Modeling Spatial Dependencies for Mining Geospatial Data : An Introduction

Sanjay Chawla, Shashi Shekhar, Wei Li Wu
Department of Computer Science
University of Minnesota
Minneapolis, MN 55455
{chawla,shekhar,wuw}@cs.umn.edu

Uygar Ozesmi
Department of Environmental Sciences
Ericyes University
Kayseri, Turkey
uygar.ozesmi-1@tc.umn.edu

February 1, 2000

Contents

Abstract

Geo-spatial data mining is a process to discover interesting and potentially useful spatial patterns embedded in spatial databases. Efficient tools for extracting information from geo-spatial data sets can be of importance to organizations which own, generate and manage large geo-spatial data sets. The current approach towards solving spatial data mining problems is to use classical data mining tools after "materializing" spatial relationships and assuming independence between different data points. However, classical data mining methods often perform poorly on spatial data sets which have high spatial autocorrelation. This approach often leads to poor results because it does not take into account the fundamental notion of spatial autocorrelation. In this paper we will review statistical techniques which can effectively model the notion of spatial-autocorrelation. We will also present a "roadmap" for extending current techniques to manage geospatial data which will the serve as basis for future research.

1 Introduction

Widespread use of spatial databases [1, 2, 3] is leading to an increasing interest in mining interesting, useful but implicit spatial patterns just as the widespread use of relational database triggered interest in classical data mining . Efficient tools for extracting information from geo-spatial data - the focus of this work, can be of importance to organizations which own, generate and manage large geo-spatial data sets. Data mining products can be a useful tool in decision-making and planning just as they are currently in the business world. Knowledge extraction from geo-spatial data has also been highlighted as a key area of research in a recently concluded NSF workshop on GIS vision for 2010 [5].

Classical data mining algorithms often perform poorly on spatial data because spatial data sets exhibit a spatial continuity property between neighboring objects. In other words the values of attributes of nearby spatial objects tend to systematically affect each other. In classical geography this property is often referred to as the first law of geography: Everything is related to everything else but nearby things are more related than distant things [6]. In spatial statistics, an area within statistics devoted to the analysis of spatial data, this is called spatial autocorrelation [7]. Ignoring spatial autocorrelation may lead to residual errors that vary systematically over space exhibiting high spatial autocorrelation. The models learnt may turn out to be not only biased and inconsistent but may also be a poor fit to the data set.

In this paper we will review techniques from spatial statistics which explicitly take into account effects of spatial autocorrelation. We will apply these techniques to an example from ecology to predict the location of bird nests in marshelands. We will show that by taking spatial autocorrelation into account the accuracy of the results show a substantial improvement. We will also highlight some of the shortcomings of spatial autocorrelated models and set the stage for future reserach.

1.1 Historical Examples and Potential Applications of Geo-Spatial Data Mining

Data mining has received a lot of attention in the general media thanks to the infamous "Diaper-Beer" example. Using data mining techniques, mainly association rules, researchers at a retail outlet discovered that "People who buy diapers in the afternoon also tend to buy beer." The researchers were not searching for this particular pattern or correlation between the two items but somehow it just "popped up". Thus, it was claimed, that data mining can search for hidden nuggets of information embedded in large volumes of data which otherwise would have been ignored.

There have been similar but more serious and valuable revelations related to spatial data. The three famous ones are [39]:

1. In 1855 the Asiatic cholera was sweeping through London. A leading epidemiologist marked on a map all the spatial locations where the cholera victims were residing. The locations formed a cluster and the centroid of the cluster turned out to be a water-pump. The government authorities turned-off the water pump and the cholera epidemic subsided.

2. The theory of Gondwanaland that the all the continents formed one land mass was postulated after R. Lenz discovered(using maps) that all the continents could be fitted together like one giant jigsaw puzzle.

3. In 1909 a group of dentists discovered that the residents of Colorado Springs had unusually healthy teeth and they attributed it to high level of natural floride in local drinking water wells.

GSDM has the potential of having a profound impact in many application areas. According to some estimates the size of data being collected is doubling every twenty months and atleast eighty percent of the data housed in databases has a spatial component. We now give some specific examples where GSDM can have an immediate effect.

Location of ammunition dump: Ammunition dumps are buried across historic battlefields and military lands throughout the United States. In many instances records which identified the location of these ammo dumps have been lost. It is important for the army to predict the location of these dumps with minimal digging [26]. This is similar to predicting potential location of oil reserves in an area to minimize unproductive drilling.

Ecological Management: The Department of Defense (DoD) is one of the biggest Federal landowners in the country. Its 425 military installations span an area of more than 25 million acres. Security considerations have effectively cut-off this area from the deleterious side effects of development. As a result some of the most pristine natural habitat can be found in land owned by the DoD. While DoD land is primarily used for military training and testing, and thus forms a unique category of land use, it is subject to various environmental protection laws. The general objective of the military land management is officially stated as: to optimize use of the land for training and testing activities, while ensuring compliance with state and federal laws and the lands' long-term sustainability as a resource asset [27]. The question we would like to pose is: Can spatial data mining be used as an effective tool for land and habitat management in general and for the DoD in particular?

Insurgency and Crime Analysis The identification of crime "hot spots" and search for explanatory variables which cause unusually high levels of crime is an area where spatial analysis can have a profound affect.

Simulation The army research labs carry out large simulations of explosions and terrorist bombings in an urban environment and their effect on buildings and other structures. These simulations have generated huge amounts of data which the army has access to. Because these simulations are usually modeled as coupled nonlinear partial differential equations it is difficult to theoretically establish the convergence of the numerical schemes exhibit Thus, in many instances these simulations diverge and the whole experiment has to be started again [28]. We want to experimentally investigate the relationship between spatial autocorrelation and convergence. By calculating the spatial autocorrelation of the variable of interest at each time step we can map the simulation run into a time series. Then we can build a classification model to predict the future behavior of the time series.

1.2 Geo-Spatial Data Mining(GSDM) and related works

Data mining [8, 9] is the process of extracting information from large volumes of data housed in databases. Data mining draws strengths from many different areas but mainly lies at the intersection of *machine learning, statistics and databases* [10]. Spatial data mining [11, 16, 12] is a subfield within data mining with exclusive emphasis on geo-spatial data. Sources of geo-spatial data abound and include satellite imagery, cartographic maps, census data and modeling runs of partial differential equations. Despite being a relatively new discipline, data mining has caught the imagination of the scientific and business world and is being extensively used as a tool in research and high level decision making. Example applications of data mining in the corporate environment are credit card fraud detection and charting the buying habits of customers. Potential scientific application of geo-spatial data mining include the optimal geographical deployment of military assets, finding interesting spatial patterns in historical climate databases and characterization of the spatial habitat of animals which appear on the list of endangered species.

Classical and Geo-Spatial data mining techniques can be partitioned into three categories: *descriptive, explanatory, and predictive.* Descriptive models characterize the distribution of the spatial phenomenon. Descriptive models are based on a set of spatial statistics and indices [7]. For example, a spatial distribution may be classified into random or clustered using nearest neighbor index or quadrat analysis [45].

Explanatory model deals with spatial associations, i.e. relationships between a phenomenon and the factors affecting its spatial distribution. For example, in order to explain why bird nests clusters occur in a certain area, roles of water bodies, vegetation weather patterns etc. may be examined. More detailed analysis may explore how each factor may influence the bird nest locations. Example techniques are based on chi-square tests, associations and spatial auto-correlation coefficients(for example Moran's I) using appropriate geographic units .

Predictive models are used to solve specific problems about predicting the values of some attributes given the value of the other attributes. For example given certain weather parameters(temperature, humidity, pressure) the meteorologist would like to know whether it will snow or not. Examples of predictive models include classification, regression, etc. In a supervised learning scenario, a data set, called the training set, is used to build a prediction model. Depending upon the type of data and domain knowledge , many techniques can be used to build the model. Examples include decision trees, linear regression, logistic regression and neural networks. The model is evaluated on its performance on test data. For example, a model about predicting snow fall can be built using historical weather data and one of the above mentioned techniques. The quality of the model will be judged on the basis of accuracy of snowfall prediction in the future. 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 [13]. This approach has fundamental shortcomings, partly because it is difficult to decide *apriori* which relationships to materialize. Researchers in spatial econometrics [18] and regional science [39] have developed comprehensive techniques to incorporate spatial information into statistical models. One of the goals of this work is to introduce these techniques to a wider audience.

1.3 What is special about GSDM

Classical data mining algorithms are often based on the assumption that variables are randomly and independently generated. This assumption implies that two events which are spatially close to one another have no effect on each other and therefore have the same mean value. This assumption is often not true for spatial data sets leading to poor performance of classical data mining techniques on spatial data sets. It is important to relax this assumption to quantify the spatial dependence and factor it into techniques for the estimation of missing values. Consider the four images shown in Figure 1. Each of these images show a different degree of interrelationship between neighboring values of an attribute. Figure 1(a) shows variation of attribute values if there is no interrelationship between neighboring value(white noise). Figure 1(d) shows an attribute whose values are substantially dependent on neighboring values. We will return to Figure 1 in Section 2 and will show how this relationship can be quantified.

In summary, challenges in geo-spatial data mining arise from following issues. First, classical data mining treats each attribute value to be independent of other values of the same attribute, whereas spatial patterns often must satisfy the constraints of continuity and high autocorrelation among nearby features. For example, population- densities, house prices, soil type, etc., of nearby locations are often related. Second, classical data mining deals with numbers and categories. In contrast, spatial data is more complex and includes extended objects such as points, lines, and polygons [15]. Finally, classical data mining works with explicit inputs, whereas spatial predicates (e.g. overlap) and attributes (e.g. distance, spatial auto-correlation) are often implicit [16].

2 An illustrative application domain

One of the authors has been involved in the development of a GIS-based spatial model for marsh-nesting bird species [20, 21]. We will use this application to highlight some of the unique and distinguishing aspects of GSDM *vis-a-vis* classical data mining.

Habitat selection and habitat suitability are key concepts in wildlife management, protection of critical habitat and conservation of sensitive and endangered species. Remote sensing and Geographical Information Systems(GIS) technology make possible the use of spatial modeling for habitat selection and suitability. Spatial models are useful in predicting microhabitat preference of species by synthesising existing theories with rigorous field research. Spatial habitat models can be used for conservation and management decisions in sensitive habitats such as coastal wetlands.

Both conservation and management of individual wetlands require small-scale understanding of micro-habitat characteristics. Habitat selection in birds occurs in a hierarchical sequence from a larger scale to a finer scale. The first level of choice is geographical region, the second is is landscape, third is habitat, and last is a path in a habitat. This study was at a micro-scale and as a consequence extensive field analysis, which generated a large quantity of data, was required.

The field work was carried over a period of two years on two marshes situated on the banks of Lake Erie in Ohio. The marshlands were partitoned into a grid framework and at each grid cell eight structural and enviormental characteristics were recorded. These were: *water depth, dominant vegetation type by durability index, vegetation stem density, stem height, distance to edge of basin, distance to open water, distance to edge by depth, distance to edge by stem density.* Besides values related to these eight layers the presence or absense of a bird-nest(red-winged blackbird) was recorded. The goal of the study was to build a model to predict the presence/absence of bird-nests as a function of the eight recorded variables. The spatial geometry of the marshland and the locations of the nests is shown in Figure 2.

We have chosen this study because of several desirable properties:

1. Spatial factors like distance to the edge of marsh and distance to the open water were deemed important in the study.

2. Models based on classical data mining techniques(logistic regression and neural networks) have not performed well despite incorporation of significant domain knowledge in selection of variables. The success rate in predicting nest locations was 20% better than a random selection.

3. Many of the explanatory variables exhibit high spatial autocorrelation suggesting that a geo-spatial data mining techniques may lead to better results.

2.1 Logistic Regression Modeling

Given a dependent variable y consisting of n observations and a matrix X of m explanatory variables the classical linear regression models the dependency of y on X using the standard linear equation

$$y_i = \beta_0 + \sum_j \beta_j X_{ij} + \epsilon_j \quad i = 1, \ldots, n \; j = 1, \ldots, m$$

where the error terms $\epsilon_i \approx N(0, \sigma^2)$ are assumed to be independent and identically distributed Gaussian random variables. In matrix form the equations can be written as

$$\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1m} \\ 1 & X_{21} & \dots & X_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \dots & X_{nm} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \vdots \\ \epsilon_n \end{pmatrix}$$

or, in a more compact vector form

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

For example, in our application domain the dependent variable, y is the presence/absence of a bird nest and the explanatory variable are the eight variables described above. When the dependent variable y is binary$(0, 1)$, the classical regression model may not be suitable for the following two reasons [19]:

1. In the classical model it is assumed that the error terms are identically and independently distributed(iid) with constant mean and variance. This cannot be true when the dependent variable is binary because the error is $X\beta$ when $y = 0$ and $1 - X\beta$ when $y = 1$.

2. There is no way to guarentee that the predicted values from classical regression model will in the $(0, 1)$ interval. To ensure that the predicted values lie with the $(0, 1)$ interval we have to formulate the model in a way such that

$$\lim_{X\beta \to +\infty} Prob(y = 1) = 1 \quad (1)$$

$$\lim_{X\beta \to -\infty} Prob(y = 1) = 0 \quad (2)$$

$$(3)$$

Two distributions that have been used to produce such outcomes are the logistic(logit) and the cumulative normal(probit) distributions. For the logit case the distribution is

$$Prob(y = 1) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

and for the probit case the distribution used is

$$Prob(y = 1) = \frac{1}{2\pi} \int_{-\infty}^{X\beta} e^{-\frac{t^2}{2}} dt$$

The logit and the probit model are called generalized linear models because if

$$\theta = Prob(y = 1) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

then, λ as defined below follows the classical linear regression model:

$$\lambda = \log(\frac{\theta}{1 - \theta}) = X\beta.$$

2.2 Experiment Design, Results and Scope for Improvement

The initial experiment design is shown in Figure 3. The purpose here was to recreate the results in the original study. We used the 1995 data to build a predictive model using logistic regression. The output of the model were the regression coefficients, β. The model built using 1995 data was tested on the data generated in 1996.

For binary dependent variable the standard way of computing the accuracy of a model is to use Receiver Operating Characteristic(ROC) curves [46]. ROC curves reveal the relationship between the true positive rate(TPR) and the false positive rate(FPR). The TPR and the FPR are defined as

$$TPR(b) = \frac{\sum I_{[\hat{y}_i > b]} I_{[y_i = 1]}}{\sum y_i} \tag{4}$$

$$FPR(b) = \frac{\sum I_{[\hat{y}_i > b]} I_{[y_i = 0]}}{\sum (1 - y_i)} \tag{5}$$

$$\tag{6}$$

Here $\hat{y}_i$ is the probability that $y_i = 1$ and I is an indicator function. Thus for cut-off probability b, $0 \leq b \leq 1$, $I_{[\hat{y}_i > b]} I_{[y_i = 1]} = 1$ if $\hat{y}_i > b$ and $y_i = 1$.

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 4(a) is the ROC curves for the model built using the 1995 data and Figure 4(b) is the ROC curve for the 1996 data for the model built with 1995 data. The following observations are worth noting.

1. The ROC curves for both the probit and the logit model are almost identical.
2. On the learning data(1995) the ROC curves are significantly above the line $TPR = FPR$. This is not true(as expected) on the 1996 test data.

2.3 Limitations of classical 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 variables show a moderate to high degree of spatial autocorrelation(see Figure 1).

For spatial data and observations the validity(or invalidity) of the independence assumption shows up in the residual errors, the ϵ_i's. When the samples are spatially related the residual error reveal a systematic variation over space, i.e., they exhibit high spatial autocorrelation. Besides leading to a poor fit of the model, ignoring the effects of spatial autocorrelation may yield biased and inconsistent estimate of the relationship between y and X. The property 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 show how spatial autocorrelation can be incorporated into regression modeling.

3 Modeling Spatial Dependencies

We will 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.

3.1 Framework

In spatial statistics autocorrelation measures are used to quantify the spatial dependence between the values of a given spatial variable. If the dependent variable or the error terms in a regression model exhibit "high" spatial autocorrelation then a suitably modified regression model can be used to quantify the relationship between dependent and explanatory variables. The solution of this model entails solving a non-linear equation in the model parameters.

3.2 Spatial Autocorrelation and Examples

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 and refer the reader to standard books on spatial statistics [17] for a description of the Geary's C 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, $x_i^{\prime}s$ 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. 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.

3.3 Spatial Error and Autoregressive Regression Models

In spatial regression the spatial dependencies of the error term or the dependent variable are directly modeled in the regression equation [18]. In the Spatial Error Model (SEM), the error terms are assumed to be spatial autocorrelated and then the regression equation becomes

$$y = X\beta + u \tag{7}$$
$$y = \rho W u + \epsilon \tag{8}$$

Here W is the neighborhood relationship contiguity matrix and ρ is a parameter that reflects the strength of spatial dependencies between the error terms.

Instead of modeling autocorrelation in the error terms we could instead assume that the dependent values y_i' are related to each other. That is

$$y_i = f(y_j) \; i \neq j.$$

Then the regression equation can be modified as follows

$$y = \rho W y + X\beta + \epsilon.$$

This model is called Spatial Autoregressive Regression (SAR). Notice when $\rho = 0$, this equation collapses to the standard regression equation. The benefits of modeling spatial autocorrelation are many:

1. The residual error would have much lower spatial autocorrelation, i.e., systematic variation. With proper choice of W, the residual error should, atleast 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. The magnitude of parameter are likely to be smaller, relative to the classical linear regression model(Section 1.1) since the y-values are partially explained by the neighboring y-values.

4. Finally, the model will have a better fit, i.e., higher R-squared statistic.

3.3.1 Solution Procedure

Estimates of the parameters ρ and β can be derived using the maximum likelihood procedure of estimating parameters. The following steps are used [19]:

1. Compute the least square estimate $b_0 = (X'X)^{-1}X'y$.

2. Estimate $b_1 = (X'X)^{-1}X'Wy$.

3. Define $e_0 = y - Xb_0$ and $e_l = Wy - Xb_l$.

4. The best estimate $\hat{\rho}$, can be derived by maximizing the likelihood function.

   $$\hat{\rho} = argmax \log |I_n - \rho W| - \frac{n}{2} \log[(e_0 - \rho e_l)'(e_0 - \rho e_l)$$

   The maximization is over the range $\frac{1}{\lambda_1} < \rho < \frac{1}{\lambda_2}$, the minimum and maximum eigenvalue of the standardized spatial weight matrix W.

The first three steps are computed only once and step 4 is an iterative procedure to compute which maximizes the log-likelihood function. It is clear that for large data sets, computing the determinant -(in step 4) is an extremely computationally expensive task raising issues of numerical accuracy and computational efficiency. The spatial contiguity matrix W is sparse and methods have been proposed to use sparse matrix techniques to speed up the solution procudure [47].

3.3.2 Probit with Spatial Autocorrelation

To deal with binary dependent variables the SAR model has to be modified using either a logistic or a cumulative normal distribution. Thus the model based on logistic distribution is

$$Prob(y = 1) = \frac{e^{\rho Wy + X\beta}}{1 + e^{\rho Wy + X\beta}}$$

and the model based on cumulative normal distribution is

$$Prob(y = 1) = \frac{1}{2\pi} \int_{-\infty}^{\rho Wy + X\beta} e^{-\frac{t^2}{2}} dt$$

We have carried out preliminary experiments using the spatial autogressive probit model. Our solution procedure follows [41] and we used the spatial econometrics matlab package(see Acknowledgements) to carry out these experiments.

The solution procedure for the spatial probit and spatial logit case is non-trivial and involves either the use of the Expectation Maximization(EM) algorithm [42] or Gibbs sampling if a Bayesian approach is used [41]. We will not describe the solution procedure here and the interested reader is urged to consult the above mentioned references.

The results of our experiments are shown in Figure 6. Clearly, by including a spatial autocorrelation term the predictive power of the model shows substantial improvement on the learning 1995 data for all cut-off probabilities. On the other hand, the results on the 1996 test data show only a small improvement over the probit model with the autocorrelation term. Clearly there is substantial room for improvement.

We have summarized all the methods that have been used to build the bird habitat model in Table 1

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-Propogation
Probit	Gen. Linear	No	Binary	TPR/FPR	Gradient Descent
Logit	Gen. Linear	No	Binary	TPR/FPR	Gradient Descent
SAR + Probit	Gen. Linear	Yes	Binary	TPR/FPR	ML/EM/Gibbs

Table 1: Different methods and their characterisitics that have been used for building the bird habitat model.

4 Critique of spatially autocorrelated models

The presence of spatial autcorrelation in the sample data is a distinguishing characterisitc of spatial data. The goal of Geo-Spatial data mining is to quantify spatial autocorrelation and incorporate it in the model building phase. As we have demonstrated, there are well developed techniques in spatial statistics, like spatial autoregression, that do just that. Despite the success of spatial autoregressive regression models there are some serious shortcomings which we list below:

1. *Design of the right contiguity matrix.* In our application domain we have worked with the four nearest neighbor or the eight nearest neighbor template to generate the contiguity matrix. This approach is simple but adhoc. For example, some birds are territorial and other form clusters. This has a substantial effect on the spatial configuration of nest sites. Designing the right contiguity matrix for a particular application is extremely important and difficult at the same time.

2. *Size of the contiguity matrix.* The size of the contiguity matrix is quadratic in the number of data records. This can lead to matrix sizes which are too large to be processed by conventional methods. There has been research related to application of sparse matrix methods but more is needed if the spatial autogressive models are to be applied to large and realistic databases. Infact one of the characteristics of data mining is to design algorithms which can scale to extremely large databases.

3. *The sensitivity of spatial autocorrelation measures to scale.* It is well known that the various measures of spatial autocorrelation are dependent on the scale of application and neighborhood choice. For example, a demographic study based on a census block and at the county level (aggregation of census blocks) may potentially lead to different and sometimes contradictory conclusions. In spatial statistics this is referred to as the Modifiable Area Unit Problem [24]. This problem may be related to the choice of discretization of numerical attributes into categorical labels in classical data mining.

4.1 A new measure for binary spatial classification

The ROC measure may not be the most suitable for binary spatial classification problems. For example consider Figure 7. Here we have shown an artificial example of actual(A) nest sites and two models which predict(P) the locations of nest sites. Both models will result in the same ROC curve but clearly the model which predicted the locations shown in Figure 7(c) is better than one shown in Figure 7(b). This is because the predicted nests in Figure 7(c) are closer to the actual nest site than those predicted in Figure 7(b). We are in the process of designing a new measure based on this observation.

5 Conclusion

We have presented an overview of statistical methods which explicitly incorporate the spatial characteristics of geographically referenced observed data. In particular we have shown how spatial autocorrelation can be included in regression analysis. For spatially correlated binary data the model has to be modified to guarentee that the predictions(rather their probability) lie in the unit interval. The parameters of the binary model are calculated using the well known method of Gibbs sampling. We carried out experiments using data from "conservation ecology" to demonstrate the usefulness of building models which take spatial effects into consideration. The results of the model led to substantial improvement in overall accuracy. We have also critqued the current methods based on spatial autocorrelation and shown why the current measures of classification may not be most suitable for spatial problems.

6 Acknowledgements

We would like to thank James Lesage(http://www.econ.utoledo.edu/l̃esage) and M. Dang(http://www.hds.utc.fr/m̃dang) for making their software available on the web. The work is sponsored in part by the United States Army High Performance Computing Reserach Center under the auspices of the Department of the Army, Army Research Laboratory Cooperative Agreement No. DAAH04-95-2-0003 and Contract No. DAAH04-95-C-0008, the contents of which do not necessarily reflect the position or the policy of government; no official endorsment shoudl be inferred. This work is also supported, in part, by the National Science Foundation under Grant No. 9631539.

References

[1] R.H. Guting. *An Introduction to Spatial Database Systems.* VLDB Journal, October 1994.

[2] S.Shekhar, S.Chawla, S. Ravada, A.Fetterer, X.Liu and C.T. Liu. *Spatial databases: Accomplishments and Research Needs. IEEE TKDE, Jan-Feb 1999.*

[3] S.Shekhar , S.Chawla. Spatial Databases: Issues, Implementation and Trends *(To be published by Prentice Hall, 2000). http://www.cs.umn.edu/ shekhar.*

[4] P. Bradley, U. Fayyad, C.Reina. Scaling EM(Expectation-Maximization) Clustering to Large Databases. *Microsoft Research, MSR-TR-98-35, 1998.*

[5] D. Mark. Geographical Information Science: Critical Issues in an Emerging Cross-Disciplinary Research Domain. *NSF Workshop, Feb. 1999.*

[6] W.R. Tobler. Cellular Geography, Philosophy in Geography, *Gale and Olsson, Eds., 379-86. Dordrecht, Reidel, 1979.*

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

[8] R. Agrawal. Tutorial Database Mining. *PODS, 75-46, 1994.*

[9] U. M. Fayyad. Knowledge Discovery in Databases: An Overview. *ILP:3-16, 1997.*

[10] H. Mannila. Data Mining: Machine Learning, Statistics, and Databases. *SSDBM:2-9, 1996.*

[11] K. Koperski, J. Adhikary, J. Han. Spatial Data Mining: Progress and Challenges. *DMKD:0-10,1996.*

[12] M. Ester, H-P Kriegel, J. Sander. Knowledge Discovery in Spatial Databases. *KI:61-74, 1999.*

[13] G. Andrienko, N. Andrienko. GIS Visualization Support to the C4.5 Classification Algorithm of KDD. *19th International Cartographic Conference Proceedings, International Cartographic Association, Ottawa, pp. 747-755, 1999.*

[14] B. Flury. A First Course in Multivariate Statistics. *Springer, 1997.*

[15] *M. Egenhofer.* What's Special about Spatial?–Database Requirements for Vehicle Navigation in Geographic Space. *SIGMOD Record, 22 (2): 398-402, June 1993.*

[16] *K. Koperski, J. Han.* Discovery of Spatial Association Rules in Geographic Information Databases. *SSD: 47-66, 1995.*

[17] *A. Cliff, J. Ord.* Spatial Autocorrelation. *London, Pion, 1973.*

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

[19] *J. LeSage.* Regression Analysis of Spatial data. *The Journal of Regional Analysis and Policy, JRAP, 27, 2: 83-94,1997.*

[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, 101:139-152, 1997.*

[21] *S.Ozesmi, U.Ozesmi.* An Artificial neural network approach to spatial habitat modeling with interspecific interaction. *Ecological Modeling, 116:15-31,1999.*

[22] *D. Papadias., N. Karacapilidis, N. Arkoumanis.* Processing Fuzzy Spatial Queries: A Configuration Similarity Approach. *International Journal of Geographic Information Science Vol. 13(2), 93-128, 1999.*

[23] *G. Karypis and V. Kumar.* A Parallel Algorithm for Multilevel Graph partitioning and Sparse Matrix Ordering. *Journal of Parallel and Distributed Computing, Vol. 48, pp. 71–95, 1998.*

[24] *S. Openshaw, P. Taylor.* A million or so correlation coefficients: three experiments on the modifiable area unit problem. *Statistical Applications in the spatial sciences. London, Pion: 127- 44, 1979.*

[25] *H.C. Huang, H.-C., N. Cressie.* Empirical Bayesian spatial prediction using wavelets. In Bayesian Inference in Wavelet Based Model. *Lecture Notes in Statistics, 141, Springer-Verlag, New York, 203-222, 1999.*

[26] *N. Radhakrishnan.* Private Conversation between the authors and the Director, CIC Directorate-ARL *.Workshop on Mining Scientific Datasets. Army HPC Research Center, September 1999.*

[27] *The Nature Conservancy.* DoD Commander's Guide to Biodiversity *April, 1996.*

[28] *R. Namburu.* Data Mining Issues in Scientific Simulation. *Workshop on Mining Scientific Datasets. Army HPC Research Center, September 1999.*

[29] *S. Shekhar, M. Coyle, D-R. Liu, b. Goyal, and S. Sarkar.* Data Models in Geographic Information Systems. *Communication of the ACM, 40(4), 1997.*

[30] *S.Shekhar, X. Liu, S.Chawla.* Modeling Direction as a spatial object. *To appear in GeoInformatica.*

[31] *S. Shekhar and D-R. Liu.* CCAM: A connectivity-Clustered Access Method for Aggregate Queries on Transportation Networks-A Summary of Results. *IEEE Transactions on Knowledge and Data Engineering, 9(1), January 1997.*

[32] *S. Shekhar and B. Amin.* Generalization by Neural Networks . *IEEE Trans. On Knowledge and Data Eng. (April), 4(2), 1992.*

[33] *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.*

[34] *S. Shekhar, Andrew Fetterer, and Brajesh Goyal.* Materialization Trade-Offs in Hierarchical Shortest Path Algorithms. *In Proc. Symposium on Large Spatial Database, 1997.*

[35] *S. Shekhar, T. A. Yang, and P. Hancock.* An Intelligent Vehicle Highway Information Management System. *Intl Jr. on Microcomputers in Civil Engineering (ISSN 0885-9507), 8(3), 1993.*

[36] *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.*

[37] *S.Shekhar, X.Liu, S.Chawla.* Battlefield Queries with Object-relational SQL. *AHPCRC Bulletin, Volume 9, No.1-2, 1999.*

[38] *M. Dang, G. Govaert.* Spatial Fuzzy Clustering using EM and Markov Random Fields. *Systems Research and Information Systems, Vol. 8, pp. 183-202, 1998.*

[39] *D.Griffith.* Statistical and mathematical sources of regional science theory: Map pattern analysis as an example. *Papers in Regional Science, 78, 21-45, 1999.*

[40] *C. Ambroise, M.Dang, G.Govaert.* Clustering of spatial data by the EM algorithm. *geoEnV I-Geostatistics for Environmental applications, A. Soares, J.G. Hernandez and R. Froidevaux(eds), 493-504. Quantitative Geology and Geostatistics, vol 9, 1996.*

[41] *J.P. LeSage.* Bayesian Estimation of spatial autoregressive models. *International Regional Science Review, 20, 113-129, 1997.*

[42] *D.P. McMillen.* Probit with spatial autocorrelation., *Journal of Regional Science,32, 335-348, 1992.*

[43] *T.Zhang, R.Ramakrishnan and Miron Livly,* BIRCH: An Efficient Data Clustering Method for Very Large Databases. *Proc. Annual SIGMOD Conf., 1996.*

[44] *J. Chiles and P. Delfiner,* Geostatistics: Modeling Spatial Uncertainty. *Wiley Series in Probability and Statistics, 1999.*

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

[46] *J. Hoeting, M. Leecaster, D.Bowden.* An improved model for spatially correlated binary responses. *Technical Report 9719, Department of Statistics, Colorado State University, 1999.*

[47] *R. Pace and R. Barry.* Sparse spatial autoregressions. *Statistics and Probability Letters, 33, 291-297, 1997.*

7 Appendix: Clustering of spatial data

Clustering is another well known data mining technique for deriving information from large data sets. It is convenient(at least initially) to frame the clustering problem in a multi-dimensional attribute space. Given n data objects described in terms of m variables each object can be represented as point in a m-dimensional space. The clustering problem is then to *identify high density groups of points from a set of non-uniformly distributed points.* For example, we would like to cluster the marsh grid locations on the basis of the attributes described above. Since the objects in this case are spatially referenced we will slightly modify our clustering objective. Namely, we would like to partition n data points into k clusters such that [38]

1. Each cluster is as homogeneous as possible.

2. Two data points which are geographically close to each other have a greater probability of belonging to the same cluster than those that are far apart.

7.1 Clustering, Mixture analysis and the EM algorithm

In the statistics literature the clustering problem is often recast in terms of *mixture models*. In a mixture model the data is assumed to be generated by a series of probability distributions where each distribution generates one cluster. The goal then is to identify the parameters of each probability distribution and their weights in the overall mixture distribution.

For example, if we assume that each cluster is governed by an m-dimensional Gaussian distribution and they are K clusters then,

$$P(\mathbf{x}|k) = \frac{1}{(2\pi)^m|\Sigma^k|)^{\frac{1}{2}}} exp(\frac{1}{2}(\mathbf{x}-\mu_\mathbf{k})^\mathbf{T}(\mathbf{\Sigma^k})^{-\mathbf{1}}(\mathbf{x}-\mu_\mathbf{k})),$$

where $k = 1, \ldots K$, u^k is the m-dimensional mean of cluster k and Σ^k is the covariance matrix. The mixture model probability function is

$$P(\mathbf{x}) = \sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{w_k P(x|k)}.$$

The coefficients w_k represent the fraction of the data set represented by the *kth* cluster. Using Bayes theorem the probability that a given data point $\mathbf{x}$ belongs to the *kth* cluster is

$$P(k|\mathbf{x}) = \frac{\mathbf{w_k P(x|k)}}{\mathbf{P(x)}}.$$

The parameters of the mixture model: u_k, Σ^k and w_k for $k = 1, \ldots, K$, may be calculated using the Expectation-Maximization(EM) algorithm. The steps of the algorithm are [4]:

1. Guess the initial model parameters: u_k^0, Σ_k^0 and w_k^0 for $k = 1, \ldots, K$.

2. At each iteration j and for each data object $\mathbf{x}$ calculate the probability that the record belongs to cluster k for $k = 1, \ldots, K$:

$$P(k|\mathbf{x}) = \frac{\mathbf{w_k^j P^j(x|k)}}{\mathbf{P^j(x)}}$$

3. Update the mixture parameters on the basis of the new estimate:

$$w_k^{j+1} = \frac{1}{n} \sum_{x \in D} P(k|\mathbf{x})$$

$$\mu_k^{j+1} = \frac{sum_{x \in D}\mathbf{xP(k|x)}}{sum_{x \in D} P(k|\mathbf{x})}$$

$$\Sigma_k^{j+1} = \frac{\sum_{x \in D} P(k|\mathbf{x})(\mathbf{x}-\mu_\mathbf{k}^\mathbf{j+1})(\mathbf{x}-\mu_\mathbf{k}^\mathbf{j+1})^\mathbf{T}}{\sum_{x \in D} P(k|\mathbf{x})}$$

4. Compute the log estimate $E_k = \sum_{x \in D} log(P^k(\mathbf{x}))$. If for some fixed stopping criterion ϵ, $|E_k - E_{k+1}| \leq \epsilon$, then stop, else set $k = k + 1$.

7.2 Neighborhood EM Algorithm

In [40] it was shown that the an equivalent interpretation of the EM algorithm could be extended to account for spatial proximity effects. The EM algorithm for mixture models is equivalent to the maximization of the the following criterion

$$D(c, \mu_k, \Sigma_k) = \sum_{k=1}^{K} \sum_{i=1}^{n} c_{ik} \log(w_k P^k(\mathbf{x_i}|\mathbf{k})) - \sum_{\mathbf{k=1}}^{\mathbf{K}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{c_{ik} log(c_{ik})}$$

where $\mathbf{c} = \mathbf{c_{ik}, i = 1, \ldots, n}$ and $k = 1, \ldots K$ define a fuzzy classification representing the grade of membership of data point $\mathbf{x_i}$ into cluster k. The c_{ik}'s satisfy the constraints $(0 < c_{ik} < 1,\ \sum_{k=1}^{K} c_{ik} = 1, \sum_{i=1}^{n} c_{ik} > 0)$.

Ambroise et. al [40] penalized the objective function $D(c, \mu_k, \Sigma_k)$ with the term

$$G(\mathbf{c}) = \mathbf{\frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ik} c_{jk} w_{ij}}$$

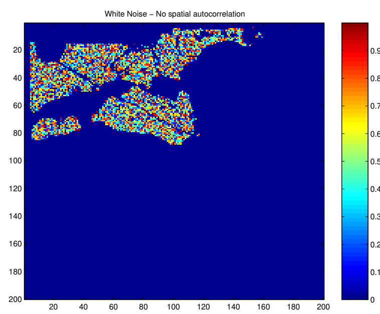
where $W = (w_{ij})$ is the contiguity matrix as defined before.

The new "spatially weighted" objective function is

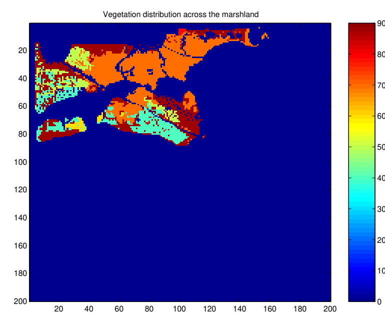
$$U(c, \mu_k, \Sigma_k) = D(c, \mu_k, \Sigma_k) + \beta G(c)$$

where $\beta \geq 0$ is a parameter to control the spatial homogeneity of the data set. Using the new criterion the spatial autocorrelation effects can be incorporated into a clustering algorithm.

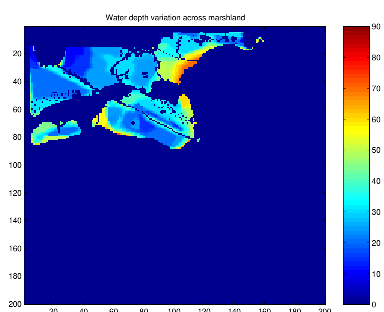
We have carried out experiments using the Neighborhood EM algorithm on the bird data set. We assume two clusters corresponding to the presence/absence of nests. When $\beta = 0$ the NEM reduces to the classical EM algorithm. We varied the β parameters and the results are shown in Figure 8. The results lead us to conclude that including spatial information in the clustering algorithm leads to a dramatic improvement of results(Figure 8(b) compared with Figure 8(a)) but overemphasising spatial information leads to "oversmoothing" and degradation in accuracy.



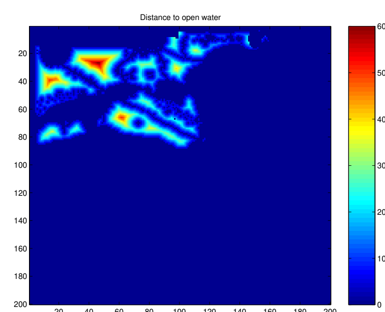
(a) A random distribution



(b) Vegetation

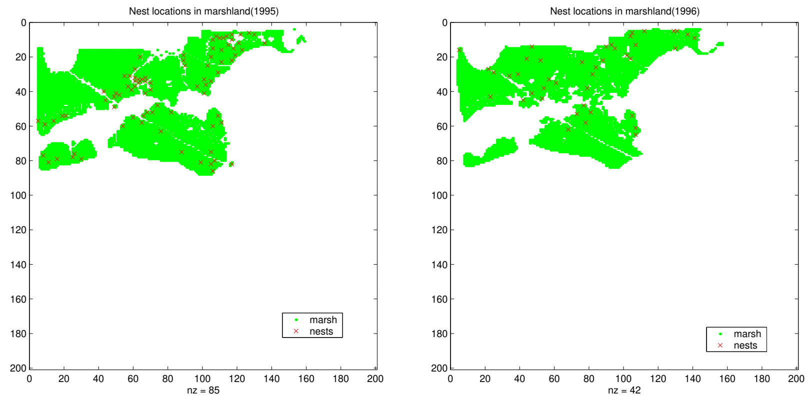


(c) Water Depth



(d) Distance to Open Water

Figure 1: The instrinsic spatial autocorrelation of the attributes is shown in these figures. (a) If the attribute values were distributed randomly then this is how it would appear. (b) Distribution of vegetation values. (c) Distribution of Depth. (d) Distribution of distance to open water.



(a) Learning Data

(b) Test Data

Figure 2: Nest locations in marshland for two consecutive years. The 1995 data will be used to build the model(Learning data) and the 1996 data will be used to validate the model(Test data).

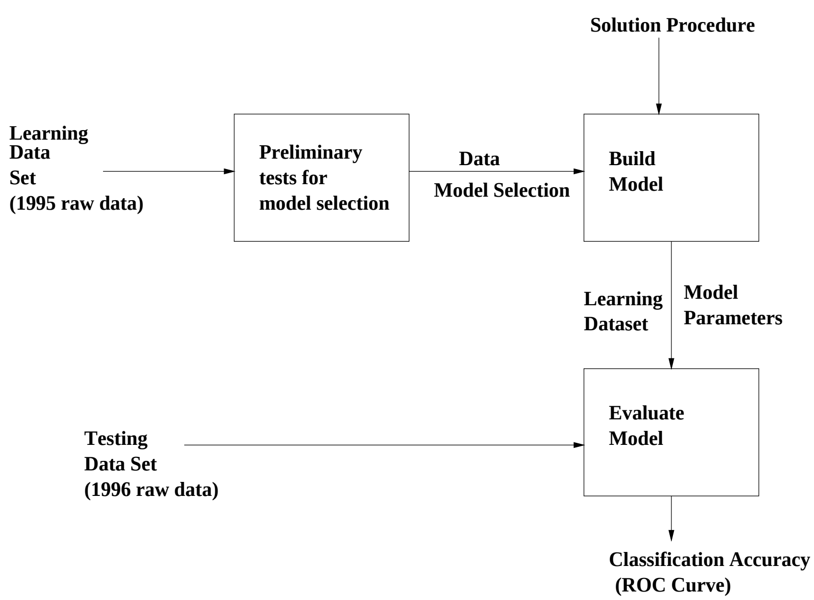
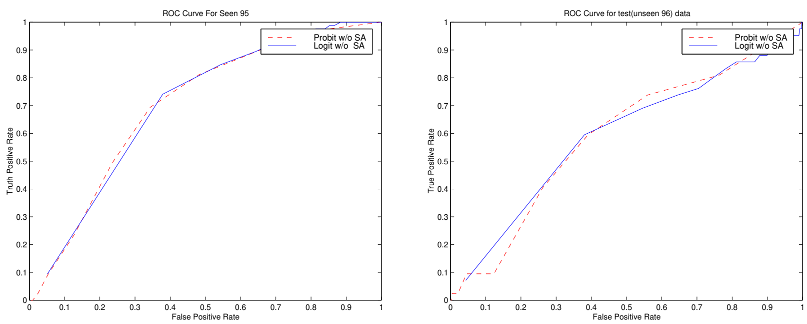


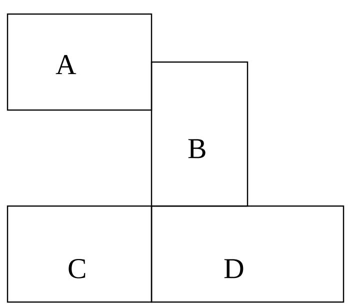
Figure 3: Design of Experiment for classical data mining



(a) Model built with 1995 data

(b) Testing the Model with 1996 data

Figure 4: Simple logistic and Probit model

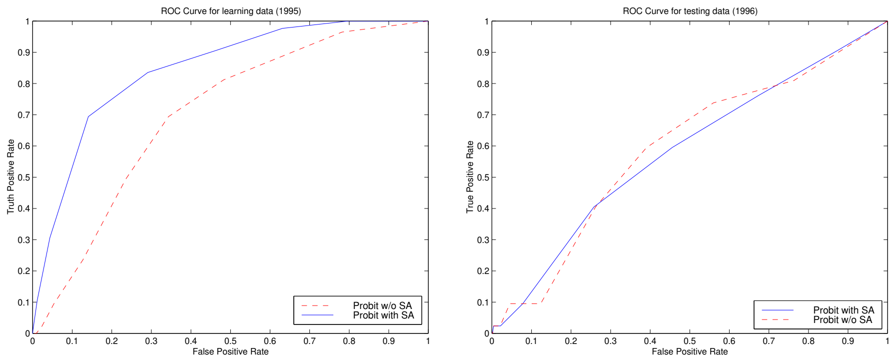


(a)

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

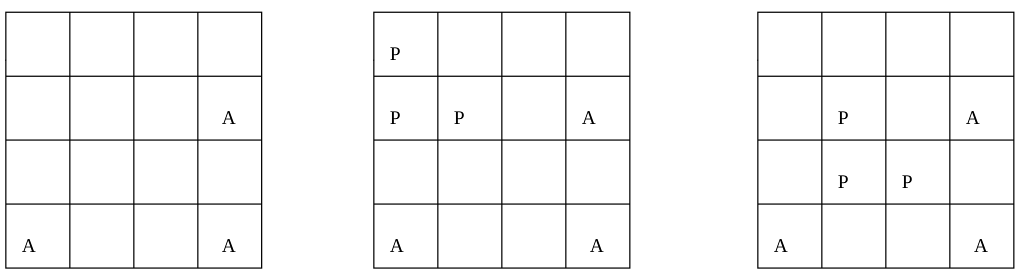
(b)

Figure 5: A spatial neighborhood and its contiguity matrix



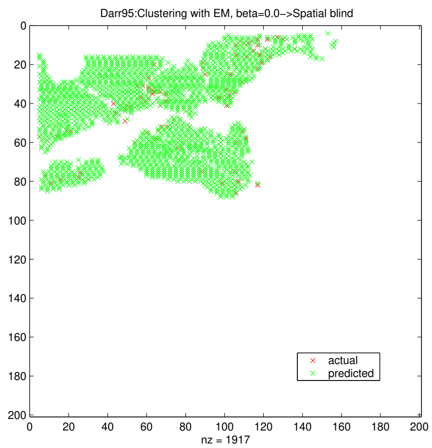
(a) Learning Data (b) Test Data

Figure 6: (a) Comparison of the probit and probit with spatial autocorrelation on the 1995 learning data. (b) Comparision of the two models on the test data. On the learning data there is substantial and systematic improvement for all levels of cut-off probability and on the test data the improvement is statistically insignificant.

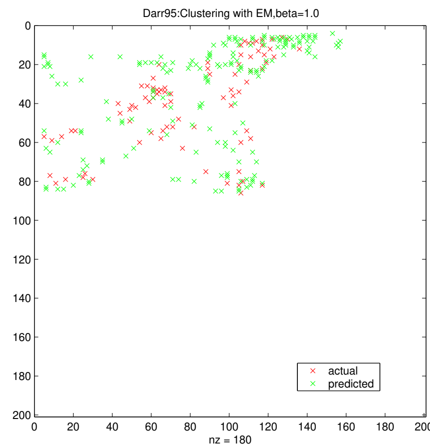


(a) (b) (c)

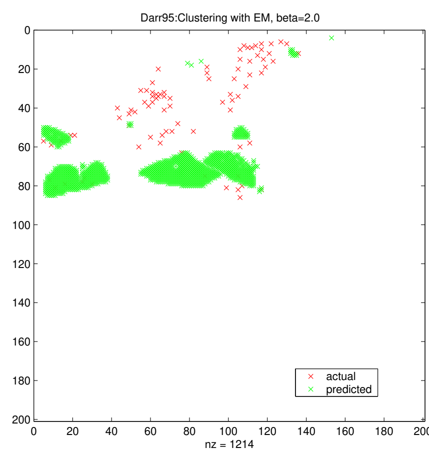
Figure 7: (a) The actual location of nests(A). (b) Locations predicted by a model. (c) Locations predicted by another model. The current measure fail to distinguish between the two models even though (c) is clearly better than (b)



(a) Spatially blind($\beta = 0.0$)



(b) Spatial($\beta = 1.0$)



(c) Spatial($\beta = 2.0$)

Figure 8: (a) As expected clustering without any spatial information leads to poor results. (b) By including spatial information($\beta = 1.0$) leads to dramatic improvement of results. (c) By overemphasising spatial information($\beta = 2.0$) again leads to poor results.