration or terminates the computation if group representatives have not changed.

Research Task T2: Explore a data parallelism-based approach: In this task we explore distributing the data across different parallel computing nodes. A key component here is the provision of spatial indexes (e.g., R-trees, distributed partitioned R-trees) to help improve the I/O cost of queries retrieving a small part of the data file. Our initial efforts include Spatial Hadoop [3], 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.

5.1.3 Spatio-Temporal Anomalies = Inconsistencies

Outliers have been informally defined as observations in a data set which appear to be inconsistent with the remainder of that set of data Barnett and Lewis [1994], or which deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism Hawkins [1980]. The identification of global outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as detection of credit card fraud and voting irregularities. This section focuses on spatial outliers, i.e., observations which appear to be inconsistent with their neighborhoods Pei et al. [2006], Sun and Chawla [2004], Wu et al. [2007]. Detecting spatial outliers is useful in many applications of geographic information systems and spatial databases. These application domains include transportation, ecology, homeland security, public health, climatology, and location-based services.

A spatial outlier Shekhar et al. [2001] is a spatially referenced object whose non-spatial attribute values differ significantly from those of other spatially referenced objects in its spatial neighborhood. Informally, a spatial outlier is a local instability (in values of non-spatial attributes) or a spatially referenced object whose non-spatial attributes are extreme relative to its neighbors, even though the attributes may not be significantly different from the entire population. For example, a new house in an old neighborhood of a growing metropolitan area is a spatial outlier based on the non-spatial attribute house age.

5.1.4 Spatio-Temporal Prediction = Recommendations

Classification and regression are similar types of patterns in data mining. Given a sample set of input-output pairs, the objective of supervised learning is to find a function that learns from the given input-output pairs, and predicts an output for any unseen input (but assumed to be generated from the same distribution), such that the predicted output is as close as possible to the desired output. For example, in remote sensing image classification, the input attribute space consists of various spectral bands or channels (e.g., blue, green, red, infra-red, thermal, etc.) The input vectors (x_i’s) are reflectance values at the i^th location in the image; and the outputs (y_i’s) are thematic classes such as forest, urban, water, and agriculture. Depending on the type of output attribute, two supervised learning tasks can be distinguished:

The fact that classical data mining techniques ignore spatial autocorrelation and spatial heterogeneity in the model-building process is one reason why these techniques do a poor job. A second, more subtle but equally important reason is related to the choice of the objective function to measure classification accuracy. For a two-class problem, the standard way to measure classification accuracy is to calculate the percentage of correctly classified objects.
Common Methods: Several previous studies Jhung and Swain [1996], Solberg et al. [1996] have shown that the modeling of spatial dependency (often called context) during the classification process improves overall classification accuracy. Spatial context can be defined by the relationships between spatially adjacent pixels in a small neighborhood. An example spatial framework and its four-neighborhood contiguity matrix is shown in Figure ???. In this section we present two spatial data mining techniques, namely the Logistic Spatial Autoregressive Model (SAR) and Markov Random Fields (MRF).

Logistic Spatial Autoregressive Model(SAR): Logistic SAR decomposes a classifier \( \hat{f}_C \) into two parts, namely spatial autoregression and logistic transformation. Spatial dependencies are modeled using the framework of logistic regression analysis. In the spatial autoregression model, the spatial dependencies of the error term, or, the dependent variable, are directly modeled in the regression equation Anselin [1988a]. If the dependent values \( y_i \) are related to each other, then the regression equation can be modified as

\[
y = \rho Wy + X\beta + \epsilon.
\]  

(1)

Here \( W \) is the neighborhood relationship contiguity matrix and \( \rho \) is a parameter that reflects the strength of the spatial dependencies between the elements of the dependent variable via the logistic function for binary dependent variables.

The massive sizes of geospatial datasets in many application domains make it important to develop scalable parameter estimation algorithms of the SAR model solutions for location prediction and classification. As noted previously, many classical data mining algorithms, such as linear regression, assume that the learning samples are independently and identically distributed (i.i.d.). This assumption is violated in the case of spatial data due to spatial autocorrelation Anselin [1988b] and in such cases classical linear regression yields a weak model with not only low prediction accuracy Chawla et al. [2001], Shekhar et al. [2002] but also residual error exhibiting spatial dependence. Modeling spatial dependencies improves overall classification and prediction accuracies significantly.

However, estimation of SAR model parameters is computationally very expensive because of the need to compute the determinant of a large matrix in the likelihood function Kazar et al. [2004], Li [1996], Pace and LeSage [2000, 2002, 2003]. The Maximum Likelihood function for SAR parameter estimation contains two terms, a determinant term and an \( SSE \) term (Equation 2). The former involves computation of the determinant of a very large matrix, which is a well-known hard problem in numerical analysis. To estimate the parameters of a ML-based SAR model solution, the log-likelihood function can be constructed, as shown in (2). The estimation procedure involves computation of the logarithm of the determinant of (log-det) a large matrix, i.e. \((I - \rho W)\).

\[
\ell(\rho | y) = -\frac{2}{n} \ln |I - \rho W| + \ln((I - \rho W)y)^T(I - x(x^Tx)^{-1}x^T)(I - x(x^Tx)^{-1}x^T)(I - \rho W)y\]  

SSE

(2)

For example, the exact SAR model parameter estimation for a very small 10,000-point spatial problem can take tens of minutes on common desktop computers. Computation costs make it difficult to use SAR for important spatial problems which involve millions of points, despite its promise to improve prediction and classification accuracy.
In the equation, \( y \) is the \( n \)-by-1 vector of observations on the dependent variable, where \( n \) is the number of observation points; \( \rho \) is the spatial autoregression parameter; \( W \) is the \( n \)-by-\( n \) neighborhood matrix that accounts for the spatial relationships (dependencies) among the spatial data; \( x \) is the \( n \)-by-\( k \) matrix of observations on the explanatory variable, where \( k \) is the number of features; and \( \beta \) is a \( k \)-by-1 vector of regression coefficients. Spatial autocorrelation term \( \rho Wy \) is added to the linear regression model in order to model the strength of the spatial dependencies among the elements of the dependent variable, \( y \).

## 6 Validation

The research ideas in this proposal will be evaluated in the following settings, using the following data sets, methodologies, and goals/questions.

**Settings.** Evaluation will be conducted in-lab, in-software system, and in-field. In-lab, simulation of the proposed ideas will be performed followed by an evaluation of the proposed ideas using different workloads. In-software system, the research ideas will be realized and validated inside the open-source database management system PostgreSQL [postgresql], along with its spatial extension PostGIS [postgis] and routing extension pgRouting [pgRouting]. In-field, the ideas will be validated via vehicle computers and using cellphone applications and webservices.

**Datasets.** All research ideas will be validated using temporally detailed roadmaps, engine measurements, and GPS-tracks. The temporally detailed roadmaps have been provided by NAVTEQ (cite navteq) (see letter of support) and 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). For engine measurement data, Oak Ridge National Laboratories has provided a letter of commitment and their interest in this topic. GPS-trace data such as is illustrated in Figure ?? is also available.

**Methodologies.** The general methodologies will entail field experiments on vehicle computers and on handheld or mobile devices, as well as software prototypes. Additional methodologies will include simulation and experimentation using benchmark data as well as analytical proofs.

**Goals/Questions.** The goals and questions will include a comparative analysis where both quality and scalability of various research ideas will be assessed. For quality, a comparison of new routes to baselines gleaned from current navigation systems using the metrics of travel time, fuel use, and GHG emissions will be done. Scalability will be evaluated by varying and observing the effect of various workload parameters.

## 7 Software Development

### References


