

# Modeling Spatial Dependencies for Mining Geospatial Data\*

*Sanjay Chawla<sup>†</sup>, Shashi Shekhar<sup>‡</sup>, Weili Wu<sup>§</sup>,  
and Uygur Ozesmi<sup>¶</sup>*

## 1 Introduction

Widespread use of spatial databases[24] is leading to an increasing interest in mining interesting and useful but implicit spatial patterns[14, 17, 10, 22]. Efficient tools for extracting information from geo-spatial data, the focus of this work, are crucial to organizations which make decisions based on large spatial data sets. These organizations are spread across many domains including ecology and environment management, public safety, transportation, public health, business, travel and tourism[2, 12].

Classical data mining algorithms[1] often make assumptions (e.g. independent, identical distributions) which violate Tobler's first law of Geography: everything is related to everything else but nearby things are more related than distant things[25]. 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[6]. Knowledge discovery techniques which ignore spatial autocorrelation typically perform poorly in the

---

\*Support in part by the Army High Performance Computing Research Center under the auspices of Department of the Army, Army Research Laboratory Cooperative agreement number DAAH04-95-2-0003/contract number DAAH04-95-C-0008, and by the National Science Foundation under grant 963 1539.

<sup>†</sup>Vignette Corporation, Waltham MA 02451. Email:schawla@vignette.com

<sup>‡</sup>Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, USA. Email: shekhar@cs.umn.edu

<sup>§</sup>Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, USA. Email: wuw@cs.umn.edu

<sup>¶</sup>Department of Environmental Sciences, Ericyes University, Kayseri, Turkey. Email:uozesmi@erciyes.edu.tr

presence of spatial data. Spatial statistics techniques, on the other hand, do take spatial autocorrelation directly into account[3], but the resulting models are computationally expensive and are solved via complex numerical solvers or sampling based Markov Chain Monte Carlo (MCMC) methods[15].

In this paper we first review spatial statistical methods which explicitly model spatial autocorrelation and we propose PLUMS (Predicting Locations Using Map Similarity), a new approach for supervised spatial data mining problems. PLUMS searches the parameter space of models using a map-similarity measure which is more appropriate in the context of spatial data. We will show that compared to state-of-the-art spatial statistics approaches, PLUMS achieves comparable accuracy but at a fraction of the cost (two orders of magnitude). Furthermore, PLUMS provides a general framework for specializing other data mining techniques for mining spatial data.

## 1.1 Unique features of spatial data mining

The difference between classical and spatial data mining parallels the difference between classical and spatial statistics. One of the fundamental assumptions that guides statistical analysis is that the data samples are independently generated: they are like the successive tosses of a coin, or the rolling of a die. When it comes to the analysis of spatial data, the assumption about the independence of samples is generally false. In fact, spatial data tends to be highly self correlated. For example, people with similar characteristics, occupation and background tend to cluster together in the same neighborhoods. The economies of a region tend to be similar. Changes in natural resources, wildlife, and temperature vary gradually over space. In fact, this property of like things to cluster in space is so fundamental that, as noted earlier, geographers have elevated it to the status of the first law of geography. This property of self correlation is called spatial autocorrelation. Another distinct property of spatial data, *spatial heterogeneity*, implies that the variation in spatial data is a function of its location. Spatial heterogeneity is measured via local measures of spatial autocorrelation[18]. We discuss measures of spatial autocorrelation in Section 2.

## 1.2 Famous Historical Examples of Spatial Data Exploration

Spatial data mining is a process of automating the search for potentially useful patterns. We now list three historical examples of spatial patterns which have had a profound effect on society and scientific discourse[11].

1. In 1855, when the Asiatic cholera was sweeping through London, an epidemiologist marked all locations on a map where the disease had struck and discovered that the locations formed a cluster whose centroid turned out to be a water-pump. When the government authorities turned-off the water pump, the cholera began to subside. Later scientists confirmed the water-borne nature of the disease.
2. The theory of Gondwanaland, which says that all the continents once formed

one land mass, was postulated after R. Lenz discovered (using maps) that all the continents could be fitted together into one piece—like one giant jigsaw puzzle. Later fossil studies provided additional evidence supporting the hypothesis.

3. In 1909 a group of dentists discovered that the residents of Colorado Springs had unusually healthy teeth, and they attributed this to high levels of natural flouride in the local drinking water supply. Researchers later confirmed the positive role of flouride in controlling tooth-decay. Now all municipalities in the United States ensure that drinking water supplies are fortified with flouride.

In each of these three instances, spatial data exploration resulted in a set of unexpected hypotheses (or patterns) which were later validated by specialists and experts. The goal of spatial data mining is to *automate* the discoveries of such patterns which can then be examined by domain experts for validation. Validation is usually accomplished by a combination of domain expertise and conventional statistical techniques.

### 1.3 An Illustrative Application Domain

We now introduce an example which will be used throughout this paper to illustrate the different concepts in spatial data mining. We are given data about two wetlands, named Darr and Stubble, on the shores of Lake Erie in Ohio USA in order to *predict* the spatial distribution of a marsh-breeding bird, the red-winged blackbird (*Agelaius phoeniceus*). The data was collected from April to June in two successive years, 1995 and 1996.

A uniform grid was imposed on the two wetlands and different types of measurements were recorded at each cell or pixel. In total, values of seven attributes were recorded at each cell. Of course domain knowledge is crucial in deciding which attributes are important and which are not. For example, *Vegetation Durability* was chosen over *Vegetation Species* because specialized knowledge about the bird-nesting habits of the red-winged blackbird suggested that the choice of nest location is more dependent on plant structure and plant resistance to wind and wave action than on the plant species.

Our goal is to build a model for predicting the location of bird nests in the wetlands. Typically the model is built using a portion of the data, called the **Learning** or **Training** data, and then tested on the remainder of the data, called the **Testing** data. For example, later on we will build a model using the 1995 data on the Darr wetland and then test it on either the 1996 Darr or 1995 Stubble wetland data. In the learning data, all the attributes are used to build the model and in the training data, one value is *hidden*, in our case the location of the nests, and using knowledge gained from the 1995 Darr data and the value of the independent attributes in the test data, we want to predict the location of the nests in Darr 1996 or in Stubble 1995.

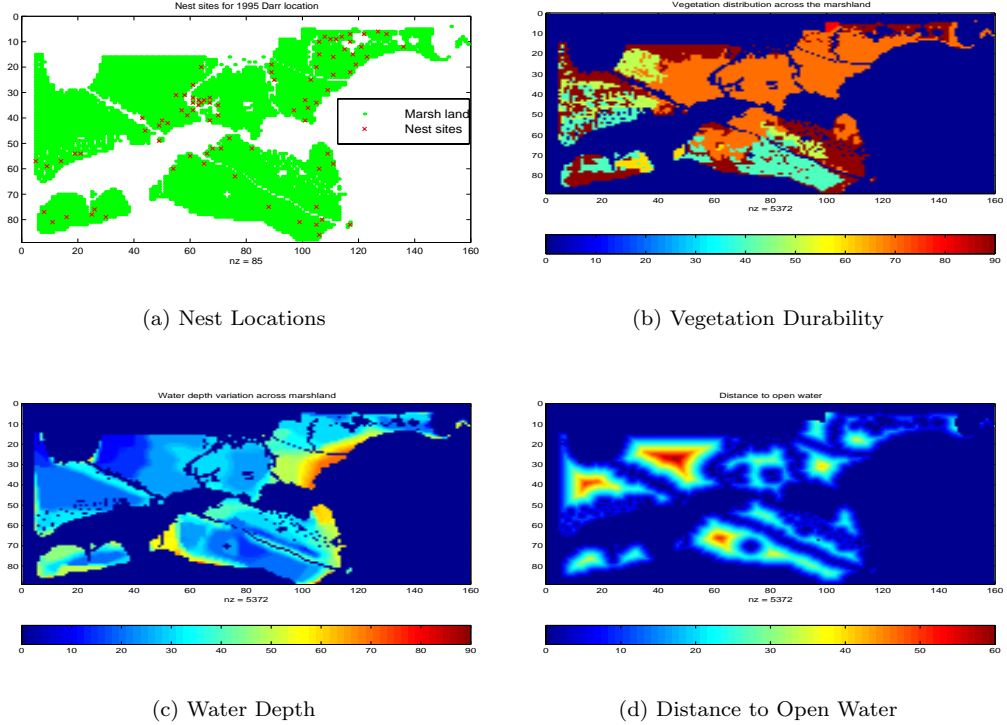
In this paper we focus on three independent attributes, namely *Vegetation Durability*, *Distance to Open Water*, and *Water Depth*. The significance of these

three variables was established using classical statistical analysis. The spatial distribution of these variables and the actual nest locations for the Darr wetland in 1995 are shown in Figure 1. These maps illustrate two important properties inherent in spatial data.

1. The value of attributes which are referenced by spatial location tend to vary gradually over space. While this may seem obvious, classical data mining techniques, either explicitly or implicitly, assume that the data is *independently* generated. For example, the maps in Figure 2 show the spatial distribution of attributes if they were independently generated. One of the authors has applied classical data mining techniques like logistic regression[20] and neural networks[19] 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. The study concluded that with the use of logistic regression, the nests could be classified at a rate 24% better than random[19].
  
2. The spatial distributions of attributes sometimes have distinct local trends which contradict the global trends. This is seen most vividly in Figure 1(b), where the spatial distribution of *Vegetation Durability* is jagged in the western section of the wetland as compared to the overall impression of uniformity across the wetland. This property is called spatial heterogeneity. In Section 2.2 we describe two measures which quantify the notion of spatial autocorrelation and spatial heterogeneity.

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. This measure may not be the most suitable in a spatial context. *Spatial accuracy*—how far the predictions are from the actuals—is as important in this application domain due to the effects of discretizations of a continuous wetland into discrete pixels, as shown in Figure 3. Figure 3(a) shows the actual locations of nests and 3(b) shows the pixels with actual nests. Note the loss of information during the discretization of continuous space into pixels. Many nest locations barely fall within the pixels labeled ‘A’ and are quite close to other blank pixels, which represent ‘no-nest’. Now consider two predictions shown in Figure 3(c) and 3(d). Domain scientists prefer prediction 3(d) over 3(c), since predicted nest locations are closer on average to some actual nest locations. The classification accuracy measure cannot distinguish between 3(c) and 3(d), and a measure of spatial accuracy is needed to capture this preference.

A simple and intuitive measure of spatial accuracy is the Average Distance to



**Figure 1.** (a) Learning dataset: The geometry of the wetland 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.

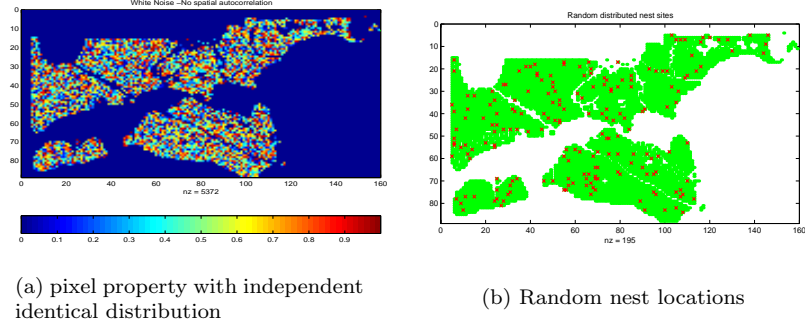
Nearest Prediction (ADNP) from the actual nest sites, which can be defined as

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

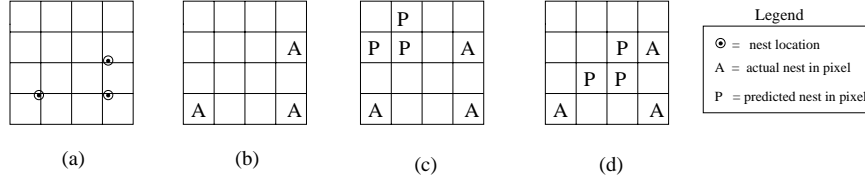
Here  $A_k$  represents the actual nest locations,  $P$  is the map layer of predicted nest locations and  $A_k.nearest(P)$  denotes the nearest predicted location to  $A_k$ .  $K$  is the number of actual nest sites. In Section 3 we will integrate the ADNP measure into the PLUMS framework. We now formalize the spatial data mining problem by incorporating notions of spatial autocorrelation and spatial accuracy in the problem definition.

#### 1.4 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 do-



**Figure 2.** *Spatial distribution satisfying random distribution assumptions of classical regression*



**Figure 3.** (a) *The actual locations of nest, (b) Pixels with actual nests, (c) Location predicted by a model, (d) Location predicted by another model. Prediction (d) is spatially more accurate than (c).*

mains including crime prevention and environmental management. The problem is formally defined as follows:

**Given:**

- A spatial framework  $S$  consisting of sites  $\{s_1, \dots, s_n\}$  for an underlying geographic space  $G$ .
- A collection of explanatory functions  $f_{X_k} : S \rightarrow R^k, k = 1, \dots, K$ .  $R^k$  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  $R^1 \times \dots \times R^K \rightarrow R^Y$ .

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

**Objective:** maximize similarity( $map_{s_i \in S}(\hat{f}^Y(f_{X_1}, \dots, f_{X_K})), map(f_Y(s_i))$ )  
 =  $(1 - \alpha)$  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 <sup>1</sup>.
2. The values of the explanatory functions,  $f_{X_1}, \dots, f_{X_K}$  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 classification" accuracy and a representation dependent "spatial classification" accuracy. The regularization term  $\alpha$  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,  $\alpha$  captures the spatial autocorrelation present in spatial data.

The study of the nesting locations of red-winged black birds [19, 20] is an instance of the location prediction problem. The underlying spatial framework is the collection of  $5m \times 5m$  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. Domain scientists prefer spatially accurate predictions which are closer to actual nests, i.e,  $\alpha > 0$ .

Finally, it is important to note that in spatial statistics the general approach for modeling spatial autocorrelation is to enlarge  $\mathcal{F}$ , the family of learning model functions (see Section 3). The PLUMS approach (See Section 3) allows the flexibility of incorporating spatial autocorrelation in the model, in the objective function or in both. Later on we will show that retaining the classical regression model as  $\mathcal{F}$  but modifying the objective function leads to results which are comparable to those from spatial statistical methods but which incur only a fraction of the computational costs.

## 1.5 Related Work and Our Contributions

Related work examines the area of spatial statistics and spatial data mining.

**Spatial Statistics:** The goal of spatial statistics is to model the special properties of spatial data. The primary distinguishing property of spatial data is that neighboring data samples tend to systematically affect each other. Thus the classical assumption that data samples are generated from independent and identical distributions is not valid. Current research in spatial econometrics, geo-statistics,

---

<sup>1</sup>The entire surface of the Earth cannot be modeled as a Euclidean space but locally the approximation holds true.

and ecological modeling[3, 16, 11] has focused on extending classical statistical techniques in order to capture the unique characteristics inherent in spatial data. In Section 2 we briefly review some basic spatial statistical measures and techniques.

**Spatial Data Mining:** Spatial data mining[9, 13, 14, 22, 4], a subfield of data mining[1], is concerned with the discovery of interesting and useful but implicit knowledge in spatial databases. Challenges in Spatial Data Mining arise from the following issues. First, classical data mining[1] deals with numbers and categories; In contrast, spatial data is more *complex* and includes extended objects such as points, lines, and polygons. Second, classical data mining works with explicit inputs, whereas spatial predicates (e.g., overlap) are often *implicit*. Third, classical data mining treats each input independently of other inputs, whereas spatial patterns often exhibit continuity and *high autocorrelation among nearby features*. For example, the population densities of nearby locations are often related. 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 these “new” spatial attributes[14]. In previous work[4] we studied spatial statistics techniques which explicitly model spatial autocorrelation. In particular we described the spatial autoregression regression (SAR) model which extends linear regression for spatial data. We also compared the linear regression and the SAR model on the bird wetland data set.

**Our contributions:** In this paper, we propose Predicting Locations Using Map Similarity (PLUMS), a new framework for supervised spatial data mining problems. This framework consists of a combination of a statistical model, a map similarity measure along with a search algorithm, and a discretization of the parameter space. We show that the characteristic property of spatial data, namely, spatial autocorrelation, can be incorporated in either the statistical model or the objective function. We also present results of experiments on the “bird-nesting” data to compare our approach with spatial statistical techniques.

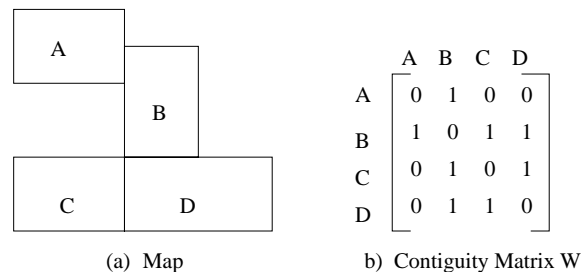
**Outline and scope of Paper:** The rest of the paper is as follows. Section 2 presents a review of spatial statistical techniques including the Spatial Autoregressive Regressive (SAR) model[15], which extends regression modeling for spatial data. In Section 3 we propose PLUMS, a new framework for supervised spatial data mining and compare it with spatial statistical techniques. In this paper we focus exclusively on classification techniques. Section 4 presents results of experiments on the bird nesting data sets and section 5 concludes the whole paper.

## 2 Basic Concepts: Modeling Spatial Dependencies

### 2.1 Spatial Autocorrelation and Examples

Many measures are available for quantifying spatial autocorrelation. Each has strengths and weaknesses. Here we briefly describe the Moran’s I measure.

In most cases, the Moran’s I measure (henceforth MI) ranges between -1 and +1 and thus is similar to the classical measure of correlation. Intuitively, a higher positive value indicates high spatial autocorrelation. This implies that like values tend to cluster together or attract each other. A low negative value indicates that



**Figure 4.** A spatial neighborhood and its contiguity matrix

high and low values are interspersed. Thus like values are de-clustered and tend to repel each other. A value close to zero is an indication that no spatial trend (random distribution) is discernible using the given measure.

All spatial autocorrelation measures are crucially dependent on the choice and design of the contiguity matrix  $W$ . The design of the matrix itself reflects the influence of neighborhood. Two common choices are the four and the eight neighborhood. 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 4. The contiguity matrix of the uneven lattice (left) is shown on the right hand-side. The contiguity matrix plays a pivotal role in the spatial extension of the regression model.

## 2.2 Spatial Autoregression Models: SAR

We now show how spatial dependencies are modeled in the framework of regression analysis. This framework may serve as a template for modeling spatial dependencies in other data mining techniques. In spatial regression, the spatial dependencies of the error term, or, the dependent variable, are directly modeled in the regression equation[3]. Assume that the dependent values  $y'_i$  are related to each other, i.e.,  $y_i = f(y_j) \ i \neq j$ . Then the regression equation can be modified as

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

Here  $W$  is the neighborhood relationship contiguity matrix and  $\rho$  is a parameter that reflects the strength of spatial dependencies between the elements of the dependent variable. After the correction term  $\rho W \mathbf{y}$  is introduced, the components of the residual error vector  $\epsilon$  are then assumed to be generated from independent and identical standard normal distributions.

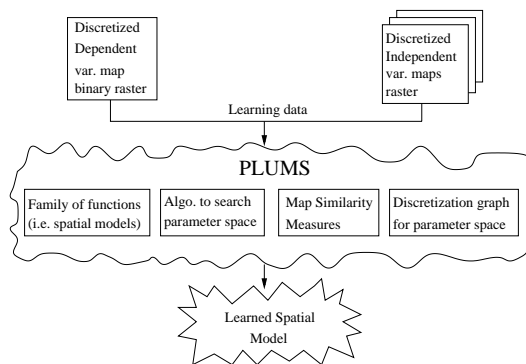
We refer to this equation as the *Spatial Autoregressive Model (SAR)*. Notice that when  $\rho = 0$ , this equation collapses to the classical regression model. The benefits of modeling spatial autocorrelation are many: (1) The residual error will have much lower spatial autocorrelation, i.e., systematic variation. With the

proper choice of  $W$ , the residual error should, at least theoretically, have no systematic variation. (2) If the spatial autocorrelation coefficient is statistically significant, then SAR will quantify the presence of spatial autocorrelation. It will indicate the extent to which variations in the dependent variable ( $y$ ) are explained by the average of neighboring observation values. (3) Finally, the model will have a better fit, i.e., a higher R-squared statistic.

As in the case of classical regression, the SAR equation has to be transformed via the logistic function for binary dependent variables. The estimates of  $\rho$  and  $\beta$  can be derived using maximum likelihood theory or Bayesian statistics. We have carried out preliminary experiments using the spatial econometrics matlab package<sup>2</sup> which implements a Bayesian approach using sampling based Markov Chain Monte Carlo (MCMC) methods[16]. The general approach of MCMC methods is that when the joint-probability distribution is too complicated to be computed analytically, then a sufficiently large number of samples from the conditional probability distributions can be used to estimate the *statistics* of the full joint probability distribution. While this approach is very flexible and the workhorse of Bayesian statistics, it is a computationally expensive process with slow convergence properties. Furthermore, at least for non-statisticians, it is a non-trivial task to decide what “priors” to choose and what analytic expressions to use for the conditional probability distributions.

### 3 Predicting Locations Using Map Similarity (PLUMS)

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 framework for spatial data mining.



**Figure 5.** *The framework for the location prediction process*

<sup>2</sup>We would like to thank James Lesage (<http://www.spatial-econometrics.com/>) for making the matlab toolbox available on the web.

### 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 produce a "learned spatial model" to predict the dependent variable using explanatory variables. PLUMS has four basic components: 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.

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 techniques such as the nearest neighbor index[7], and the principal component analysis of a pair of raster maps.

### 3.2 Greedy Search algorithm of PLUMS

---

#### 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 three explanatory variables is (a,b,c). Its neighborhood may

include the following parameter value tuples:  $(a+\delta, b, c)$ ,  $(a-\delta, b, c)$ ,  $(a, b+\delta, c)$ ,  $(a, b-\delta, c)$ ,  $(a, b, c+\delta)$ , and  $(a, b, c-\delta)$  given a uniform grid with cell-size  $\delta$  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 the neighbors have higher values for the map similarity measure, the neighbor with the highest value of map similarity measure is chosen. This process is repeated until no neighbor has a higher map similarity measure value, i.e., a local maxima has been found. Clearly, this search algorithm can be improved using a variety of ideas including gradient descent[5] and simulated annealing[23]. A simple function family is the family of generalized linear models, e.g., logistic regression[15], 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, 16], geo-statisticians[12] use Co-Kriging and ecologists 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 6. 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 6 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 (ADNP), ...
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*

## 4 Experiment Design and Evaluation

We carried out experiments to compare the classical regression and spatial autoregressive regression (SAR) models[4] and an instance of the PLUMS framework.

**Goals:** The goals of the experiments were (1) to evaluate the effects of including the spatial autoregressive term,  $\rho W \mathbf{y}$ , in the logistic regression model and (2) compare the accuracy and performance of an instance of PLUMS with spatial regression models.

The 1995 Darr wetland data was used as the learning set to build the classical and spatial models. The parameters of the classical logistic and spatial regression model were derived using maximum likelihood estimation and MCMC methods (Gibbs Sampling). The two models were evaluated based on their ability to predict

		Generalized Linear		Generalized Linear with Autocorrelation		Non-Linear with Autocorrelation	
		Search	Greedy (G)	Simulated Annealing(SA)	G	SA	G
Classification accuracy measure ( $\alpha=0$ )	Discretization						
	Non-Uniform(NU) Uniform(U)						
Spatial accuracy measure ( $\alpha=1$ )	U	PLAN A	PLAN (1)	PLAN (2)		PLAN (3)	
	NU						
Map similarity measure ( $0<\alpha<1$ )	U	PLAN (4)					
	NU	PLAN (5)					

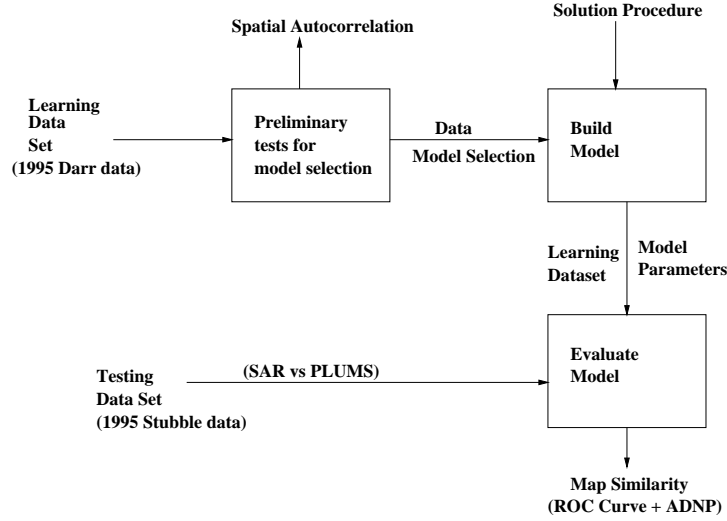
**Figure 6.** Space of design choices for PLUMS components: function family, map-similarity measure, search algorithms and discretization. **G** refers to Greedy search and **SA** refers to Simulated Annealing. **U** and **NU** refer to uniform and non-uniform grid based discretization of parameter space respectively.

the nest locations on the test data. Classification accuracy, which we describe below, was used to evaluate the two models. Then we compare these two models with PLUMS in terms of performance and spatial accuracy (ADNP).

The experimental setup is shown in Figure 7. The data sets used for the learning portion of the experiments, i.e., to predict locations of bird-nests, is shown in Figure 1. Explanatory variables in these data-sets are defined over a spatial grid of approximately 5000 cells. The 1995 data acquired in the Stubble wetland served as the testing data sets. This data is similar to the learning data except for the spatial locations.

We also evaluated PLUMS(A), an instance of PLUMS implementation A shown in Figure 6. PLUMS(A) was 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 was used.

**Metric of Comparison for Classical Accuracy:** We compared the classification accuracy achieved by classical and spatial logistic regression models on the test data. Receiver Operating Characteristic (ROC) [8] curves were used 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



**Figure 7.** *Experimental Method for evaluation spatial autoregression*

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.

**Metric of Comparison for Spatial Accuracy:** We compared spatial accuracy achieved by PLUMS, classical regression and Spatial Autoregressive Regression (SAR) by using 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  $A_k$  stands for 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.

**Result of comparison between PLUMS, Classical regression and SAR models:**

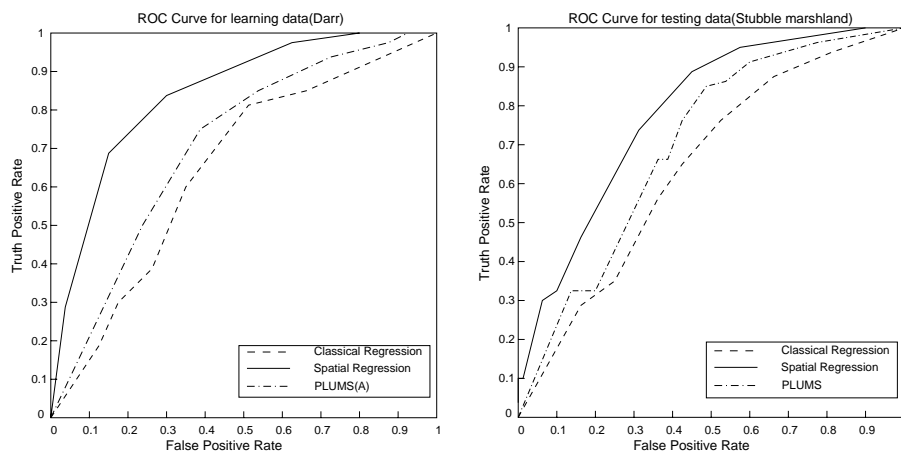
The results of our experiments are shown in Table 2. As can be seen, PLUMS(A) and SAR achieve similar spatial accuracy on test data sets, but PLUMS(A) needs two orders of magnitude less computational time to learn. The run-times for

Data set		PLUMS	Classical	SAR
Learning	spatial accuracy	16.90	47.16	13.96
Testing	spatial accuracy	19.19	41.43	19.30
Learning	Run-time(Seconds)	80	10	19420 <sup>1</sup>

**Table 2.** *Learning time and spatial accuracies*

learning the location-prediction models for the three methods are shown in Table 2. We note that spatial regression takes two orders of magnitude more computation time relative to PLUMS using the public domain code[16] despite the sparse matrix techniques[21] used in the code.

Figure 8(a) illustrates the ROC curves for the three models built using the Darr learning data and Figure 8(b) displays the ROC curve for the Stubble test data. It is clear that using spatial regression 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 6 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.

## 5 Future Work and Conclusion

In this paper we have proposed PLUMS (Predicting Locations Using Map Similarity), a framework for mining spatial data. We have shown how spatial autocorrelation, the characteristic property of spatial data, can be incorporated in the PLUMS framework. When compared with state-of-the-art spatial statistics methods of predicting bird-nest locations, PLUMS achieved comparable spatial accuracy while incurring only a fraction of the cost. Our future plan is to bring in other data mining techniques, including clustering and association rules, within the PLUMS framework. We also plan to investigate other search algorithms and new map-similarity measures.

<sup>1</sup>10,000 draws for Gibbs sampling, 1000 burn-outs

# Bibliography

- [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] Sanjay Chawla, Shashi Shekhar, Weili Wu, and Uygur Ozesmi. Extending Data Mining for Spatial Applications: A Case Study in Predicting Nest Locations. *2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2000)*, Dallas, TX, May 2000.
- [5] Vladimir Cherkassky and Filip Mulier. *Learning From Data Concepts, Theory, and Methods*. John Wiley & SONS Inc., 1998.
- [6] N.A. Cressie. *Statistics for Spatial Data (Revised Edition)*. Wiley, New York, 1993.
- [7] P.J. Diggle. *Statistical analysis of spatial point patterns*. Academic Press, 1993.
- [8] J.P. Egan. *Signal Detection Theory and ROC analysis*. Academic Press, New York, 1975.
- [9] 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.
- [10] 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.
- [11] 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.

- [12] Issaks, Edward, and Mohan Srivastava. *Applied Geostatistics*. Oxford University Press, Oxford, 1989.
- [13] E. Knorr and R. Ng. Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining. *IEEE TKDE*, 8(6):884–897, 1996.
- [14] 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.
- [15] 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.
- [16] J.P. LeSage. Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, (20):113–129, 1997.
- [17] D. Mark. Geographical information science: Critical issues in an emerging cross-disciplinary research domain. In *NSF Workshop*, February 1999.
- [18] H. Miller. Potential contributions of spatial analysis to geographic information systems for transportation(gis-t. *Geographical Analysis*, 31:373–399.
- [19] 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.
- [20] 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.
- [21] R. Pace and R. Barry. Sparse spatial autoregressions. *Statistics and Probability Letters (Publisher: Elsevier Science)*, (33):291–297, 1997.
- [22] 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.
- [23] S. Shekhar and B. Amin. Generalization by neural networks. *IEEE Trans. on Knowledge and Data Eng.*, 4(2), 1992.
- [24] 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*, 11(1), Jan-Feb 1999.
- [25] W.R. Tobler. *Cellular Geography, Philosophy in Geography*. Gale and Olsson, Eds., Dordrecht, Reidel, 1979.