Project Summary

The goal of geo-spatial data mining is to discover interesting and useful but implicit geo-spatial patterns. A historical example of geo-spatial pattern relates to the 1855 epidemic of Asiatic Cholera in London, England. At that time, it was not known that cholera is a water-borne disease. A leading epidemiologist marked the locations of residences of cholera victims on a map of London. Surprisingly, the locations formed a nebulae, the centroid of which turned out to be a water-pump. When the water-pump was shut-down, the epidemic began to subside. Thus using simple spatial exploratory analysis, both the source and the carrier of the epidemic were simultaneously identified.

Geo-spatial data mining is crucial to organizations making geo-spatial decisions. For example, public health organizations are interested in geo-spatial patterns in the spread of infectious diseases. Private companies are interested in geo-spatial patterns in consumer spending for marketing as well as logistical reasons. A special kind of geo-spatial data mining problem is Location Prediction. Public safety organizations are interested in *location prediction* of crimes to plan police patrols. Ecological and environmental organizations may be interested in *location prediction* for protecting bio-diversity.

The current approach towards solving spatial data mining problems is to use classical data mining tools after "materializing" spatial relationships. For example, materializing distance-to-nearest-water-pump for residences of cholera patients would allow classical data mining techniques(e.g. regression) to identify distance to water pumps as an important explanatory variables. However, many classical data mining techniques including regression assume that the learning samples are drawn from identical and independent distributions(i.i.ds). In other words, the data about a cholera patient is independent(spatially) of the data describing other patients. This assumption is not true for spatial attributes, e.g. residence location or distance to water pump. The property of samples affecting other sample values in the neighborhood is called spatial autocorrelation.

In addition, classical data mining maximizes classification accuracy even though *spatial accuracy* may be more important for location prediction.

Invalid i.i.d assumptions will make classical techniques yield a weaker model(low R-square) and may overestimate the influence of various factors. Spatial statistics, a branch of classical statistics, has explored new parametric model, e.g. spatial autocorrelation based regression(SAR), to account for spatial autocorrelation. These models improve both the classification and spatial accuracy. However, they are computationally extremely expensive, and do not scale up to large datasets.

The goal of this proposal is to define and explore efficient techniques for location prediction. This proposal formally defines the location prediction problem. It describes the experience of using classical techniques along with their limitations. It proposes PLUMS, an efficient new technique for geo-spatial data mining driven by "map-similarity" measures which is a linear combination of classification and spatial accuracy measures. Preliminary results show that: (1) The proposed technique yields comparable spatial accuracy using order of magnitude less computational resources and (2) PLUMS can be a scalable technique for large scale location prediction problems where spatial accuracy is more important than classification accuracy.

An intriguing question raised by this proposal is whether PLUMS can achieve comparable classification accuracy as well as spatial accuracy relative to spatial statistical methods(e.g. SAR) using orders of magnitude less computational resources ? The proposed work aims to explore the behavior of PLUMS in greater detail by studying the effect of alternative choices related to design issues(e.g., function families, map similarity measures, discretization, incremental computation) on the solution quality. It proposes specific performance tuning tasks to make PLUMS scalable.

TABLE OF CONTENTS

For font-size and page-formatting specification, see GPG Section II.C.

Appendix Items:

1. Map data from Stubble marsh land
2. Map similarity and spatial accuracy measure

[0]Proposers may select any numbering mechanism for the proposal. The entire proposal, however must be paginated. Complete both columns only if the proposal is numbered consecutively.

Contents

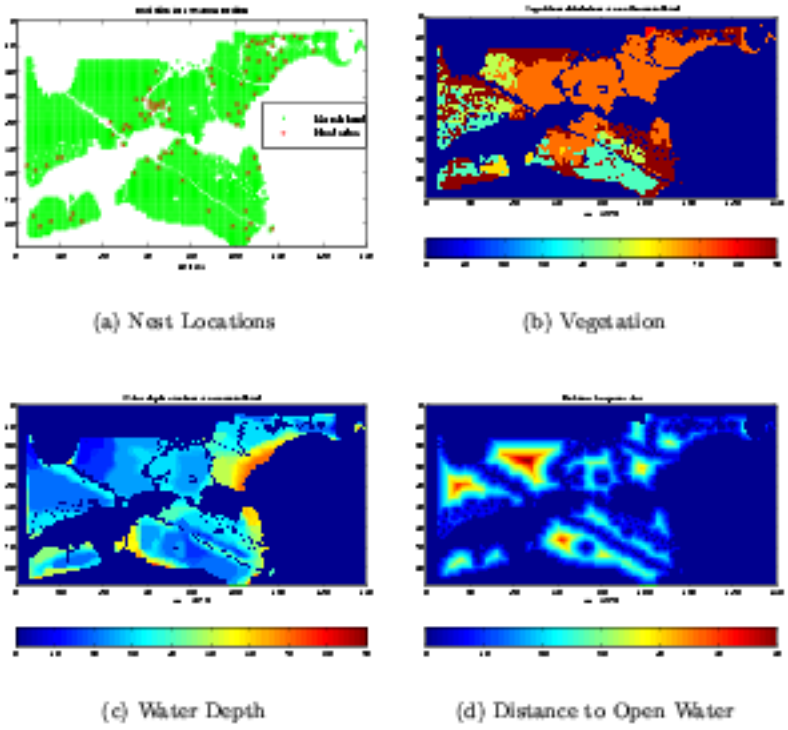
1 Introduction

Widespread use of spatial databases [24, 48, 54, 71] is leading to an increasing interest in mining interesting and useful but implicit spatial patterns[34, 40, 21, 47]. Efficient tools for extracting information from geospatial data, the focus of this work, are crucial to organizations which make decisions based on large geospatial data sets. These organizations are spread across many domains including ecology and environment management, public safety, transportation, public health, business logistics, travel and tourism. [2, 26, 30, 36, 46, 64, 72].

Classical data mining algorithms [1, 19] often make assumptions(e.g. independent, identical distributions), which violate the first law of Geography: everything is related to everything else but nearby things are more related than distant things [12, 66]. In other words, the values of attributes of nearby spatial objects tend to systematically affect each other. In spatial statistics, an area within statistics devoted to the analysis of spatial data, this is called spatial autocorrelation [13]. Ignoring spatial autocorrelation may lead to residual errors that vary systematically over space exhibiting high spatial autocorrelation. The models derived may turn out to be not only biased and inconsistent but may also be a poor fit to the data set. Spatial statistical methods [3, 37], however, are computationally expensive due to their reliance on contiguity matrices which can be larger than the spatial datasets being analyzed.

Many spatial data mining problems, e.g. location prediction, can benefit from new computationally efficient techniques, which maximize spatial accuracy instead of classification accuracy. This proposal develops computationally efficient techniques for the location prediction problem to maximize map similarity, which is a combination of classification accuracy and spatial accuracy.

1.1 An Illustrative Application Domain



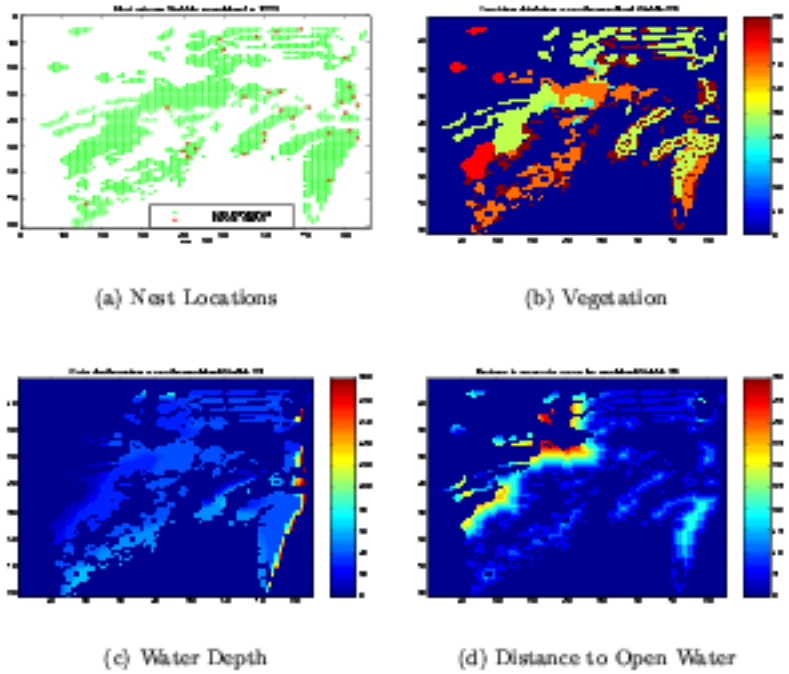
(a) Nest Locations (b) Vegetation

(c) Water Depth (d) Distance to Open Water

Figure 1: (a) Learning dataset: The geometry of the marshland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the marshland, (c) The spatial distribution of *water depth*, and (d) The spatial distribution of *distance to open water*.

The availability of accurate spatial habitat models is an important tool for wildlife management, protection of critical habitat and endangered species. Since the underlying process governing the interaction between wildlife and environmental factors is complex, statistical models are built to gain some insight on the basis of data collected during field work. One of our close collaborators has been involved in the development of spatial model for the nesting locations of a marsh-nesting bird species [43, 44]. We will use this application, and the accompanying data, to explain the location predication problem and its unique aspects *vis-a-vis* classical data mining.

The learning and testing datasets that we will use was collected in 1995 and 1996 from two marshlands(Darr and Stubble) located on the shores of Lake Erie in Ohio. For the purpose of data collection, a local coordinate system was established for each marshland and a regular grid consisting of approximately 5000 cells was superimposed. The cells of the grid had a square geometries of size 5 meters by 5 meters. In each cell the values of several structural and environmental factors were recorded, including *water depth*, *dominant vegetation durability index and distance to open water*. These three factors play the role of most significant explanatory variables. At each cell was also recorded the fact whether a bird-nest(red-winged blackbird) was present or not. The presence of the nest played the role of dependent variable. The geometry of the Darr marshland, locations of the nests and spatial distribution of the explanatory variables are shown in Figure 1. The corresponding maps for the Stubble marshland are shown in Figure 2.



(a) Nest Locations (b) Vegetation

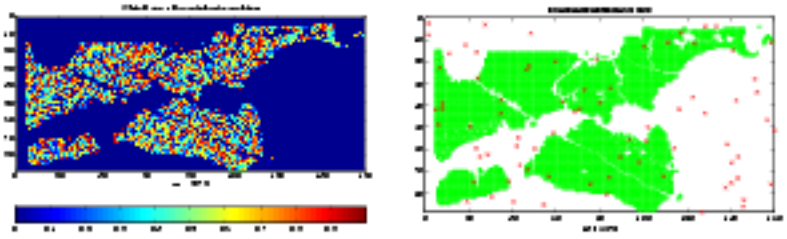
(c) Water Depth (d) Distance to Open Water

Figure 2: (a) The geometry of the marshland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the marshland, (c) The spatial distribution of *water depth*, and (d) The spatial distribution of *distance to open water*.

Our collaborators have applied classical data mining techniques like logistic regression[44] and neural networks[43] to build spatial habitat models. Logistic regression was used because the dependent variable is binary(nest/no-nest) and the logistic function "squashes" the real line onto the unit-interval. The values in the unit-interval can then be interpreted as probabilities. They concluded that using logistic regression the nests could be classified at a 24% rate better than random[43]. The use of neural networks actually decreased the classification accuracy[43] but led to a better understanding of the interaction between the explanatory and the dependent variable.

Detailed discussions with our collaborators reveal two important reasons why, despite extensive domain

knowledge, the results of classical data mining are not "satisfactory". First, classical techniques, e.g. logistic regression, make assumption about identical independent distribution for the properties of each pixel, ignoring spatial autocorrelation. Figure 3(a) shows a spatial distribution consistent with assumption of classical regression. It looks like "white noise" as properties of pixel are generated from independent identical distributions. Note that the maps of explanatory variable in Figure 1 have much more gradual variation indicating high spatial autocorrelation. Figure 3(b) shows a random distribution of nest locations which is quite different from the distribution of actual nests shown in Figure 1(a).



(a) pixel property with independent identical distribution

(b) Random nest locations

Figure 3: Spatial distribution satisfying distribution assumptions of classical regression

A second, more subtle but equally important reason is the objective function of classification measure accuracy. For a two-class problem the standard way to measure classification accuracy is to use the total square error. This measure may not be the most suitable for spatial data. Spatial accuracy is as important in this application domain due to the effects of discretizations of continuous marsh into discrete pixels, as shown in Figure 4. Figure 4(a) shows the actual locations of nests and 4(b) shows the pixels with actual nests. Note the loss of information during the discretization of continuous space into pixels. Many nest location barely fell within the pixels labeled 'A' and were quite close to other pixels with label of no-nest. Now consider two predictions shown in Figure 4(c) and 4(d). Domain scientists prefer prediction 4(d) over 4(c), since predicted nest locations are closer on average to some actual nest locations. Classification accuracy measure cannot distinguish between 4(c) and 4(d), and one needs a measure of spatial accuracy to capture this preference.

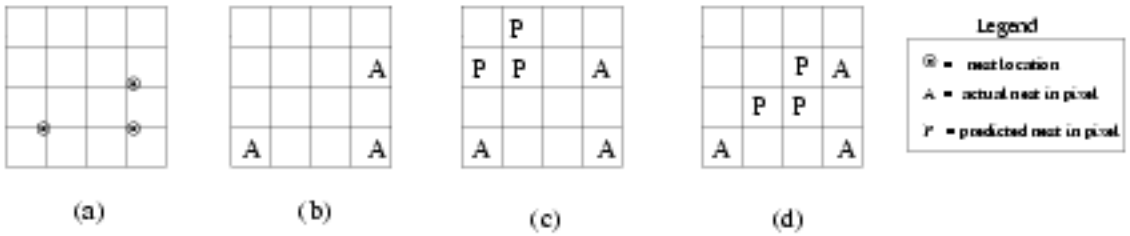


Figure 4: (a)The actual locations of nest, (b)Pixels with actual nests, (c)Location predicted by a model, (d)Location predicted by another mode. Prediction(d) is spatially more accurate than (c).

1.2 Location Prediction: Problem Formulation

The Location Prediction problem is a generalization of the nest location prediction problem. It captures the essential properties of similar problems from other domains including crime prevention and environmental management. The problem is formally defined as follows:

Given :

- A spatial framework S consisting of sites $\{s_1, \ldots, s_n\}$ for an underlying geographic space G.

- A collection of explanatory functions $f_{X_k} : S \rightarrow R^k, k = 1, \ldots K$. R^k is the range of possible values for the explanatory functions.
- A dependent function $f_Y : S \rightarrow R^Y$
- A family $\mathcal{F}$ of learning model functions mapping $R^1 \times \ldots R^K \rightarrow R^Y$.

Find : A function $\hat{f}^Y \in \mathcal{F}$.

Objective : maximize similarity$(map_{s_i \in S}(\hat{f}^Y(f_{X_1}, \ldots, f_{X_K})), map(f_Y(s_i)))$
$= (1-\alpha)$classification_accuracy$(\hat{f}^Y, f_Y) + (\alpha)$spatial_accuracy$((\hat{f}^Y, f_Y)$

Constraints :

1. Geographic Space S is a multi-dimensional Euclidean Space [1].
2. The values of the explanatory functions, the f_{X_k}'s and the response function f_Y may not be independent with respect to those of nearby spatial sites, i.e. spatial autocorrelation exists.
3. The domain R^k of the explanatory functions is the one-dimensional domain of real numbers.
4. The domain of the dependent variable, $R^Y = \{0, 1\}$.

The above formulation highlights two important aspects of location prediction. It explicitly indicates that (i) the data samples may exhibit spatial autocorrelation and, (ii) an objective function i.e., a map similarity measure is a combination of classification accuracy and spatial accuracy. The *similarity* between the dependent variable f_Y and the predicted variable $\hat{f}^Y$ is a combination of the traditional accuracy" and a representation dependent "spatial classification" accuracy. The regularization term α controls the degree of importance of **spatial accuracy** and is typically domain dependent. As $\alpha \rightarrow 0$, the map similarity measure approaches the traditional classification accuracy measure. Intuitively, α captures the spatial autocorrelation dependent in the data.

The study of nesting location of red-winged black bird [43, 44] is an instance of the location prediction problem. The underlying spatial framework is the collection of 5m*5m pixels in the grid imposed on marshes. Explanatory variables, e.g. water depth, vegetation durability index, distance to open water, map pixels to real numbers. Dependent variable, i.e. nest locations, maps pixels to a binary domain. The explanatory and dependent variables exhibit spatial autocorrelation, e.g. gradual variation over space, as shown in Figure 1 and 2. Domain scientist prefer spatially accurate predictions which are closer to actual nests, i.e, $\alpha > 0$.

1.3 Impact of Proposed Research

The proposed research will develop new techniques for discovering spatial patterns. These techniques will be based on map similarity measures incorporating spatial properties of application domains. These techniques also will be scalable and computationally efficient to quickly discover candidate patterns for further exploration via more rigorous techniques. Availability of such techniques will help researches and policy makers in many organizations, as explained in a letter of support from our collaborators in Ecology (See Attachment in section I). Location prediction has the potential of having a profound impact in many application areas including public health, public safety, ecology, environment and political-science. The spatial tracking of the spread of disease is an important concern for governments at the national, state and local level [72]. Location prediction can be important tool for health officers as it provides the ability to predict areas most vulnerable to outbreaks of specific contagious diseases.

Location prediction may be helpful in finding local patterns, e.g. hot spots, in crime databases for assigning police officers to different locations. The Crime Mapping Laboratory[14] at the National Institute of Justice has identified hot spot analysis as a critical area for research. Use of classical data mining techniques will yield weaker models and weaker insights, since spatial data exhibits autocorrelation and discretization make total per pixel square error less appropriate.

[1] The entire surface of the Earth cannot be modeled as a Euclidean space but locally the approximation holds true.

1.4 Summary of Proposed Research

The proposed project has three main objectives. The first objective is to develop, evaluate and implement a set of scalable geo-spatial data mining techniques for location prediction. This includes development of map similarity and spatial accuracy measures, families of parametric functions with spatial auto-correlation terms, discretization of parameter space and search algorithms to find parameter values optimizing map similarity. We will explore map similarity and spatial accuracy measures for raster maps as well as vector maps. We will also study techniques to estimate generalization error [69] during learning towards avoiding over-fitting problems. Finally, we will characterize the family of problems which can benefit from the new techniques developed in this research.

The second objective is to integrate these location prediction techniques into a comprehensive software system with software tools to provide kernel function primitives of a spatial data mining system. We may integrate our techniques with popular data mining software packages e.g. S-Plus [70] and other spatial analysis software packages.

The third objective is to evaluate this system using a spatial habitat modeling application that serves researchers and policy makers in the area of ecology and environment management.

2 Results from Prior Support

NSF funded a project titled "Databases for Spatial Graph Management Systems" (IRI-9631539) for P.I. in the recent past. This section provides a project summary, objectives, indication of success and project impact.

Summary: Databases for spatial graph management are very important for a large number of applications including transportation, utilities (e.g. gas, electricity) and urban management. This project addressed the needs of spatial graph management in the areas of physical database design and query processing. The main objective was to develop, evaluate and implement a novel spatial graph storage and access method, called CCAM, based on graph-connectivity. CCAM assigns the nodes of a graphs to disk pages via the graph partitioning approach to maximize the CRR, i.e., the chances that a pair of connected nodes are allocated to a common page of the file. CCAM includes a secondary index structure for facilitating fast searches based on node-id. CCAM supports operations of, insert(), delete(), create(), and find(), as well as new operations, get-a-successor() and get-successors(), which retrieve one or all successors of a node to facilitate graph analysis algorithms. Another objective was to develop scalable spatial graph-clustering algorithms as well as incremental reorganization strategies to enhance CCAM. This project also explored strategies to enhance existing geometric access methods for managing connectivity properties along with proximity properties. This techniques developed in the project were evaluated with benchmark graphs from Advanced Traveler Information System applications. The research plan included following tasks, (a) Incremental Reorganization Policies and Page-Access Graph, (b) Managing Connectivity Properties along with Proximity Properties, (c) An External Graph-Partitioning Algorithm, (d) Improved Cost Model and Partitioning Algorithms for Get-Successors(), and (e) Experimental Evaluation .

Accomplishments: The project accomplished the following : (i) Development of CCAM [39, 57, 59], a clustering method based on min-cut graph partitioning idea as well as performance evaluation characterizing its advantages over geometric clustering methods for spatial graph queries; (ii) Development and evaluation of three incremental reorganization strategies for CCAM; (iii) Techniques to take advantage of graph-connectivity information within conventional geometric access methods to enhance their performance; (iv) Development and implementation of prototype of CCAM storage and access method for spatial graphs, and (v) Experimental evaluation with ATIS spatial graphs, e.g. major roads in Minneapolis and Twin Cities. All illustrative map showing its major results appeared on the back cover of the proceedings of NSF/IDM PI workshop 1999. It also lead to improved understanding of clustering and access methods for spatial graphs. One of the interesting insight showed that CCAM was optimized for get-a-successor() operation and one needed a min-cut hyper-graph partitioning technique to be optimal for get-successors() operations. We worked on a novel hyper-graph partitioning algorithm, HMETIS, to experimentally evaluate the gains

obtained from moving to a hyper-graph model. The project led to exploration of graph partitioning for other problems in databases including declustering problems, and join-index based join algorithms. Finally, the project led to several refereed journal and conference publications [32, 58, 59, 62, 63], as well as refereed conference papers [31, 39, 57, 60, 61].

Impact: The project partially supported the Ph. D. thesis of Dr. Duen-Ren Liu, Xuan Liu and C.T. Lu and the M. S. thesis of Mr. Rajat Aggarwal as well. Dr. Liu is currently with the faculty of Inst. of Info. Management at National Chiao Tung University in Taiwan. Mr. Aggarwal is currently with Lattice Semiconductor Corporation. The project supported development of a regular graduate course on Spatial Databases and a seminar course on Databases for spatial graph management at the Computer Sc. Dept. in the University of Minnesota. Each course was attended by about 15 students from Computer Sc. as well as application areas such as Transportation.

CCAM attracted attention of organization developing "network engines" to manage large spatial graphs. P.I. was invited by Environmental Systems Research Inst., the largest company in GIS software industry, to advise on the storage methods for spatial graphs such as road-map. United Nations Development Program selected the P.I. to act as an expert advisor for the spatial database component in the local level development project, 1997-98. National Center for Geographic Information and Analysis invited P.I. as an expert on "navigable road-map databases" for advising California Dept. of Transportation center for interoperability.

The project is leading to exploration of graph partitioning for other problems in databases including declustering problems [58, 60, 62, 63], and join-index based join algorithms [61]. It also lead to min-cut hyper-graph partitioning algorithms needed to optimize CCAM for get-successors() operations. We developed a novel hyper-graph partitioning algorithm, HMETIS [31, 32], in collaboration with Prof. V. Kumar, which benefits VLSI circuit partitioning.

3 Preliminary Results

Recall that we proposed a general problem definition for the Location Prediction problem, with the objective of maximizing "map similarity", which combines spatial accuracy and classification accuracy. In this section, we propose the PLUMS approach and describe preliminary results from a pair of experiments which compare PLUMS with related work.

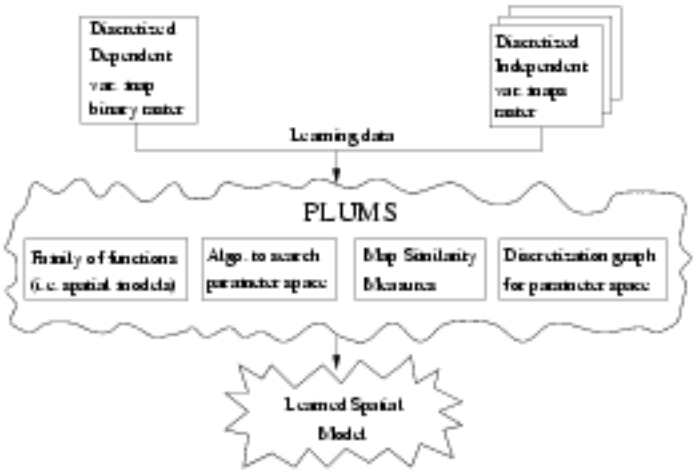


Figure 5: The framework for the location prediction process

3.1 Proposed Approach: Predicting Locations Using Map Similarity(PLUMS)

Predicting Locations Using Map Similarity(PLUMS) is the proposed supervised learning approach. Figure 5 shows the context and components of PLUMS. It takes a set of maps for explanatory variables and a map for the dependent variable. The maps must use a common spatial framework, i.e. common geographic space and common discretization, and produces a "learned spatial model" to predict the dependent variable using explanatory variables. PLUMS has four basic components, namely, a map similarity measure, a family of parametric functions representing spatial models, a discretization of parameter space, and a search algorithm. PLUMS uses the search algorithm to explore the parameter space to find the parameter value tuple which maximize the given map similarity measure. Each parameter value tuple specifies a function from the given family as a candidate spatial model. model functions.

A simple map similarity measure focusing on spatial accuracy for nest-location maps(or point sets in general) is the average distance from an actual nest site to the closest predicted nest-site. Other spatial accuracy and map similarity measures can be defined using nearest neighbor index [15], principal component analysis of a pair of raster maps [51] etc. as discussed in proposed work.

Algorithm 1 greedy-search-algorithm

```
parameter-value-set find-A-local-maxima(parameter-value-set PVS, discretization-of-parameter-space SF,
                                        map-similarity-measure-function MSM, learning-map-set LMS) {
    parameter-value-set best-neighbor, a-neighbor;
    real best-improvement=1, an-improvement;
    while(best-improvement > 0) do {
        best-neighbor = PVS.get-a-neighbor(SF);
        best-improvement = MSM(best-neighbor,LMS) - MSM(PVS,LMS);
        foreach a-neighbor in PVS.get-all-neighbors(SF) do {
            an-improvement = MSM(a-neighbor,LMS) - MSM(PVS,LMS);
            if(an-improvement > best-improvement) {
                best-neighbor = a-neighbor; best-improvement = an-improvement;
            }
        }
        if (best-improvement > 0) then PVS=best-neighbor;
    } /* found a local maxima in parameter space */
    return PVS;
}
```

A special case of PLUMS using greedy search is described in Algorithm 1. The function "find-A-local-maxima", takes a seed value-tuple of parameters, a discretization of parameter space, a map-similarity function and a learning data set consisting of maps of explanatory and dependent variables. It evaluates the parameter-value tuple in the immediate neighborhood of current parameter-value tuple in the given discretization. An example of a current parameter-value tuple in a red-winged-black bird application with 3 explanatory variables is (a,b,c). Its neighborhood may include the following parameter value tuples: (a+δ,b,c), (a-δ,b,c),(a,b+δ,c),(a,b-δ,c),(a,b,c+δ), (a,b,c-δ) given a uniform grid with cell-size δ discretization of parameter space. A more sophisticated discretization may use non-uniform grids. PLUMS evaluates the map similarity measure on each parameter value tuple in the neighborhood. If some of neighbors have higher values for the map similarity measure, the neighbor with highest value of map similarity measure is chosen. This process is repeated and it ends when no neighbor has a higher value of map similarity measure, i.e., a local maxima has been found. Clearly, this search algorithm can be improved using a variety of ideas including gradient descent [9, 20] and simulated annealing [53, 69] etc. We plan to explore alternative search algorithms in the proposed work. A simple function family is the family of generalized linear models, e.g. logit, probit [37] with or without autocorrelation terms. Other interesting families include non-linear functions. In the spatial statistics literature many functions have been proposed to capture the spatial autocorrelation property. For example, Econometricians use the family of spatial autoregression models [3, 38], Geo-statisticians [30] use Co-Kriging and Ecologists [28] use the Auto-Logistic models. Table 1 summarizes several special cases of PLUMS by enumerating various choices for the four components.

The design space of PLUMS is shown in Figure 7. Each instance of PLUMS is a point in the four dimensional conceptual space spanned by *similarity measure*, *family of functions*, *discretization of parameter space* and *external search algorithm*. For example, the PLUMS implementation labeled **A** in Figure 7

corresponds to the spatial accuracy measure(ADNP), generalized linear model(for the family of functions), a greedy search algorithm and uniform discretization.

PLUMS Component Choices	
Component	Choices
Map similarity	avg. distance to nearest prediction from actual, nearest neighbor index, ...
Search algorithm	greedy, gradient descent, simulated annealing, ...
Function family	generalized linear(GL) (logit, probit), non-linear, GL with autocorrelation
Discretization of parameter space	Uniform, non-uniform, multi-resolution, ...

Table 1: PLUMS Component Choices

3.2 Experimental Evaluation

We have carried out experiments to compare the classical regression, spatial autoregressive regression(SAR) models and an instance of the PLUMS framework. The data-set used for the learning portion of the experiments, i.e., to predict locations of bird-nests, is shown in Figure 1. The data-set used for testing the models is shown in Figure 2. The two data-sets are similar except for their spatial locations. Explanatory variables in these data-sets are defined over a spatial grid of approximately 5000 cells.

Before we show and analyze the results we briefly describe the classical and the SAR models. More information can be obtained from [3, 38]. In classical linear regression the relationship between the dependent variable $\mathbf{y}$ and the explanatory variables $\mathbf{X}$ is governed by the equation,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Since the dependent variable is binary the models have to be transformed so that the value of the dependent variable lie in the $[0, 1]$ interval where they can be interpreted as probabilities. The transformation can be effected by using logit or probit model [37, 42] which are instances of family of regression models called the Generalized Linear Models(GLIM). The standard assumptions on $\boldsymbol{\epsilon}$, the error term, is that the errors are i.i.d. This assumption, as we have shown, is invalid in the presence of spatial autocorrelation. One way of incorporating the spatial autocorrelation property is to replace the linear regression model with the spatial autoregressive regression(SAR) model:

$$\mathbf{y} = \rho W\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

This model assumes that the y variable is dependent on the X as well as itself. Conceptually the relationship between the y can be described as $y_i = f(y_j)\ i \neq j$. The self-relationship(autocorrelation) between the y_i's is defined in the contiguity matrix W. The contiguity matrix captures the apriori information about the spatial domain and the structure of the relationship between the y_i's. The simplest such relationship is that the y_i's are only dependent on their immediate neighbors. For example, two common choices are the four and the eight neighborhood relationship. Thus given a lattice structure and a point S in the lattice, a four-neighborhood assumes that S influences all cells which share an edge with S. In an eight-neighborhood it is assumed that S influences all cells which either share an edge or a vertex. An eight neighborhood contiguity matrix is shown in Figure 6. The contiguity matrix of the uneven lattice(left) is shown on the right hand side. Note that the size of W for our dataset of 5000 pixel is approximately 25 million. However only about 20,000 entries are non-zero.

Figure 6: A spatial map with 4 objects and its contiguity matrix W

For the purpose of the preliminary evaluations, we also evaluate PLUMS(A), an instance of PLUMS mplementation A in Figure 7. PLUMS(A) is implemented using a greedy search algorithm described in Algorithm 1. We use a map-similarity based purely on spatial accuracy (i.e. $\alpha = 1$), measured by average distance of nearest predicted location from an actual location. A uniform discretization of parameter space is used.

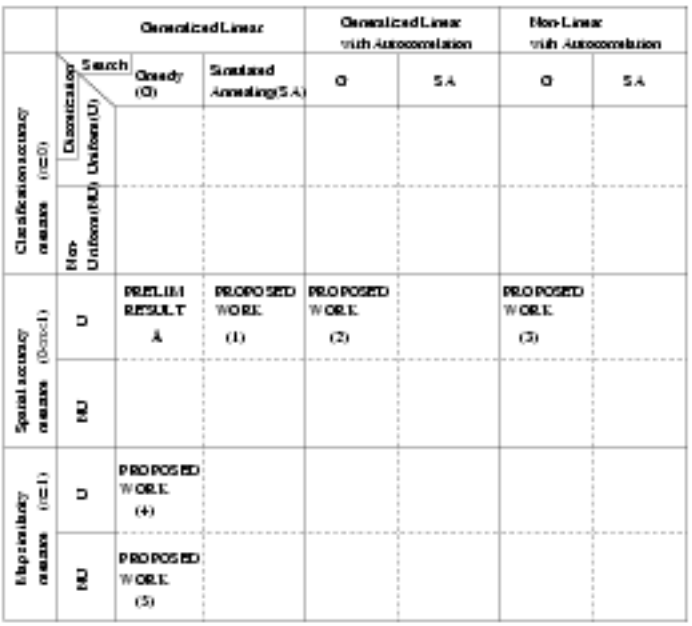


Figure 7: Space of design choices for PLUMS components:*function family*, *map-similarity measure*, *search algorithms* and *discretization*. **G** refers to Greedy search and **SA** refers to Simulated Annealing. **U** and **NU** refer to uniform and non-uniform grid based discretization of parameter space respectively.

3.2.1 Comparison of spatial accuracy and learning-time

We now compare logistic regression, SAR and PLUMS(A) on spatial accuracy as measured by ADNP(Average Distance to Nearest Prediction), which is defined as

$$ADNP(A, P) = \frac{1}{K} \sum_{k=1}^{K} d(A_k, A_k.nearest(P)).$$

Here the A_k's are the actual nest locations, P is the map layer of predicted nest locations and $A_k.nearest(P)$ denotes the nearest predicted location to A_k. K is the number of actual nest sites. The units for ADNP is the number of pixels in the experiment. The results of our experiments are shown in Table 2. From the experiments it is clear that PLUMS(A) achieves similar spatial accuracy on test datasets as SAR, while it needs order of magnitude less computational time to learn.

Data set		PLUMS(A)	Probit	Logit	SAR(with Probit)
Learning	spatial accuracy	16.90	47.16	47.19	13.96
Testing	spatial accuracy	19.19	41.43	41.5	19.30
Learning	Run-time(Seconds)	80	10	10	19420 [1]

Table 2: Learning time and spatial accuracies for learning and test data set

[1] 10,000 draws for Gibbs sampling, 1000 burn-outs

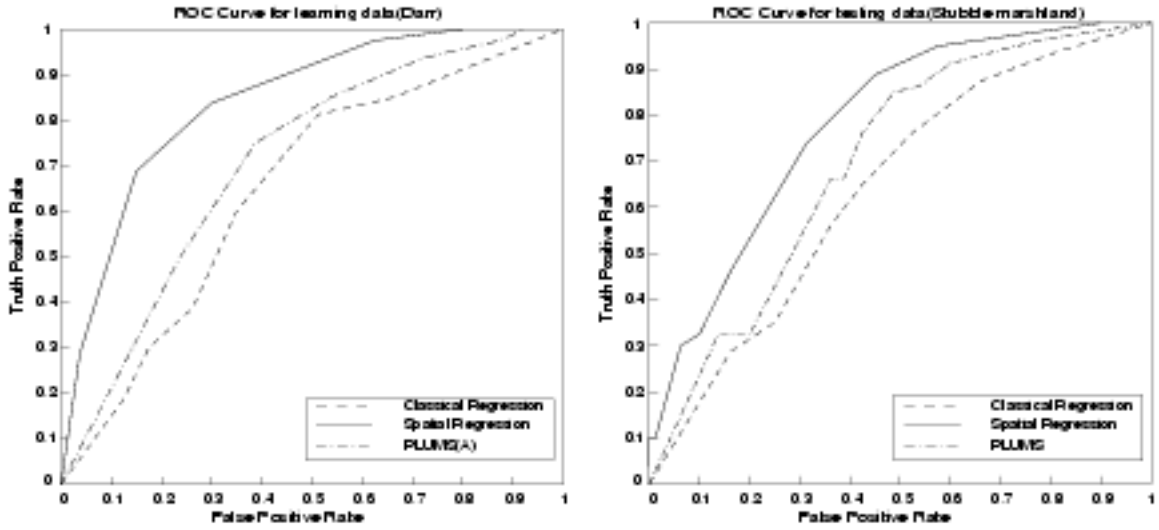
The run-time for learning location prediction models for 3 methods are shown in Table 2. We note that SAR takes two orders of magnitude more computation time relative to PLUMS(A) using the public domain code [38] despite the sparse matrix techniques [45] used in the code.

3.2.2 Comparison of classification accuracy

Classification accuracy achieved by classical regression, SAR(both modified with Probit) and PLUMS(A) are compared on predicting the location of the nests. We use the Receiver Operating Characteristic(ROC) [17] curves to compare classification accuracy. ROC curves plot the relationship between the true positive rate(TPR) and the false positive rate(FPR).

For each cut-off probability b, $TPR(b)$ measures the ratio of the number of sites where the nest is actually located and was predicted divided by the number of actual nest sites. The FPR measures the ratio of the number of sites where the nest was absent but predicted divided by the number of sites where the nests were absent. The ROC curve is the locus of the pair $(TPR(b), FPR(b))$ for each cut-off probability. The higher the curve above the straight line $TPR = FPR$ the better the accuracy of the model.

Figures 8(a) is the ROC curves for the model built using the Darr learning data and Figure 8(b) is the ROC curve for the Stubble test data. It is clear that by using the SAR resulted in better predictions at all cut-off probabilities relative to PLUMS(A), a simple and naive implementation of PLUMS. Alternative smarter implementations of PLUMS enumerated in Figure 7 need to be explored to close the gap.



(a) ROC curves for Learning Data

(b) ROC curves for Test Data

Figure 8: (a) Comparison of PLUMS(A) with other methods on the Darr learning data. (b) Comparison of the models on the test data.

3.3 Interpretation of Preliminary Results

Our preliminary work with classical data mining methods (e.g. logistic regression) and spatial statistical methods (e.g. SAR) for location prediction shows the need for new spatial data mining techniques. Classical regression methods are based on the independent identically distribution assumption and classification accuracy measures. They do not capture spatial auto-correlation properties or spatial accuracy goals of location prediction problems. Spatial statistical methods are based on spatial autocorrelation, but are suited for small spatial datasets with small contiguity matrices, which are much larger than the original spatial

datasets. We have preliminary results on appropriateness of PLUMS, a map similarity driven spatial data mining techniques for location prediction problems. PLUMS(A), a simple implementation of PLUMS, is promising for a large scale location prediction problems where quality of solution is measured in terms of spatial accuracy. Our experiments with location prediction on a real-world spatial dataset (e.g. red-winged blackbird habitats in Darr and Stubble marshes) indicate that PLUMS is likely to outperform existing spatial data mining methods in terms of spatial accuracy for location prediction using orders of magnitude less computational resources.

The intriguing questions are the following. Can other implementations of PLUMS achieve similar if not better classification accuracy as well as spatial accuracy relative to spatial statistical methods(e.g. SAR) using an order of less computational resources? Can PLUMS scale to extremely large Geo-spatial data-sets containing Terabytes of satellite image and aerial survey data for larger geographic areas with finer spatial resolution of measurements. Many of the proposed tasks addresses the first question by exploring other design choices(e.g. 1,2,3,4,5 in figure 7) for PLUMS towards improving the spatial as well as classification accuracy. Other tasks relate to performance tuning towards scaling up to terabyte data sets. We propose to develop scalable algorithms for improving the spatial accuracy of PLUMS derived models, as well as implement and evaluate them in this proposal.

In addition, we have been working on developing effective and scalable data mining techniques [53, 56] spatial data models [55], spatial indexing [61], and spatial query processing strategies [55, 56] for different application domains, including ecology [6] transportation [55, 59] and terrain visualization [62]. Our team is capable of addressing the proposed research issues.

4 Proposed Work

Our preliminary work has shown that simple implementation of PLUMS achieves comparable spatial accuracy as orders of magnitude more expensive spatial statistical methods. The goals of proposed work is to redesign PLUMS to achieve comparable classification accuracy as well and scale up to larger geo-spatial datasets. For achieving these goals, we propose the following tasks: (i) New measures of map similarity, (ii) Efficient search algorithm, (iii) I/O performance tuning, (iv) Formalization, and (v) Experimental Evaluation.

4.1 New Measures for map similarity

Map similarity measure is a combination of classification accuracy and spatial accuracy. Driving the greedy learning algorithm in PLUMS by a map similarity measure(with $0 < \alpha < 1$) rather than a pure spatial similarity measure is likely to improve the classification accuracy achieved by PLUMS. It may require careful redesign of spatial accuracy measure to minimize any conflict with classification accuracy. The comparison of classical regression and the SAR model indicate that it is possible for a learning model to improve both spatial accuracy and classification accuracy using spatial auto-correlation probably at the expense of higher computational costs.

A more interesting hypothesis relates to the trade-off between map similarity measures and function families towards modeling spatial autocorrelation. It is hypothesised that a suitable map similarity function can model spatial autocorrelation eliminating the need for including the autocorrelation term in the modeling function. We plan to explore a variety of map similarity functions towards this purpose. The family of spatial accuracy measure explored will include the vector-GIS measures, e.g. nearest neighbor index (NNI)[15] as well as raster-GIS measures, e.g. principal component analysis [51]. We plan to generalize the spatial-statistical measures of spatial autocorrelation to measures of spatial cross correlation towards design of appropriate spatial accuracy measures. We plan to be in close discussion with spatial domain scientist, e.g. Prof. Uygar Ozsemi, to ensure that the spatial accuracy measures under consideration are meaningful to application domains.

Finally, an important issue of interest related to map similarity is the raster-vector conversion. The current PLUMS(A) implementation iterates between the following two distinct steps. First, the algorithm searches a potentially large parameter space in search for a tuple of parameter values which may potentially

maximize the spatial accuracy measure. The parameter value tuple determined at this step instantiates a member of the function family, which defines a probability surface distributed over the spatial domain. The second step involves the conversion of the *raster* probability surface to a *vector* map of nest locations. There are many strategies available to perform a raster to vector conversion. In our preliminary results we take the highest N values of the probability surface, where N is the number of the actual nest locations. The spatial accuracy measure is calculated between the actual and the predicted nest locations. We plan to explore other strategies for converting the raster probability surface into a vector layer of predicted nest locations. For example, consider the probability surface as a collection of pairs $z_i = (s_i, p_i)$ where $s_i = (x_i, y_i)$ are the sites where the variables were sampled and p_i is the probability at site s_i. Define a distance function $d\ d(z_i, z_j) = \sqrt{(\gamma((x_i - x_j)^2 + (y_i - y_j)^2) + \delta((p_i - p_j)^2))}$ where γ and δ are positive parameters. Define a density [25] function f for each site as $f(s, s_i) = \sum_{j=1}^{j=n} \exp^{-\frac{d(s,s_j)}{\sigma^2}}$ where σ is a user-defined parameter to normalize the density function f. The function f captures the zone of influence of each location on the probability surface. We want to choose those s_i's which have large zones of influence.

4.2 Efficient Search Algorithms

Preliminary implementation of PLUMS is based on greedy search of the parameter space. We are exploring other search paradigms to speed up and to find solutions better than local maximas of map similarity. The search algorithm can be speeded up via incremental computation of map similarity functions as well as using non-uniform discretizations of parameter space. PLUMS changes only one parameter at a time in exploring the neighboring parameter tuples. It is possible to estimate map similarity of any neighbor by examining only the maps of corresponding explanatory variable and the dependent variables for many function families. It should also be feasible to estimate the map similarity of all neighbors by scanning all the maps only once. Non-uniform discretization of parameter space can be based on gradient search [20] if map similarity measure is differentiable. The parameter δ can be made sensitive to the rate of change in map similarity function, while guarding the possibility of overshooting the target maxima.

Finally we will explore search algorithms which allow better exploration of parameter space. Multiple trials of greedy search with different starting points is an approach. Simulated Annealing [68] is another approach which the P.I. has explored in previous work [53].

4.3 I/O Performance Tuning: Scaling up to large data sets

Many applications of spatial data mining are characterized by large datasets, too large to fit, all at once, inside typical main memories. We plan to redesign PLUMS to scale upto larger geo-spatial datasets. The I/O bottleneck in PLUMS will arise in the computation of map similarity measures. Appropriate chunking [50] of maps in pages to compute the map similarity in a single scan is the first step. We also propose to explore map compression techniques exploiting spatial autocorrelation to further speed up computation of classification and spatial accuracy with minimal approximation error. We will evaluate Quad-tree[49] and wavelet[41] based spatial compression schemes before exploring new schemes based on spatial autocorrelation.

4.4 Formalization

We have proposed a new objective of map similarity, which we believe captures the underlying semantics of the location prediction problem. We plan to evaluate the proposed objective from a theoretical viewpoint based on statistical learning theory. This will help us identify the limitations of the map similarity measure and provide a data independent perspective. Of particular interest are the non-parametric statistical issues like regularization [7] of the objective function and generalization error estimation during learning [8, 69].

4.5 Experimental Evaluation

We have identified a a benchmark spatial dataset(Figure 1 and Figure 2) as well as a design space of PLUMS for improving classification and spatial accuracy. This design space is shown in Figure 7. The first component is related to the map similarity measure. In preliminary results we have examined classification accuracy($\alpha = 0$) and spatial accuracy($\alpha = 1$). We plan to explore map similarity ($0 < \alpha < 1$) by comparing PLUMS implementations **A** and **4** in Figure 7. The second component of PLUMS is the family of functions for the learning model. We plan to experiment with the family of generalized linear function(with spatial autocorrelation term) via comparison of PLUMS implementations **A,2** and **3**. The third component is the discretization of the parameter space. Our preliminary work has focused on a uniform-grid approach and we plan to evaluate the adaptive non-uniform discretizations via comparisons of implementations **4** and **5**. Finally we plan to evaluate search algorithms which can go beyond local maxima. This component will be explored on the basis of implementation **A** and **1**.

An orthogonal set of experiments will evaluate I/O scalability of PLUMS resulting from map chunking, map compression and incremental computations of map similarity. For this purpose we will seek a large geo-spatial dataset which cannot reside in main memory all at once. Candidate datasets include those from crime hot-spot analysis [14]. University of Minnesota Center for Urban and Regional Affairs as well as the Minneapolis police department have such data sets.

5 Research Plan and Schedule, Available Resources, Technology Transfer and Diversity

The proposed research will be accomplished according to the following schedule.

- *New measure for map similarity:* Design of a set of measures for map similarity will take 6 months. We will exploit the techniques from Remote Sensing for raster maps and the techniques from Geographic Information Systems for vector maps.
- *Efficient Search Algorithms:* This task will take 6 months. Greedy-search and simulated annealing algorithms have already been implemented using MATLAB and evaluated for small (Megabyte) spatial data sets. Computation of map similarity measures appears to be the computational bottleneck. We plan to explore incremental algorithms for compute map similarity of neighbors in parameter space. We also plan to re-implement our algorithms in C and use profiling techniques to identify and remove computational bottlenecks.
- *I/O Performance Tuning:* This task will take 6 months. Minimizing Disk access will be critical for efficiently computing map similarity for large maps (Gigabytes). Incremental computation of map similarity can reduce disk I/O by eliminating the need to see maps for all explanatory variables but one. Additional savings may come from techniques like appropriate map chunking, map compression, multi-resolution representation etc.
- *Formalization of the new proposed measure:* It is hard to estimate the time for formalization. We will work with colleagues in statistical learning theory to formalize the statistical assumption behind PLUMS to characterize the nature of spatial application which can best benefit from our proposed work.
- *Experimental Evaluation:* This task will take 6 months. We plan to identify competing methods and heuristics for location prediction from Geo-statistics (e.g. Kriging for spatial interpolation), Remote Sensing (e.g. contextual classification using Markov Random Fields) and Spatial Econometrics (e.g. spatial autocorrelation regression) and compare them with PLUMS. We plan to use datasets from red-winged blackbird's habitat analysis and urban crime hot spot analysis for comparison. We plan to measure learning-time, as well as accuracy of location prediction for alternative methods.

Available Resources: We have access to state-of-the-art computer resources. We have a NSF infrastructure grant in place currently and are likely to be awarded another one related to data mining. The equipment from these grants can use those to carry out the research. We also have access to resources from the Army High Performance Research Center(AHPCRC) and the Computer Science Department.

Technology Transfer : The results of this research will be disseminated via our web-site (http://www.cs.umn.edu/research/shashi group) and via publications in premier refereed conferences and journals. The technology developed during this project will be evaluated by our collaborators in Ecology before being made available to professionals interested in location prediction. The main audience is a set of researchers and policy makers who would like to predict similar rather than identical maps because of the high granularity of decision making and the prohibitive cost of current spatial statistics techniques and inadequacy of the classical data mining techniques in modeling spatial autocorrelation properties.

Diversity: P.I. has been participating in a Summer Institute Program for undergraduate from historically black colleges and universities held annually at the Army High Performance Research Center at the University of Minnesota. P.I. has mentored a dozen students over last 3 years. P.I. is the faculty in charge of organizing this institute in Summer 2000. P.I. plans to engage students with projects related to geo-spatial data mining in order to expose them to this rapidly developing area.

6 Comparison with Related Work

Related work includes spatial statistics and spatial data mining.

Spatial Statistics: The purposes of spatial statistical models can be divided into three categories: descriptive, explanatory, and predictive. Descriptive models characterize the distribution of the spatial phenomenon. Often description is based on a set of spatial statistics and indices. For example, spatial distributions (e.g. nest locations in a marsh) may be classified into random (see Figure 3(b)) or clustered (see Figure 1(a), 2(a)) using spatial autocorrelation (e.g Moran's I coefficient), nearest neighbor index or quardat analysis [4, 13, 15].

Explanatory models deal with spatial associations, i.e. relationships between a phenomenon and the factors affecting its spatial distribution. For example, in order to explain why red-winged black birds nest in a certain area of a marsh, roles of vegetation durability index, water depth and distance to edge etc. may be examined. More detailed analysis may explore how each factor may influence the nest locations. Example techniques are based on chi-square tests and spatial correlation coefficients using appropriate geographic units.

Predictive models may be used subsequently for prediction or simulation of alternative management strategies. For example, near future nesting locations may be predicted given the current conditions and growth factor of significant factors (e.g vegetation durability index, distance to edge etc.) under certain assumptions. Alternatively, these models may explore what may happen if certain conditions are changed via new management strategies. Example techniques include regression using appropriate geographic units, structural factors (e.g. local features of the geographic unit) as well as spatial factors (e.g. absolute location, distance to certain features and neighborhood effects such as spatial autocorrelation) [10, 11, 16, 22, 27, 29, 65]. Current research in Spatial Econometrics [3, 38], Geo-statistics [30] and Ecological modeling [28] has focused on enlarging the family of functions in order to include functions which capture the unique characteristics inherent in spatial data.

Spatial Data Mining: Spatial data mining [18, 33, 34, 35, 47, 52], a subfield of data mining [1, 19], is concerned with discovery of interesting and useful but implicit knowledge in spatial databases. Common patterns [1] discovered by data mining algorithms include descriptive patterns (e.g. clustering[33]), explanatory patterns (e.g. association rules[35]) and predictive patterns (e.g. classification rules and decision trees). The foundations of data mining algorithms are in statistics and machine learning. One of the goals of data mining algorithms is to scale up to analyze very large datasets which may not fit in the main memory. A

proposed data mining *desiderata* [5] includes the following scalability goals: (i) **One-Scan:** The algorithm should require only one-scan of the database, whenever possible. In the context of location prediction this is a very severe restriction because of the neighborhood influence of spatial entities, (ii) **Anytime Algorithm:** The algorithm is able to produce the "best" results at anytime during the computation, (iii) **Limited RAM:** The algorithm works with limited RAM and buffer allocated by the user, (iv) **Incremental:** The algorithm proceeds in an incremental fashion: In the presence of a new data the algorithm can use previous results without starting the computation afresh. (v) **Forward-only cursor:** The data being processed may be a result of an expensive join algorithm over a potentially distributed data warehouse. Thus the algorithm must operate with a forward-only cursor over a view of the database.

Challenges in Spatial Data Mining arise from following issues. *First*, classical data mining[1, 67] deals with numbers and categories. In contrast, spatial data is more *complex* and includes extended objects such as points, lines, and polygons. Appropriate spatial modeling is required for analyzing and mining spatial dataset. It may involve all feature types (points, lines, and polygons). Choice of geographic unit is a key decision for polygonal features. Polygonal geographic units may be arbitrary (e.g. a grid), based on existing boundaries (e.g. administrative or political) or derived from data distribution (e.g. areas homogeneous with respect to significant factors).

Second, classical data mining works with explicit inputs, whereas spatial predicates (e.g. overlap) and attributes (e.g. distance, spatial auto-correlation) are often *implicit*. In the presence of spatial data the standard approach in the data mining community is to materialize spatial relationships as attributes and rebuild the model with the "new" spatial attributes [35, 34]. One needs to decide *apriori* which relationships to materialize.

Third, classical data mining treats each input to be independent of other inputs, whereas spatial patterns often exhibit continuity and *high autocorrelation among nearby features*. For example, population density of nearby locations are often related. Researchers in spatial econometrics [3] and regional economics [23, 37] have developed techniques to incorporate spatial autocorrelation information [10, 11, 16, 21, 26, 28, 64] into regression models using a contiguity matrix. These techniques are computationally expensive and require large memory resources as the size of contiguity matrix is much larger than the size of the spatial datasets.

Fourth, the measure of evaluation, e.g. spatial accuracy of predicted location , maybe substantially different from classical measures, e.g. classification accuracy measured as total square error over pixels. For example, the average distance between predicted nests and actual nests may be an appropriate measure for the predicting locations of red-winged black birds. There are few well-known spatial data mining algorithms which are driven by such spatial measures. Techniques from optimization and search algorithm literature may be exploited to address this issue as shown in this proposal.

References

[1] R. Agrawal. Tutorial on database mining. In *Thirteenth ACM Symposium on Principles of Databases Systems*, pages 75–76, Minneapolis, MN, 1994.

[2] P.S. Albert and L.M. McShane. A generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data. *Biometrics (Publisher: Washington, Biometric Society, etc.)*, 51:627–638, 1995.

[3] L Anselin. *Spatial Econometrics: methods and models*. Kluwer, Dordrecht, Netherlands, 1988.

[4] J.E. Besag. Spatial Interaction and Statistical Analysis of Latice Systems. *Journal of Royal Statistical Society, Ser. B (Publisher: Blackwell Publishers)*, 36:192–236, 1974.

[5] P. Bradley, U. Fayyad, and C.Reina. Scaling em (expectation-maximization) clustering to large databases. Technical Report MSR-TR-98-35, Microsoft Research, November 1998.

[6] S. Chawla, S. Shekhar, W-L Wu, and U. Ozesmi. Modeling spatial dependencies for mining geospatial data: An introduction. In *Geographic data mining and Knowledge Discovery(GKD) (Ed. Harvey Miller and Jiawei Han), under contract with Taylor and Francis, URL: http://www.geog.utah.edu/ ~hmiller/gkd_text*.

[7] Vladimir Cherkassky and Filip Mulier. *(Chapter 3)Learning From Data Concepts, Theory, and Methods*. John Wiley & SONS Inc., 1998.

[8] Vladimir Cherkassky and Filip Mulier. *(Chapter 4) Learning From Data Concepts, Theory, and Methods*. John Wiley & SONS Inc., 1998.

[9] Vladimir Cherkassky and Filip Mulier. *Learning From Data Concepts, Theory, and Methods*. John Wiley & SONS Inc., 1998.

[10] A. Cliff and J. Ord. *Spatial Autocorrelation*. London Pion Ltd., 1973.

[11] A.D. Cliff and J.K. Ord. *Spatial processes: Models and Applications*. London Pion Ltd., 1981.

[12] P. Could. *The Geographer at Work*. Routledge and Kegan Paul, London, 1985.

[13] N.A. Cressie. *Statistics for Spatial Data (Revised Edition)*. Wiley, New York, 1993.

[14] Dan Sadler. Exploring Crime Mapping. National Inst. of Justice Crime Mapping Research Center web-site http://www.ojp.usdoj.gov/cmrc/briefingbook/welcome.html.

[15] P.J. Diggle. *Statistical analysis of spatial point patterns*. Academic Press, 1993.

[16] J. Dobson. Spatial logic in paleogeography and the explanation of continental drift. In *Annals, Association od American Geographers*, pages 187–266, 1992.

[17] J.P. Egan. *Signal Detection Theory and ROC analysis*. Academic Press, New York, 1995.

[18] M. Ester, H-P Kriegel, and J. Sander. Knowledge discovery in spatial databases. In *Advances in Artificial Intelligence, 23rd Annual German Conference on Artificial Intelligence*, pages 61–74, Bonn, Germany, September 1999.

[19] U. M. Fayyad. Knowledge discovery in databases: An overview. In *Inductive Logic Programming, 7th International Workshop, ILP-97, Lecture Notes in Computer Scienc*, volume 1297, pages 3–16. Springer, September 1997.

[20] B. Flury. *A First Course in Multivariate Statistics (Section 7.5: Simple Logistic Regression)*. Springer, 1997.

[21] C. Greenman. Turning a map into a cake layer of information. *New York Times*, January 20th (http://www.nytimes.com/library/tech/00/01/circuits/arctiles/20giss.html) 2000.

[22] Griffith and ed. Daniel A. *Spatial Statistics: Past, Present and Future*. Institute of Mathematical Geography, Ann Arbor, MI, 1990.

[23] D. Griffith. Statistical and mathematical sources of regional science theory: Map pattern analysis as an example. *Papers in Regional Science (Publisher: Springer)*, (78):21–45, 1999.

[24] R.H. Guting. An Introduction to Spatial Database Systems. *Very Large Data Bases Journal (Publisher:Springer Verlag)*, October 1994.

[25] A. Hinneburg and D.A. Keim. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. *Proc. Int. Conf. Knowledge Discovery and Data Mining(KDD'98), New York, NY*, pages 58–65, 1998.

[26] M. Hohn and L. Gribko A.E. Liebhold. A Geostatistical model for Forecasting the Spatial Dynamics of Defoliation caused by the Gypsy Moth, Lymantria dispar (Lepidoptera:Lymantriidae). *Environmental Entomology (Publisher: Entomological Society of America)*, 22:1066–1075, 1993.

[27] H.C. Huang, H.-C., and N. Cressie. Statistical analysis of spatial point patterns. in bayesian inference in wavelet based model. *Lecture Notes in Statistics (Publisher: Springer Verlag)*, 141:203–222, 1999.

[28] F. Huffer and H. Wu. Markov chain monte carlo for autologistic regression models with application to the distribution of plant species. *Biometrics (Publisher: Washington, Biometric Society, etc.)*, 54(3):509–535, 1998.

[29] W. Isard. *Location and Space Economy*. MIT Press, Cambridge, MA, 1985.

[30] Issaks, Edward, and Mohan Srivastava. *Applied Geostatistics*. Oxford University Press, Oxford, 1989.

[31] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar. Multilevel hypergraph partitioning: Application in VLSI domain. *Proceedings ACM/IEEE Design Automation Conference*, 1997.

[32] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar. Multilevel hypergraph partitioning: Application in VLSI domain. *IEEE Transactions on Very Large Scale Integration(VLSI) Systems*, 7(1):69–79, March 1999.

[33] E.M. Knorr and R.T. Ng. Extraction of spatial proximity relationships and commonalities in spatial data mining. *Proc. 2rd Int. Conf. Knowledge Discovery and Data Mining(KDD'96),Portland, Oregon*, pages 347–350, 1996.

[34] K. Koperski, J. Adhikary, and J. Han. Spatial data mining: Progress and challenges. In *Workshop on Research Issues on Data Mining and Knowledge Discovery(DMKD'96)*, pages 1–10, Montreal, Canada, 1996.

[35] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Advances in Spatial Databases, Proc. of 4th International Symposium, SSD'95*, pages 47–66, Portland, Maine, USA, 1995.

[36] P. Krugman. *Development, geography, and economic theory*. MIT Press, Cambridge, MA, 1995.

[37] J. LeSage. Regression Analysis of Spatial data. *The Journal of Regional Analysis and Policy (Publisher: Mid-Continent Regional Science Association and UNL College of Business Administration)*, 27(2):83–94, 1997.

[38] J.P. LeSage. Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, (20):113–129, 1997.

[39] D.R. Liu, S. Shekhar, and M. Coyle. An Evaluation of Access Methods for Spatial Networks. In *Proceedings of the Tenth International Conference on Data Engineering, IEEE*, 1994.

[40] D. Mark. Geographical information science: Critical issues in an emerging cross-disciplinary research domain. In *NSF Workshop*, Feburary 1999.

[41] Y. Mayer. *Wavelets: Algorithms and Applications*. SIAM, Philadelphia, 1993.

[42] D.P. McMillen. Probit with spatial autocorrelation. *Journal of Regional Science (Publisher: Springe)*, (32):335–348, 1992.

[43] S. Ozesmi and U. Ozesmi. An Artificial neural network approach to spatial habitat modeling with interspecific interaction. *Ecological Modelling (Publisher: Elsevier Science B. V.)*, (116):15–31, 1999.

[44] U. Ozesmi and W. Mitsch. A spatial habitat model for the Marsh-breeding red-winged black-bird(agelaius phoeniceus l.) In coastal lake Erie wetlands. *Ecological Modelling (Publisher: Elsevier Science B. V.)*, (101):139–152, 1997.

[45] R. Pace and R. Barry. Sparse spatial autoregressions. *Statistics and Probability Letters (Publisher: Elsevier Science)*, (33):291–297, 1997.

[46] R.J.Haining. *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge, U.K., 1989.

[47] John F. Roddick and Myra Spiliopoulou. A bibliography of temporal, spatial and spatio-temporal data mining research. *ACM Special Interest Group on Knowledge Discovery in Data Mining(SIGKDD) Explorations*, 1999.

[48] S. Shekhar and S. Chawla. *Spatial Databases: Issues, Implementation and Trends*. (Under Contract)Prentice Hall, 2000.

[49] H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, 1990.

[50] S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. In *Tenth International Conference on Data Engineering*, pages 328–336. IEEE Computer Society, February 1994.

[51] R. Schowengerdt. *Remote Sensing:Models and Methods for Image Processing*. Academic Press, 1997.

[52] E.C. Shek, R.R. Muntz, E. Mesrobian, and K. Ng. Scalable exploratory data mining of distributed geoscientific data. *Proc. 2rd Int. Conf. Knowledge Discovery and Data Mining(KDD'96),Portland, Oregon*, 1996.

[53] S. Shekhar and B. Amin. Generalization by neural networks. *IEEE Trans. on Knowledge and Data Eng.*, 4(2), 1992.

[54] S. Shekhar, S. Chawla, S. Ravada, A.Fetterer, X.Liu, and C.T. Lu. Spatial databases: Accomplishments and Research Needs. *IEEE Transactions on Knowledge and Data Engineering*, Jan-Feb 1999.

[55] S. Shekhar, Andrew Fetterer, and Brajesh Goyal. Materialization trade-offs in hierarchical shortest path algorithms. In *Proc. Symposium on Large Spatial Database(Springer Verlag)*, 1997.

[56] S. Shekhar and B. Hamidzadeh. Learning transformations rules for semantic query optimization: A data-driven approach. *IEEE Trans. on Knowledge and Data Eng. (Special Issue on Discovery in Databases)*, 5(6), 1993.

[57] S. Shekhar and D. R. Liu. Connectivity-clustered access methods for networks and network computations: A summary of results. In *Proc. Intl. Conf. on Data Engineering, IEEE*, 1995.

[58] S. Shekhar and D. R. Liu. Partitioning similarity graphs: A framework for declustering problems. *An International Journal: Information Systems(Publisher: Pergamon Press)*, 21(6), September 1996.

[59] S. Shekhar and D-R. Liu. Ccam: A connectivity-clustered access method for aggregate queries on transportation networks-a summary of results. *IEEE Trans. on Knowledge and Data Eng.*, 9(1), January 1997.

[60] S. Shekhar and D.R. Liu. A Similarity Graph-Based Approach to Declustering Problem and its Applications Toward Parallelizing Grid Files. In *Proceedings of the Eleventh International Conference on Data Engineering, IEEE*, pages 373–381, March 1995.

[61] S. Shekhar and S. Ravada. Parallelizing the refinement step of spatial join. In *Intl. Conf. on Geographic Info. Systems*. ACM, November 1997.

[62] S. Shekhar, S. Ravada, V. Kumar, D. Chubb, and G. Turner. Declustering and Load-Balancing Methods for Parallelizing Geographic Information Systems. *To appear in IEEE Transaction on Knowledge and Data Engineering. A short version of this paper is published in Proceedings of the 4th International Symposium on Large Spatial Databases (SSD95). Also available as Technical Report TR 95-076, Department of Computer Science, University of Minnesota. URL:http://www.cs.umn.edu/Research/shashi-group/paper_ps/rqtkde96.ps.*

[63] S. Shekhar, S. Ravada, V. Kumar, D. Chubb, and G. Turner. Parallelizing a gis on a shared address space architecture. *IEEE Computer (Special Issue on Multiprocessors)*, 29(12), December 1996.

[64] S. Shekhar, T. A. Yang, and P. Hancock. An intelligent vehicle highway information management system. *Intl Jr. on Microcomputers in Civil Engineering (Publisher: Blackwell Publishers)*, 8(3), 1993.

[65] M. Tiefelsdorf and B. Boots. The exact distribution of Moran's I. *Environment and Planning(Publisher: London Pion Ltd.)*, 27:985–999, 1995.

[66] W.R. Tobler. *Cellular Geography, Philosophy in Geography*. Gale and Olsson, Eds., Dordrecht, Reidel, 1979.

[67] U.M.Fayyad, J.G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In *Advances in knowledge Discovery and Data Mining*, pages 1–34, Menlo Park, 1996. AAAI Press.

[68] P.J.M van Laarhoven. *Simulated Annealing:Theory and Applications*. Wiley, New York, 1987.

[69] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1997.

[70] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer, New York, 1997.

[71] M.F. Worboys. *GIS: A Computing Perspective*. Taylor and Francis, 1995.

[72] Y. Yasui and S.R. Lele. A Regression Method for Spatial Disease Rates: An Estimating Function Approach. *Journal of the American Statistical Association*, 94:21–32, 1997.

7 Biographical Sketch

Shashi Shekhar's current research interests include spatial databases, geographic information systems (GIS) and intelligent transportation systems. He has published over 100 research papers in refereed journals, conferences, workshops and edited books. He is currently a editorial board member of the IEEE Transactions on Knowledge and Data Engineering, and of the IEEE-Computer-Society Computer Science and Engineering Practice Board. He served as the Program co-chair of the ACM Intl. Workshop on Advances in GIS, 1996, and a vice-chair for IEEE Intl. Conf. on Tools with AI, 1995. He is completing a book on "Spatial Databases" and has served as a research advisor on this topic to the Environmental Systems Research Inst. (the dominant company in GIS software market) as well to the United Nations. He has given 30 invited talks at various research institutions. Shekhar's accomplishments includes databases for managing spatial graphs (e.g. road-maps), parallelization of GIS, routing algorithms for Advanced Traveler Information Systems, and archival of traffic measurements. His group has developed, CCAM, one of the most efficient clustering and indexing methods for large road maps as well as algorithms for path evaluation as well as for computing shortest paths. His sponsors include the National Science Foundation, NASA, Army Research Laboratories, Control Data Inc., US DOT, FHWA, MN/DoT and the ITS Institute. He is a senior member of IEEE, and a member of ACM. More details are available on http://www.cs.umn.edu/~shekhar.

Contact Information:
Address: 4-192, EE/CS Bldg., 200 Union St. SE, Minneapolis, MN 55455.
Phone: (612) 624-8307 :: Fax: (612) 625-0572
Electronic: shekhar@cs.umn.edu, http://www.cs.umn.edu/~shekhar

Professional Preparation

1985	B.S.,	Computer Science,	Indian Inst. of Tech. (IIT), Kanpur (India).
1987	M.S.,	Computer Science,	University of California, Berkeley.
1989	M.S.,	Business Administration,	University of California, Berkeley.
1989	Ph.D.,	Computer Science,	University of California, Berkeley.

Appointments

1995–	Associate Professor, University of Minnesota, Minneapolis, Minnesota.
1989–95	Assistant Professor, University of Minnesota, Minneapolis, Minnesota

Five Publications Related to the Research Project

1. Modeling Spatial Dependencies for Mining Geo-spatial Datasets: An Introduction, (w/ S. Chawla et al.), in *Geographic data mining and knowledge discovery (Ed. Harvey Miller and Jiawei Han), under contract with Taylor and Francis, URL: http://www.geog.utah.edu/~hmiller/gkd_text*. A summary appears in NSF specialist meeting on Discovering geographic knowledge in data-rich environments (3/1999), http://www.spatial.maine.edu/~max/varenius/KDreport.pdf
2. Spatial Databases: Accomplishments and Research Needs, (w/ S. Chawla et al.), IEEE Trans. on Knowledge and Data Eng., vol. 11, no. 1, Jan.-Feb. 1999.
3. A Scalable, Highly Parallel Formulation of the Back-propagation Algorithm for Hypercubes and Related Architectures, (w/ V. Kumar et al.), IEEE Trans. on Parallel and Distr. Systems, October 1994.
4. Learning Transformation Rules for Semantic Query Optimization: A Data-Driven Approach, (with B. Hamidzadeh), IEEE Transactions Knowledge and Data Eng. (Special Issue on Discovery in Databases), Oct. 1993.
5. Generalization by Neural Networks, (with B. Amin), IEEE Transactions on Knowledge and Data Eng., special issue on self-organizing data and knowledge representations, Volume 4, Number 2, April 1992.

Five other Recent Publications

1. Multilevel Hypergraph Partitioning: Applications in VLSI Domain, (with George Karypis, Rajat Aggarwal, and Vipin Kumar), IEEE Transactions on VLSI Systems, Vol. 7, No. 1, March 1999.

2. Declustering and Load-Balancing Methods for Parallelizing Geographic Information Systems, (with S. Ravada, V. Kumar, D. Chubb, G. Turner), IEEE Transactions in Knowledge and Data Eng., vol. 10, no. 4, July-Aug. 1998 (A summary of results appeared in Intl Symp. on Large Spatial Databases 1995, Springer Verlag LNCS 951).

3. Experiences with Data Models in Geographic Information Systems, (with D. Liu, et al.), Communications of the ACM, April 1997, Vol. 40, No. 4.

4. CCAM: A Connectivity-Clustered Access Method for Networks and Network Computations, (with D. Liu), IEEE Transactions on Knowledge and Data Eng., January 1997, Vol. 9, No. 1 (A summary of results appeared in IEEE Intl. Conf. on Data Eng. 1995.)

5. Parallelizing a GIS on a Shared Address Space Architecture, (with S. Ravada, V. Kumar, D. Chubb and G. Turner), IEEE Computer, (Special issue on Shared Memory Multiprocessors), Volume 29, No. 12, December 1996.

Synergistic Activities

- Completing a textbook on Spatial Databases under contract with Pretice Hall (1999-2000) based on new courses on Spatial Databases and Spatial Data Mining developed at University of Minnesota.
- Active participation in broadening the participation of groups underrepresented in science via supervising over two dozen undergraduate students from historically black colleges in Army High Performance Computing Research Center annual summer workshops (1997-present), NSF Research Experience for Undergraduates (1999) and Undergraduate Research Opportunity Programs (1991-present). Server as Director, Undergraduate Studies (1995-97) and chair of departmental curriculum committee (1999-present).
- Editorial Board, IEEE Transactions on Knowledge and Data Eng. (1996-present), IEEE-CS Science and Eng. Pratices Publication Board (1995-97).
- Program Chair, ACM Intl. Conf. on Geographic Info. Systems (1996). Co-chair or vice-chair for many conferences including IEEE Intl. Conf. on Tool with AI (1995), Mini-track on Neural Networks in HICCS (1996), AAAI Workshop on Integrating Symbolic AI with Neural Networks (1992).
- Technical Advisor to United Nations Development Program (1997-98), Environmental Systems Research Inst. (Redlands, CA, 1995), Minnesota Dept. of Transportation Strategic Research Initiatives (1993-94).

Collaborators and Other Affiliations

- In past 48 months I have collaborated with Prof. V. Kumar, Prof. J. Srivastava, Prof. T. Burke, Prof. M. Bauer, Prof. M. Donath, Prof. A. Tripathi, Prof. G. Karypis, Prof. Prof. J. Nieber, Prof. A. Tewfik (all University of Minnesota); Dr. Radhakrishnan, Dr. Phil Emmerman, Dr. Raju Numburu, Doug Chubb, G. Turner, J. Gurny (all Army Research Lab.); Marthand Nookala, Jim Wright, Jim Aswegan (all MNDOT), Prof. M. Egenhofer (Univ. of Maine), and Dr. Pradeep Sinha (Intertech Systems).
- My thesis advisor were Prof. C. V. Ramamoorthy and Prof. L. A. Zadeh (all University of California, Berkeley).
- I supervised Ph.D. thesis of Prof. T. A. Yang (U. Conneticut), Prof. B. Hamidzadeh (U. British Columbia), Prof. Duen Ren Liu (Taiwan), Dr. Mark Coyle (Siebel), and Dr. Siva Ravada (Oracle). I supervised post-doctoral works of Dr. S. Chawla. Following individual visited my research laboratory for 3-weeks to a year: Prof. B. Y. Hwang (Korea), Prof. H. Diwakar (Pune U., India), Dr. F. Polat (Bilkent U., Turkey), Prof. I. Singh (India).

8 Budget Justification

The requested budget includes two years of support for one graduate student and 1 month (per year) of the time of Shashi Shekhar. It includes support for supplies and travel to disseminate the research results at a premier refereed conference. Finally, it includes small budgets for material and supplies include photocopying, printing, as well as services and other provisions which are normally not covered by indirect overhead charges yet are essential to facilitate the research activities.

Need for additional equipment is minimal, since P.I. has access to an excellent computing environment. Last year, P.I. was part of a group, which was awarded an infra-structure grant for a proposal titled *Research in Networked Information Systems* by NSF to acquire special equipment such as high-performance workstations, hang-bandwidth networks and high-capacity storage (disk-arrays). In addition, P.I. was part of a group which submitted an infra-structure proposal titled *Cluster Computing for Knowledge Discovery in Diverse Data Sets* for acquiring a parallel computer to support knowledge discovery tasks including spatial data mining. Informally, we have learned that this proposal is likely to be funded. Thus P.I. has access to computing infra-structure, which is likely to be adequate for the proposed research. This will facilitate experimentation with spatial datasets of large sizes. It will also provide facilities for visualizing the spatial patterns and the results of the spatial data mining techniques for location prediction.

9 Current and Pending Support (March 2000)

9.1 Pending Support for Shashi Shekhar

- Co-P.I., Cluster Computing for Knowledge Discovery in Diverse Data Sets, $74,516, submitted to National Science Foundation.
- Co-P.I., Integrative Education and Training in Genomic Science and Engineering, $2,593,078 submitted to National Science Foundation.
- Co-P.I., Cluster computing for knowledge discovery in Diverse Datasets. Submitted to National Science Foundation. Approx. $ 76,000, for infrastructure.

9.2 Current Support for Shashi Shekhar

- P.I., High Performance Spatial Visualization of Traffic Data, $100,000, Federal Highway Authority Inst. for Intelligent Transportation Systems, 1/2000 - 12/2000.
- P.I., Spatial Databases and Spatial Data Mining $84,000, Army Research Lab. / AHPCRC, 1/2000 -1/2001.
- Co-P.I, Research in Networked Information Systems, $97,000 for infrastructure, National Science Foundation (NSF), with Prof. A. Tripathi et al., Jan. 1999 - Dec. 2001.
- Co-P.I., Institutionalizing MTPE Data for Land and Environment Management, $1,334,552 National Aeronautics and Space Agency, with Prof. T. Burk et al., 9/1997- 8/2001.
- Co-PI, A New Approach to Assessing Road User Charges, $770,000, Federal Highway Administration and the State DOTs of California, Illinois, Indiana, Iowa, Michigan, Minnesota, Nevada, S. Dakota, Texas, and Wisconsin, with Prof. M. Donath et al., July 1999 - December 2001.

FACILIITES and EQUIPMENT

Need for additional equipment is minimal, since P.I. has access to an excellent computing environment. Last year, P.I. was part of a group, which was awarded an infra-structure grant for a proposal titled *Research in Networked Information Systems* by NSF to acquire special equipment such as high-performance workstations, hang-bandwidth networks and high-capacity storage (disk-arrays). In addition, P.I. was part of a group which submitted an infra-structure proposal titled *Cluster Computing for Knowledge Discovery in Diverse Data Sets* for acquiring a parallel computer to support knowledge discovery tasks including spatial data mining. Informally, we have learned that this proposal is likely to be funded. Thus P.I. has access to computing infra-structure, which is likely to be adequate for the proposed research. This will facilitate experimentation with large spatial datasets. It will also provide facilities for visualizing the spatial patterns and the results of the spatial data mining techniques for location prediction.

The computer science department at the University of Minnesota has the following computer equipment installed:

- 1 - Sun SPARCserver 1000, 6 CPUs, 64 MB RAM. (Caesar)
- 1 - Sun SPARCserver 1000e, 4 CPUs, 256 MB RAM (augustus)
- 1 - SGI Challenge M (yocto)
- 1 - Ancor FC-266 Fibrechannel switch.
- 2 - Sun SPARCsever 1000's, each with 4 CPUs, 128 MB RAM, 24 GB Sun disk arrays, Ciprico 8 GB disk arrays, and 8 networks each including an ATM and Fibrechannel link. (Dilbert and Dogber
- (Dilbert and Dogbert) 1 - SGI Challenge S with 32 Meg RAM, and three networks with one ATM link. Print server and ITLabs mail hub. (gregaran)
- 1 - Sun IPX's serving as NIS (neptune)
- 1 - Sun IPX serving as a license server. (license)
- 1 - NeXT cube serving as Webster Server. (nakamichi)
- SGI Challenge S, 150 MHz R4400, 64 MB RAM, 4 GB disk space
- SGI Challenge M, 150 MHz R4400, 80 MB RAM, 13 GB disk space