Predicting Locations Using Map Similarity(PLUMS): A Framework for Spatial Data Mining *

Sanjay Chawla
Vignette Corporation
Waltham, Massachusetts
chawla@cs.umn.edu

Shashi Shekhar
Computer Science
University of Minnesota
Minneapolis, MN 55455, USA.
shekhar@cs.umn.edu

Weili Wu
Computer Science
University of Minnesota
Minneapolis, MN 55455, USA.
wuw@cs.umn.edu

ABSTRACT

Spatial data mining is a process to discover interesting, potentially useful and high utility patterns embedded in spatial databases. Efficient tools for extracting information from spatial data sets can be of importance to organizations which own, generate and manage large spatial data sets. The current approach towards solving spatial data mining problems is to use classical data mining tools after "materializing" spatial relationships. However, the key property of spatial data is that of spatial autocorrelation. Like temporal data, spatial data values are influenced by values in their immediate vicinity. Ignoring spatial autocorrelation in the modeling process leads to results which are a poor-fit and unreliable. In this paper we will propose PLUMS(Predicting Locations Using Map Similarity), a new approach for supervised spatial data mining problems. PLUMS searches the space of solutions 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 computational cost. Furthermore, PLUMS provides a general framework for specializing other data mining techniques for mining spatial data.

1. INTRODUCTION

Widespread use of spatial databases [14, 30, 33, 37] is leading to an increasing interest in mining interesting and useful but implicit spatial patterns[19, 24, 12, 29]. 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 logistics, travel and tourism. [2, 15, 17, 21, 28, 34, 38].

Classical data mining algorithms [1, 10] 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 [5, 35]. 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 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 [22].

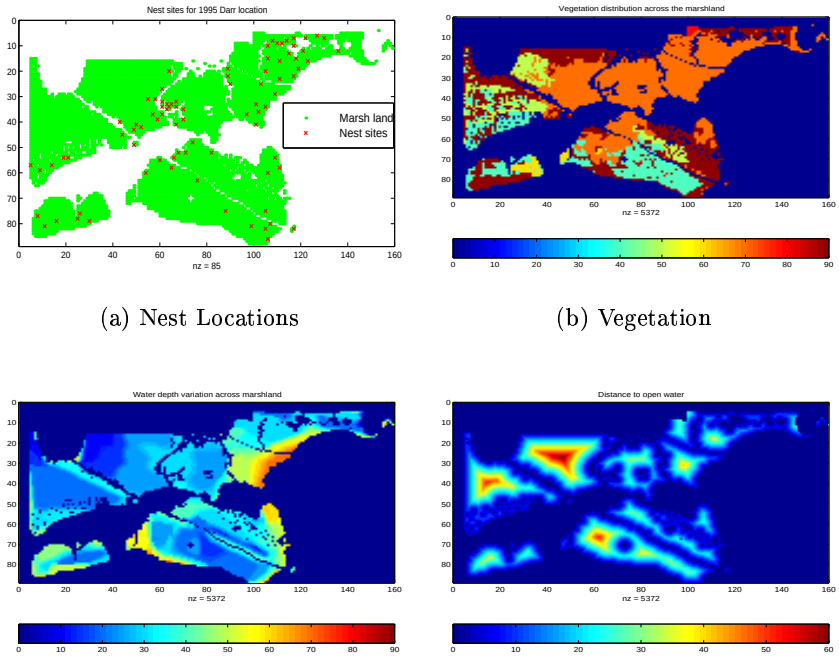
In this paper we will propose PLUMS(Predicting Locations Using Map Similarity), a new approach for supervised spatial data mining problems. PLUMS searches the parameter space of models using a map-similarity measure which is more appropriate in the context of spatial data. We will show that compared 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 to specialize other data mining techniques for mining spatial data.

1.1 An Illustrative Application Domain

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 authors has been involved in the development of spatial model for the nesting locations of a marsh-nesting bird species [25, 26]. 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 be used was collected in 1995 and 1996 from two wetlands(Darr and Stubble) located on the shores of Lake Erie in Ohio. For

*This work was supported 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 9631539.



(a) Nest Locations

(b) Vegetation

(c) Water Depth

(d) Distance to Open Water

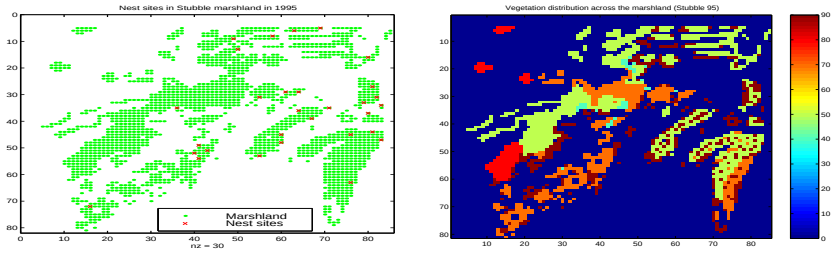
Figure 1: (a) Learning dataset: The geometry of the wetland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the wetland, (c) The spatial distribution of *water depth*, and (d) The spatial distribution of *distance to open water*.

the purpose of data collection, a local coordinate system was established for each wetland and a regular grid consisting of approximately 5000 cells was superimposed. The cells of the grid had square geometries of size 5 meters by 5 meters. In each cell the values of several structural and environmental variables 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 wetland, locations of the nests and spatial distribution of the explanatory variables are shown in Figure 1. The corresponding maps for the Stubble wetland are shown in Figure 2.

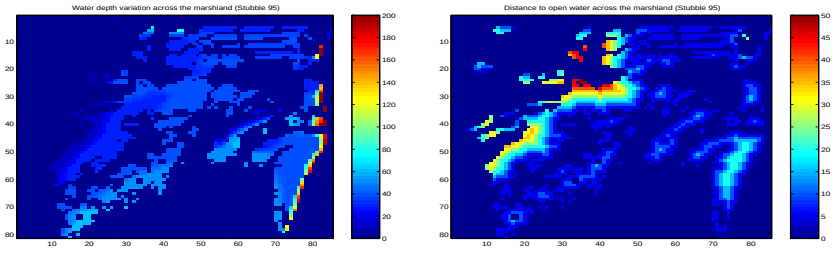
One of the authors has applied classical data mining techniques like logistic regression[26] and neural networks[25] to build spatial habitat models. Logistic regression was used because the dependent variable is binary(nest/no-nest) and the logistic function "squashes" the real line onto the unit-interval. 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[25]. The use of neural networks actually decreased the classification accuracy[25] but led to a better understanding of the interaction between the explanatory and the dependent variable.

There are 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 independent distributions 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 and 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 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 calculate the percentage of correctly classified objects. 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 discretization 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.



(a) Nest Locations (b) Vegetation



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

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

A simple and intuitive measure of spatial accuracy is the Average Distance to 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 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. We now formalize the spatial data mining problem by incorporating notions of spatial autocorrelation and spatial accuracy in the problem definition.

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 $\hat{f}^Y \in \mathcal{F}$.

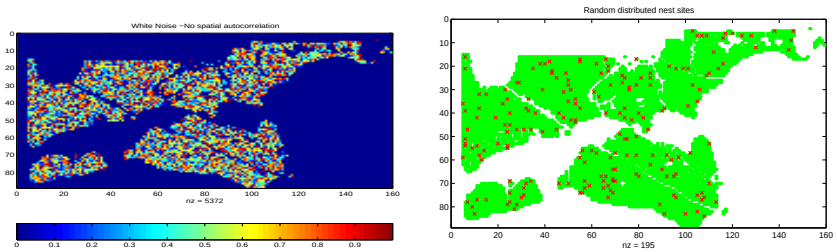
Objective : maximize similarity$(map_{s_i \in S}(\hat{f}^Y(f_{X_1}, \ldots, f_{X_K})), map(f_Y(s_i)))$
$= (1-\alpha)$ classification_accuracy$(\hat{f}^Y, f_Y) +$
(α)spatial_accuracy$((\hat{f}^Y, f_Y)$

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

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



(a) pixel property with independent identical distribution

(b) Random nest locations

Figure 3: Spatial distribution satisfying distribution assumptions of classical regression

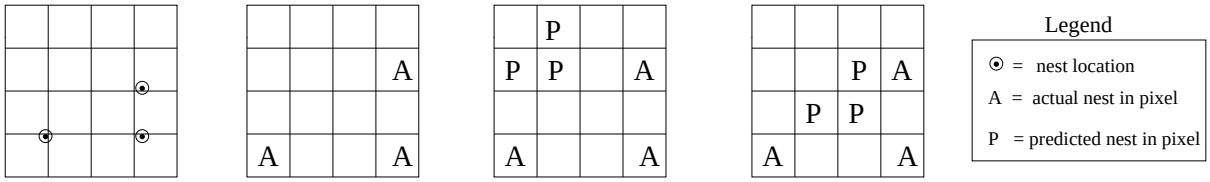


Figure 4: (a)The actual locations of nest, (b)Pixels with actual nests, (c)Location predicted by a model, (d)Location predicted by another mode. Prediction(d) is spatially more accurate than (c).

degree of importance of **spatial accuracy** and is typically domain dependent. As $\alpha \to 0$, 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 [25, 26] is an instance of the location prediction problem. The underlying spatial framework is the collection of 5mX5m 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, $\alpha > 0$.

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 2.3). The PLUMS [2] approach(See Section 3) allows flexibility of incorporating spatial autocorrelation in the model, the objective function or 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 incur only a fraction of the computational costs.

1.3 Related Work and Our Contributions

Related work includes 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 and Ecological modeling [3, 23, 13] has focused on extending classical statistical techniques in order to capture the unique characteristics inherent in spatial data. In Section 2 we will briefly review some basic spatial statistical measures and techniques.

Spatial Data Mining: Spatial data mining [9, 18, 19, 20, 29], a subfield of data mining [1, 10], is concerned with 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 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. 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 [20, 19].

Our contributions: In this paper we will propose a new framework for spatial data mining. This framework con-

[2] An interesting piece of trivia is that there is actually a PLUM bird island just off the coast of Boston, Massachusetts.

sists of a combination of statistical model, a map similarity measure along with a search algorithm and a discretization of the parameter space. We will show that the characteristic property of spatial data, namely, spatial autocorrelation, can be incorporated in the statistical model or the objective function. We will also conduct experiments on the "bird-nesting" data to compare our approach with spatial statistical techniques. The rest of the paper is as follows. In Section 2 we will briefly review some important spatial statistical concepts. In Section 3 we will propose PLUMS, a new framework for spatial data mining. Experiments carried out to compare PLUMS and spatial statistical methods will be elaborated upon in Section 4. We will close in Section 5 with some comments and directions for future work.

2. BASIC CONCEPTS: MODELING SPATIAL DEPENDENCIES

2.1 Logistic Regression Modeling

Given an $n-$vector $\mathbf{y}$ of observations and an $n \times m$ matrix $\underline{X}$ of explanatory data, classical linear regression models the relationship between y and $\underline{X}$ as

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

Here $\mathbf{X} = [1, \underline{X}]$ and $\beta = (\beta_0, \ldots, \beta_m)^t$. The standard assumption on the error vector ϵ is that each component is generated from an independent and and identical and normal distribution, i.e, $\epsilon_i = N(0, \sigma^2)$.

When the dependent variable is binary, as is the case in the "bird-nest" example, the model is transformed via the logistic function and the dependent variable is interpreted as the probability of finding a nest at a given location. Thus, $Prob(y = 1) = \frac{e^{X\beta}}{1+e^{X\beta}}$. This transformed model is referred to as **logistic** regression.

The fundamental limitation of classical regression modeling is that it assumes that the sample observations are independently generated. This may not be true in the case of spatial data. As we have shown in our example application, the explanatory and the independent variables show a moderate to high degree of spatial autocorrelation(see Figure 1). The inappropriateness of the independence assumption shows up in the residual errors, the ϵ_i's. When the samples are spatially related, the residual errors reveal a systematic variation over space, i.e., they exhibit high spatial autocorrelation. This is a clear indication that the model was unable to capture the spatial relationships existing in the data. Thus the model is a poor fit to the data. Incidentally the notion of spatial autocorrelation is similar to that of time autocorrelation in time series analysis but is more difficult to model because of the multi-dimensional nature of space. We now introduce a statistic which quantifies spatial autocorrelation.

2.2 Spatial Autocorrelation and Examples

There are many measures available for quantifying spatial autocorrelation. Each have their own strengths and weaknesses. Here we will briefly describe the Moran 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 is indicative of high spatial autocorrelation. This implies that like values tend to cluster together or attract each other. A low negative value is an indication that 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. The exact definition of MI is given in the Appendix.

All spatial autocorrelation measures are crucially dependent on the choice and design of the contiguity matrix W. The design of the matrix itself is predicated on determining "what constitutes a neighborhood of influence?" 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 5. The contiguity matrix of the uneven lattice(left) is shown on the right hand side. The contiguity matrix plays a crucial role in the spatial extension of the regression model.

2.3 Predicting Locations Using Spatial Statistics

We now show how spatial dependencies are modeled in the framework of regression analysis. This 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

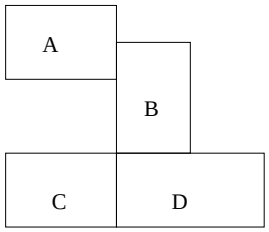
$$\mathbf{y} = \rho W\mathbf{y} + \mathbf{X}\beta + \epsilon.$$

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

We will refer to this equation as the **Spatial Autoregressive Model(SAM)**. Notice 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 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 it 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., higher R-squared statistic(See the Appendix for a dramatic example).

As in the case of classical regression, the SAM equation has to be transformed via the logistic function for binary dependent variables. The estimates of ρ and β can be derived using maximum likelihood theory or Bayesian statistics. We have carried out preliminary experiments using the spatial econometrics matlab package [3] which implements a Bayesian approach using sampling based Markov Chain

[3]We would like to thank James Lesage(http://www.econ.utoledo.edu/lesage) for making the matlab toolbox available on the web.



(a) Map

	A	B	C	D
A	0	1	0	0
B	1	0	1	1
C	0	1	0	1
D	0	1	1	0

b) Contiguity Matrix W

Figure 5: A spatial neighborhood and its contiguity matrix

Monte Carlo(MCMC) methods [23]. 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, and at least for non-statisticians, it is very difficult to decide what "priors" to choose and what are the appropriate analytic expressions 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.

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

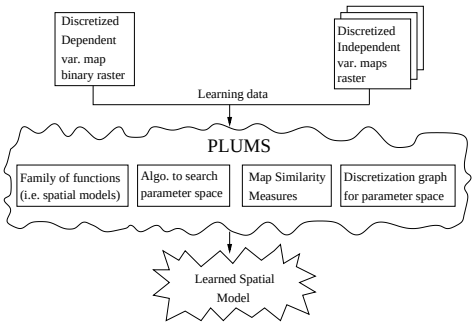
Predicting Locations Using Map Similarity(PLUMS) is the proposed supervised learning approach. Figure 6(a) shows the context and components of PLUMS. It takes a set of maps for explanatory variables and a map for the dependent variable. The maps must use a common spatial framework, i.e. common geographic 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.

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 [7], principal component analysis of a pair of raster maps [31] etc.

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 [4, 11] and simulated annealing [32, 36] etc. A simple function family is the family of generalized linear models, e.g. logistic regression [22] 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

```
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;
}
```

Algorithm 1: greedy-search-algorithm



(a) PLUMS Framework

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

(b) Space of Design Choice

Figure 6: (a)The framework for the location prediction process. (b)Space of Design Choice for PLUMS

spatial autoregression models [3, 23], Geo-statisticians [17] use Co-Kriging and Ecologists [16] 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(b). 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 **??** corresponds to the spatial accuracy measure(ADNP), generalized linear model(for the family of functions), a greedy search algorithm and uniform discretization.

4. EXPERIMENT DESIGN AND EVALUATION

Goals: The goals of the experiments are (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 experimental setup is shown in Figure 7. 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 the nest locations on the test data. Classification accuracy, which we describe next, was used to evaluate the two models. Then we compare these two models with PLUMS in terms of performance and spatial accuracy(ADNP).

Metric of Comparison for Classification accuracy: Classification accuracy achieved by classical and spatial logistic regression are compared on the test data. We use the Receiver Operating Characteristic(ROC) [8] 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.

Metric of Comparison for Spatial Accuracy Spatial accuracy achieved by PLUMS, classical regression and SAM(Spatial Autoregressive Model) are compared based on 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.

Result of Comparison between Classical and Spatial Regression (SAM) models: We use the 1995 Stubble wetland data to make comparison between the two models. The result is shown in Figure 8. Clearly, by including a spatial autocorrelation term, there is substantial and systematic improvement for all levels of cut-off probability on both the learning data(1995 Darr) and test data(1995 Stubble). However, the performance of SAM model is very slow and not scalable. The choice of contiguity matrix w is non-trivial, but very crucial to SAM model.

Result of comparison between PLUMS, Classical regression and SAM models: We carried out experiments to compare PLUMS with classical and spatial regression models. For this we also used the 1995 data acquired in the Stubble wetland. 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 SAM, while it needs order of magnitude less computational time

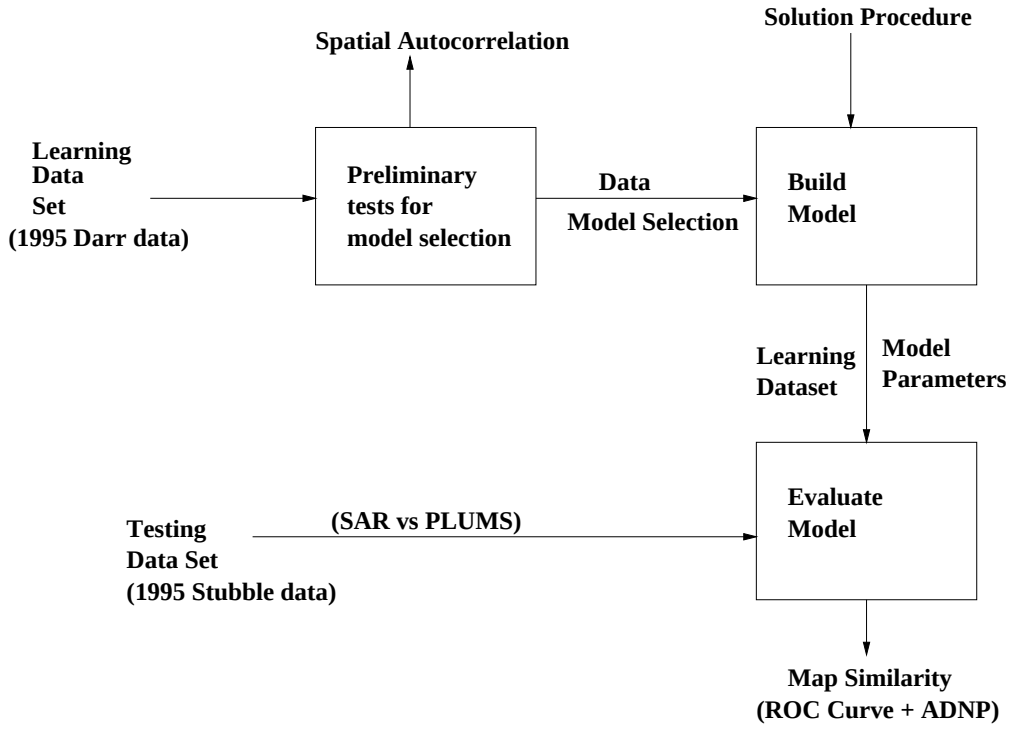


Figure 7: Experimental Method for evaluation spatial autoregression

to learn.

The run-time for learning 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 [23] despite the sparse matrix techniques [27] used in the code.

Figures 9(a) is the ROC curves for the three models built using the Darr learning data and Figure 9(b) is the ROC curve for the Stubble test data. It is clear that by 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 **??** need to be explored to close the gap.

5. FUTURE WORK AND CONCLUSION

In this paper we have proposed PLUMS(Predicting Locations Using Map Similarity), a framework for spatial data mining. 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 method in predicting bird-nest locations, PLUMS achieved comparable spatial accuracy while incurring only a fraction of the cost. Furthermore, PLUMS provides a template for specializing other data mining techniques for spatial data.

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 , new map-similarity measures and non-uniform parameter spaces and determine their dominance zones.

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

6. ADDITIONAL AUTHORS

Additional authors: Uygar Ozesmi (Department of Environmental Sciences, Ericyes University, Kayseri, Turkey, email: `uygar.ozesmi-1@tc.umn.edu`)

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

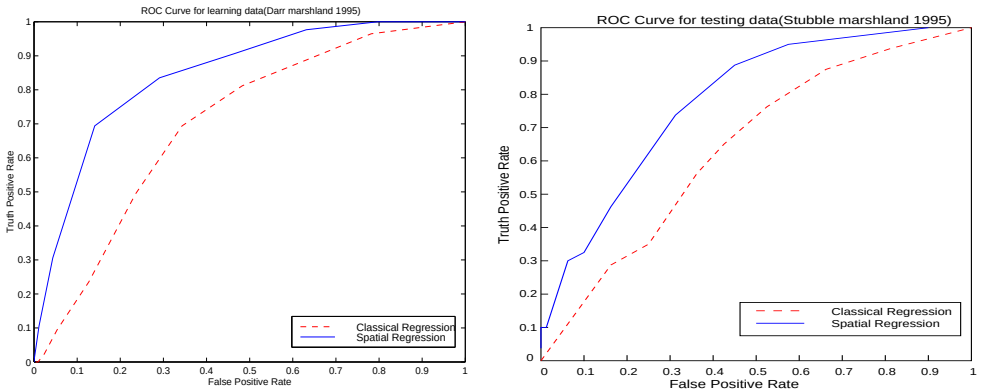
[5] P. Could. *The Geographer at Work.* Routledge and Kegan Paul, London, 1985.

[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*



(a) Learning Data

(b) Test Data

Figure 8: (a) Comparison of the logistic and logistic with spatial autocorrelation on the 1995 Darr wetland learning data. (b) Comparison of the two models on the 1995 Stubble wetland testing data.

Conference on Artificial Intelligence, pages 61–74, Bonn, Germany, September 1999.

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

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

[18] E. Knorr and R. Ng. Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining. *IEEE TKDE*, 8(6):884–897, 1996.

[19] 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.

[20] 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.

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

[22] 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.

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

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

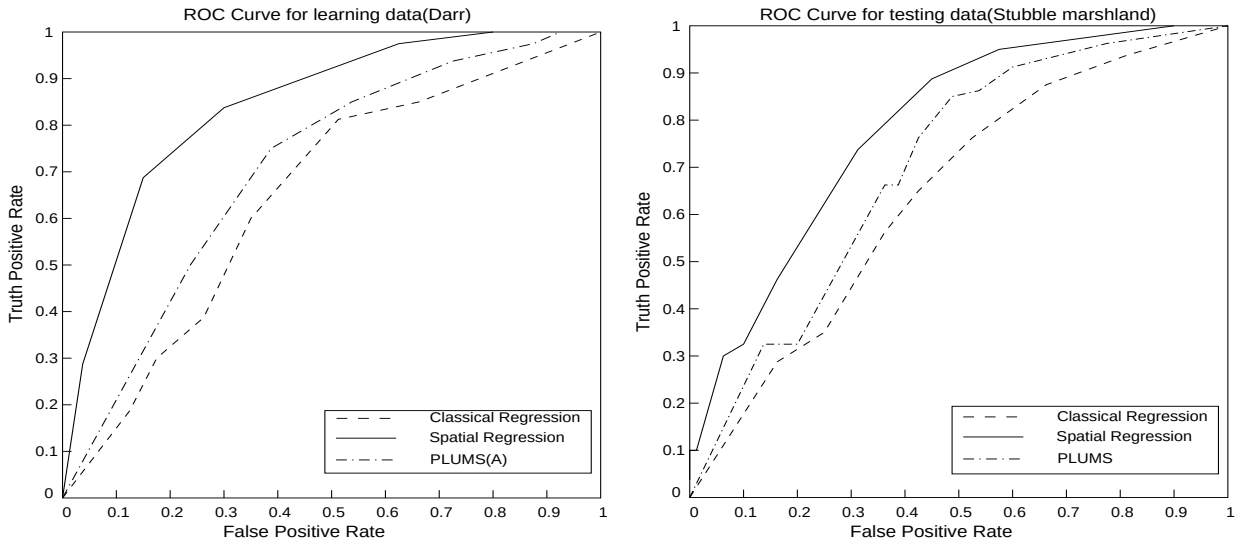
[25] S. Ozesmi and U. Ozesmi. An Artificial neural network approach to spatial habitat modeling with interspecific interaction. *Ecological Modelling (Publisher: Elsevier Science B. V.)*, (116):15–31, 1999.

[26] 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.

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

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

[29] John F. Roddick and Myra Spiliopoulou. A bibliography of temporal, spatial and spatio-temporal data mining research. *ACM Special Interest Group on*



(a) ROC curves for Learning Data

(b) ROC curves for Test Data

Figure 9: (a) Comparison of PLUMS(A) with other methods on the Darr learning data. (b) Comparison of the models on the test data.

Knowledge Discovery in Data Mining(SIGKDD) Explorations, 1999.

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

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

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

[33] 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.

[34] 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.

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

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

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

[38] 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.

8. APPENDIX:SPATIAL AUTOCORRELATION

8.1 Moran's I measure

There are many measures available for quantifying spatial autocorrelation. Each have their own strengths and weaknesses. The two most well known measures are Moran's I and Geary's C measure. Here we will briefly describe the Moran 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 is indicative of high spatial autocorrelation. This implies that like values tend to cluster together or attract each other. A low negative value is an indication that high and low values are interspersed. Thus like values are de-clustered and tend to repel each other. A smooth surface will have a high spatial autocorrelation and a chessboard-like surface a high negative spatial autocorrelation. A value close to zero is an indication that no spatial trend (random distribution) is discernible using the given measure.

The formula for MI is

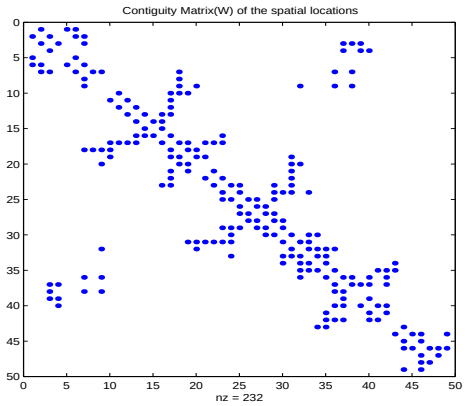
$$MI = \frac{n}{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}} \cdot \frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}$$

where n is the number of data points, $x_i's$ are the data values, $\bar{x}$ is the mean and W is the design or contiguity matrix. All spatial autocorrelation measures are crucially dependent on the choice and design of the contiguity matrix W.

8.2 Summary of different methods

We summarized all the methods that have been used to build the bird habitat model in Table 3.

8.3 Example of including the spatial autoregressive term



(a) The contiguity matrix of locations

	R-square	Moran I (residual)
Ordinary Regression	0.5521	0.23
Spatial Auto Regression	0.6518	0.04

(b) R-square and Moran I of residual

Figure 10: (a)Crime data set in 49 neighborhoods in Columbus Ohio [3],where number of crime incidents is dependent variable, and the explanatory variables include mean income and mean house value.(b)Coefficient of determination(R^2) and Moran I results of this data set via ordinary regression model and Spatial Auto Regression(SAM) model. Clearly the SAM model provides a better fit than the classical regression model.

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

Data set		PLUMS	Classical	SAM
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 [1]

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

Method Name	Model Type	Spatial AC	Dependent Var. Type	Accuracy Measure	Solution Procedure
Linear Regression	Linear	No	Numeric	Total Square Error(TSE)	Closed Form
Neural Networks	NonLinear	No	Numeric/ Categorical	TSE	Gradient Descent Back-Propagation
Probit	Gen. Linear	No	Binary	TPR/FPR	Gradient Descent
Logit	Gen. Linear	No	Binary	TPR/FPR	Gradient Descent
SAM + Probit	Gen. Linear	Yes	Binary	TPR/FPR	ML/EM/Gibbs

Table 3: Different methods and their characteristics that have been used for building the bird habitat model.