Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 00-001

Modeling Spatial Dependencies for Mining Geospatial Data: A Statistical Approach

Sanjay Chawla, Shashi Shekhar, and Wei Li Wu

January 10, 2000

Modeling Spatial Dependencies for Mining Geospatial Data :A Statistical Approach *†

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

January 7, 2000

Contents

1 Introduction **4**
1.1 Geo-Spatial Data Mining 4
1.2 Classical regression and its limitations 6
1.3 What is special about Geo-Spatial Data Mining 6
1.4 The Diaper-Beer analogue in spatial data mining 7

2 Modeling Spatial Dependencies **8**
2.1 Framework 8
2.2 Spatial Autocorrelation 8
2.3 Spatial Error and Autoregressive Regression Models 9
2.3.1 Solution Procedure 10
2.4 Spatially correlated binary data 11

3 Preliminary Results **12**
3.1 Spatial Autoregression for binary response data 14

4 Clustering of spatial data **15**
4.1 Clustering, Mixture analysis and the EM algorithm 15
4.2 Neighborhood EM Algorithm 16

* *Keywords*: spatial data mining, spatial autoregression, autocorrelation, EM algorithm, regression, probit, logit, contiguity matrix, Gibbs sampling, Bayesian regression,clustering, mixture model

† Computer Science Technical Report:00-001

5 Future Work **17**
5.1 Spatial Dependence Modeling: Defining an appropriate Neighborhood Matrix 17
5.2 Scaling up to High Performance algorithms for large Data Sets: Exploiting the sparse matrix structure 17
5.3 The sensitivity of spatial autocorrelation measures to scale and neighborhood choice 18
5.4 Evaluation of methods in different application domains . . . 19

6 Example application domains for spatial data mining **19**

7 Conclusion **20**

8 Acknowledgements **21**

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 overview statistical techniques which can effectively model the notion of spatial-autocorrelation. We will also present a "roadmap" for extending classical data mining techniques to manage geo-spatial 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 a marshes. We will show that by taking spatial autocorrelation into account the accuracy of the results show a substantial improvement. Finally we will chart a "road map" for further research in spatial data mining.

1.1 Geo-Spatial Data Mining

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, 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. For example, a spatial distribution may be classified into random or clustered using nearest neighbor index or quadrant analysis.

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 for all three methods(descriptive, explanatory and predictive) is to materialize spatial relationships as attributes and rebuild the model with the "new" spatial attributes [13]. This approach has fundamental shortcomings, as we will illustrate in the case of regression.

1.2 Classical regression and its limitations

Given a vector of n observations and a matrix X of explanatory variables the classical linear regression model encapsulates their interrelationship using the standard linear equation

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

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

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

or, in a more compact vector form

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

For spatial data and observations the assumptions on the residual errors, the ϵ_i's which are inherited by the y_i's may not be valid. The residual error may have systematic variation over space, i.e., exhibit high spatial autocorrelation. Ignoring the effects of spatial autocorrelation may yield biased and inconsistent estimate of the relationship between $\bar{y}$ and $\tilde{X}$. In particular we may not be able to assume that the y_i's are independent of each other. This property is called *spatial autocorrelation* and though similar to the concept of time autocorrelation in time series analysis it more difficult to model because of the multi-dimensional nature of space. Ignoring spatial autocorrelation may yield biased and inconsistent estimate of the relationship between $\vec{y}$ and $\tilde{X}$. In addition it may lead to a poor fit of the model.

1.3 What is special about Geo-Spatial Data Mining

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.

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].

1.4 The Diaper-Beer analogue in spatial data mining

Data mining has received a lot of attention in the general media thanks mainly to the infamous "Diaper-Beer" example. Using data mining techniques, mainly association rules, researchers at a giant 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 the drinking water.

2 Modeling Spatial Dependencies

We will now show how spatial dependencies are modeled in the framework of regression analysis. This will serve as a prototype for our proposed approach towards other data mining techniques.

2.1 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.

2.2 Spatial Autocorrelation

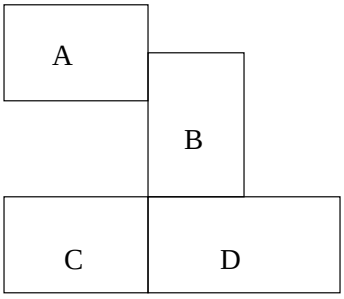
There are many measures available for quantifying spatial autocorrelation. Each have their own strengths and weaknesses. The two most well known measures are Moran's I and Geary's C measure. Here we will briefly describe the Moran I measure and refer the reader to standard books on spatial statistics [17] for a description of the Geary's C measure.

The Moran's I measure (henceforth MI) ranges between -1 and +1. 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

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

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. 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. The two types of neighborhood are shown in Figure 1. 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.



(a)

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

(b)

Figure 1: A spatial neighborhood and its contiguity matrix

2.3 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

$$y = X\beta + u \quad (1)$$

$$y = \rho W u + \epsilon \quad (2)$$

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 would 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.

2.3.1 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 it will be useful to extend previous to speed up the solution procedure.

2.4 Spatially correlated binary data

For the situation where the dependent variable y is binary$(0, 1)$, the conventional Spatial Autoregressive Regression(SAR) cannot be used for the following two reasons [19]:

1. In the SAR 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 $\rho W y + X\beta$ when $y = 0$ and $1 - \rho W y - X\beta$ when $y = 1$.

2. There is no way to guarentee that the predicted values from the SAR 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

$$\lim_{\rho W y + X\beta \to +\infty} Prob(y = 1) = 1 \quad (3)$$

$$\lim_{\rho W y + X\beta \to -\infty} Prob(y = 1) = 0 \quad (4)$$

$$(5)$$

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

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

and for the probit case the distribution used is

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

3 Preliminary Results

The current approach towards solving spatial data mining problems is to "materialize" spatial relationships into variables(e.g. distance to water) and then use a relevant classical data mining technique with "extra" spatial explanatory variable. For example, consider the following scenario: We are given a database of homes in and around a major metropolis. The database includes the price of the home (the dependent variable for this example) and some other explanatory variables like the age, size, address and type of the home. If we build a regression model based on the database and use the model to predict the value of homes given the values of the explanatory variables then our model may not be accurate. Part of the inaccuracy may be due to the fact that we are not taking into account the spatial information latent in the explanatory variables. For instance, given the address of a home we can easily calculate how far each home is from the metropolis downtown, parks, lakes, public transportation etc. to check if these have an influence on the price of a home. The current approach in data mining is to encapsulate the distance information into new variables, e.g., distance to downtown and then rebuild the model with this extra explanatory variable. Unfortunately this approach has many shortcomings: * How does one decide which spatial information is important? In the above example, as we mentioned, it is common knowledge that distance to downtown influences house prices. But so does adjacency to a golf course and orientation of the house (A house facing north in the higher latitudes of the northern hemisphere is clearly less preferable to a similar house facing south). It is clearly computationally expensive if not intractable to "materialize" all the possible spatial relationships and incorporate them into the model. They are simply too many possibilities and even then the computational cost of computing spatial relationships is too high. * Lack of familiarity with the application domain: If the data mining analyst is not familiar with the application domain then even "materializing" spatial relationships may not be possible because of the lack of awareness of potentially useful spatial relationships.

In contrast our approach follows the two-fold path: 1. We provide simple tests which determine if spatial autocorrelation play an important role in the problem at hand. 2. If the tests lead us to conclude that spatial

autocorrelation are important then we apply a spatially generalized data mining tool which can account for spatial information with resorting to the "materialization" of spatial relationships.

To validate our approach we chose an application in the domain of "conservation ecology". Regression analysis [20] and neural networks [21] was used to build a spatial habitat model for the marsh nesting bird species Red Wing Blackbird. The goal was to predict the nest locations (the dependent variable) based on the following explanatory variables: Distance to the edge of the marsh, Distance to open water, Water depth, Vegetation height, Stem Density and Vegetation Durability. The domain knowledge of our collaborators motivated them to consider two spatial variables: Distance to the edge of the marsh and Distance to open water as explanatory variables. As we have mentioned for a data mining analyst not familiar with the application the choice of these spatial variables is clearly non-intuitive and not obvious. The spatial geometry of the marshland and the locations of the nests is shown in Figure 2.

We have chosen this study to highlight our approach because of the following 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. The authors have conducted extensive field surveys and have collected data which spans a period of three years (Extensive "ground truthing").

3. Models based on classical regression have not performed well despite incorporation of significant domain knowledge in selection of variables by our collaborators.. The success rate in predicting nest locations was 20

4. Many of the explanatory variables exhibit high spatial autocorrelation suggesting that a spatial regression model may be lead to better results.

We have computed the Moran I coefficient of the explanatory variables(Table 1) . An image of the variable is also shown below. It is clear that all the variables(except vegetation durability, which incidentally was considered accurate by the authors of the study) show a relatively high spatial autocorrelation. It is also clear that from the images of the explanatory variables an intuitive sense of spatial autocorrelation can be deduced. Thus

we believe that the above candidate application is a strong candidate for the use of spatial regression model.

Explanatory Variable	Four Neighbor	Eight Neighbor
Distance to Edge	0.7606	0.9032
Distance to Open Water	0.7342	0.8022
Depth	0.6476	0.7408
Vegetation Height	0.7742	0.8149
Stem Density	0.6267	0.7653
Vegetation Durability	0.3322	0.4851

Table 1: Moran I coefficient of explanatory variables to predict nest locations for the Red Winged Blackbird

3.1 Spatial Autoregression for binary response data

We also conducted experiments using the probit spatial autoregressive model. It is well known that the likelihood function for the parameters for the probit model does not have a closed form solution. We used the well known Gibbs sampling methodology to obtain the value of the parameters ρ and β.

We compared the effects of including a spatial autogressive term with the plain "vanilla" probit model. The comparison was done on the basis of Receiver operating characteristic(ROC) curves. ROC curve reveal the relationship between the true positive rate(TPR) and the false positive rate(FPR). The TPR and the FPR are defined as

$$TPR = \frac{\sum I_{[yp>b]} I[y_i = 1]}{\sum y_i} \tag{6}$$

$$FPR = \frac{\sum I_{[yp>b]} I[y_i = 0]}{\sum (1 - y_i)} \tag{7}$$

$$\tag{8}$$

where $0 \leq b \leq 1$. TPR measures the ratio of the number of sites where the nest is actually located and the nest is predicted above a threshold probability value b versus the number of actual nest sites. The FPR measures the ratio where the nest were absent but were predicted at the given threshold versus the number of sites where the nests were absent. The results of plot of TPR vs. FPR is shown in Figure 3.

4 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.

4.1 Clustering, Mixture analysis and the EM algorithm

In the statistics literature the clustering problem is often recasted 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,

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

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:

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

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

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

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 $k = k + 1$.

4.2 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 $\mathbf{c = c_{ik}, i = 1, \ldots, n}$ 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

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

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 4, 5 and Figure 6.

5 Future Work

5.1 Spatial Dependence Modeling: Defining an appropriate Neighborhood Matrix

The standard contiguity matrix is based on the notion of adjacency. We want to investigate how other spatial relationships based on direction, topology and distance [22] can be incorporated in the contiguity matrix. At the modeling level the obvious way is to either have a separate contiguity matrix for each spatial relationship or have a generalized contiguity matrix which combines the effect of all categories of spatial relationships. Even when the choice of the spatial relationship for the design of the contiguity matrix is straightforward, the issue of implementation is non-obvious. For example, in the case of the bird habitat model the underlying geometry of the marshes is fragmented because of the many water channels which fissure through them. Should the channels be considered as open water? If the distance criterion is used, should nest locations which are close but separated by a channel be considered as neighbors? These are some of the issues that we propose to investigate regarding the design of the contiguity matrix.

5.2 Scaling up to High Performance algorithms for large Data Sets: Exploiting the sparse matrix structure

Many applications of spatial data mining have very large data sets. If sampling is not an option then traditional statistical techniques become computationally expensive for large data sets. One of the goals of data mining is to develop computationally efficient algorithms which can scale up to large

sets. For spatial data mining the problem of large data sets is compounded because of the neighborhood influence. The algorithm has to make multiple passes over the data set in order to fully account for the effects of neighboring objects. Thus design of efficient and scalable algorithms is a key area of research in spatial data mining.

Our research will focus on the following key tasks: * We have identified the spatial autocorrelation as the key parameter in spatial data mining. Thus it is imperative that efficient algorithms for computing the spatial autocorrelation are part of any spatial data mining "toolbox". Furthermore, in dynamic spatial processes(for example, time dependent partial differential equations) methods for incrementally computing the spatial autocorrelation are needed. At the moment we are not aware of any methods which compute the spatial autocorrelation in an incremental fashion. * When dependent variables are assumed to be spatially related then one has to take recourse to approximate algorithms to solve for the model parameters. These algorithms are similar to those used for the solution of the eigenvector problem for matrices. Efficient algorithms which exploit the sparsity of the contiguity matrix can be formulated. We have already developed algorithms based on graph partitioning which can efficiently solve the eigenvector problem for large and sparse matrices [23].

5.3 The sensitivity of spatial autocorrelation measures to scale and neighborhood choice

It is well known [24] 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. This problem is clearly related to the choice of discretization of numerical attributes into categorical labels.

Our tests to determine whether spatial data mining tools are relevant to the application are dependent on spatial autocorrelation tests and our sensitive to scale and neighborhood choice. This is clearly illustrated in Table 1 where the choice of four neighborhood and eight neighborhood can lead to different (though not contradictory) results. We will thoroughly investigate the effects of scale and neighborhood choice on different measures of spatial autocorrelation.

A recent theoretical study [25] in spatial statistics has used wavelets to investigate multiscale spatial phenomenon. We will investigate this study as

it is known that use of wavelets does tend to ameliorate the effects of scale to a certain extent.

5.4 Evaluation of methods in different application domains

In order to validate our model selection and demonstrate the effectiveness of algorithms we plan to carry out extensive experimentation on a wide variety of large geo-spatial data sets. Our design of experiment will be structured along the following tasks:

- The development of a benchmark set of applications, data sets and patterns. Some of these are discussed below.
- Experiment design by selection of candidates (e.g. proposed techniques and its competitors), comparison metrics (e.g. performance, classification accuracy, etc.), and independent variables (e.g. data-set size, presence of noise, variety of data types, etc.).
- Instrumentation and measurements: Analysis of the data gathered to characterize dominance zones of proposed techniques.

6 Example application domains for spatial data mining

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 is an area where spatial analysis is extensively used. We would like to validate our research on available crime data sets.

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.

7 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 also showed how the mixture model can be used for the clustering spatial data. In particular the objective function of the EM algorithm can be penalized to incorporate a spatial term. 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 also laid out a roadmap for future research and listed a set of example applications which can potentially benefit from spatial data mining.

8 Acknowledgements

We would like to thank James Lesage(http://www.econ.utoledo.edu/lesage) and M. Dang(http://www.hds.utc.fr/mdang) for making their software available on the web.

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] flur-97 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.*

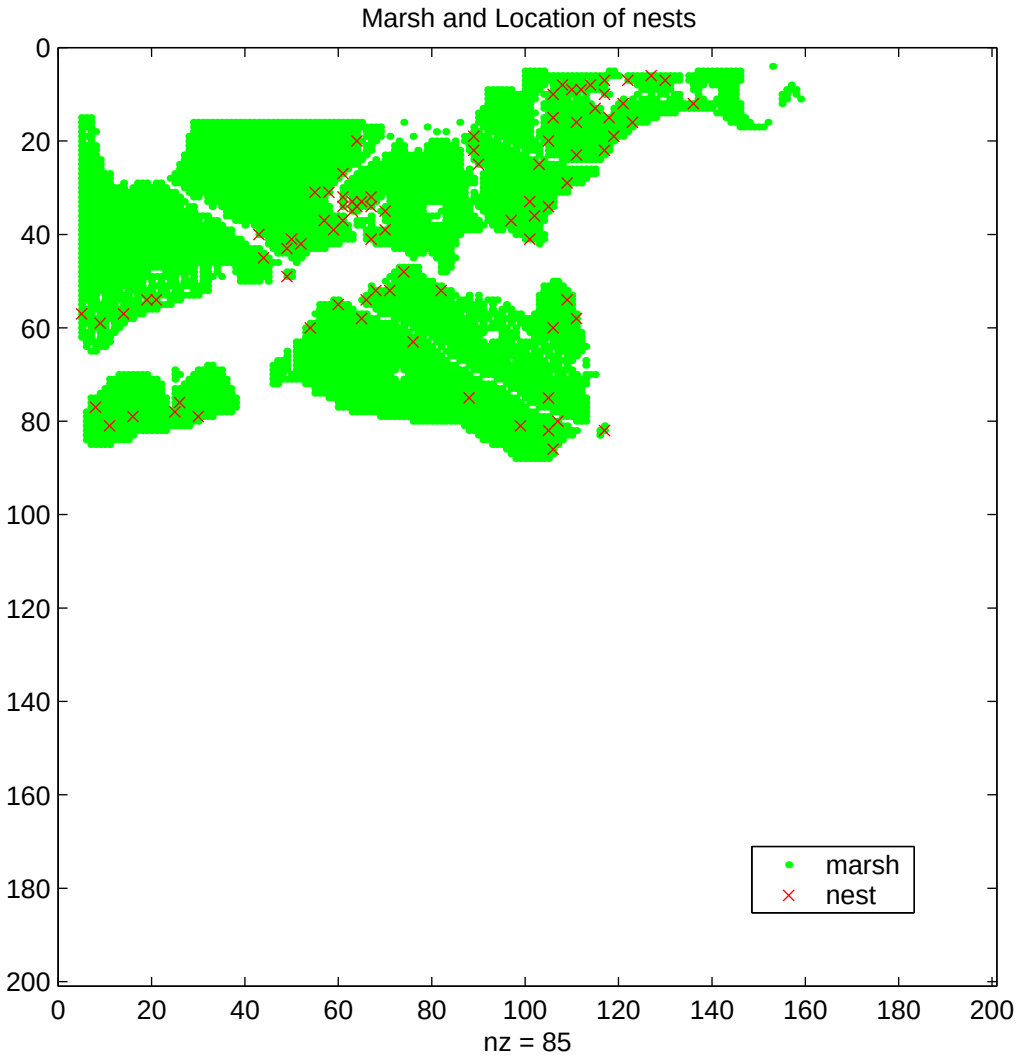


Figure 2: The marshland and the location of nests

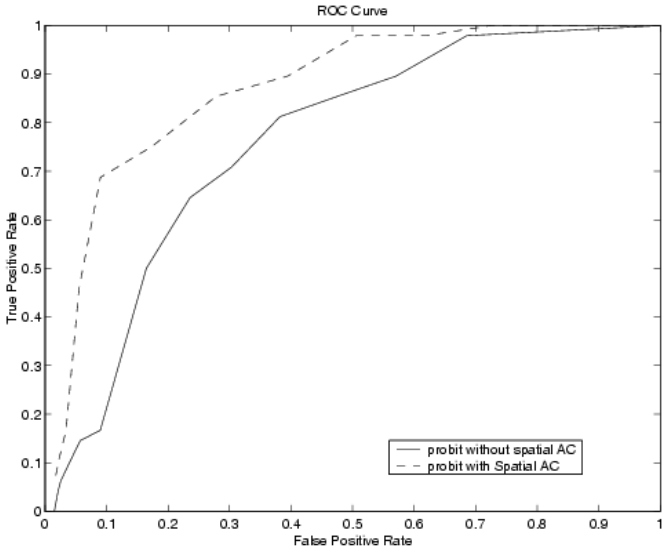


Figure 3: The ROC curves for the probit and the probit with spatial autoregression. The further the curve away from the line TPR=FPR, the better the prediction power of the model. The spatial probit model is better than the probit model at all cut-off probabilities.

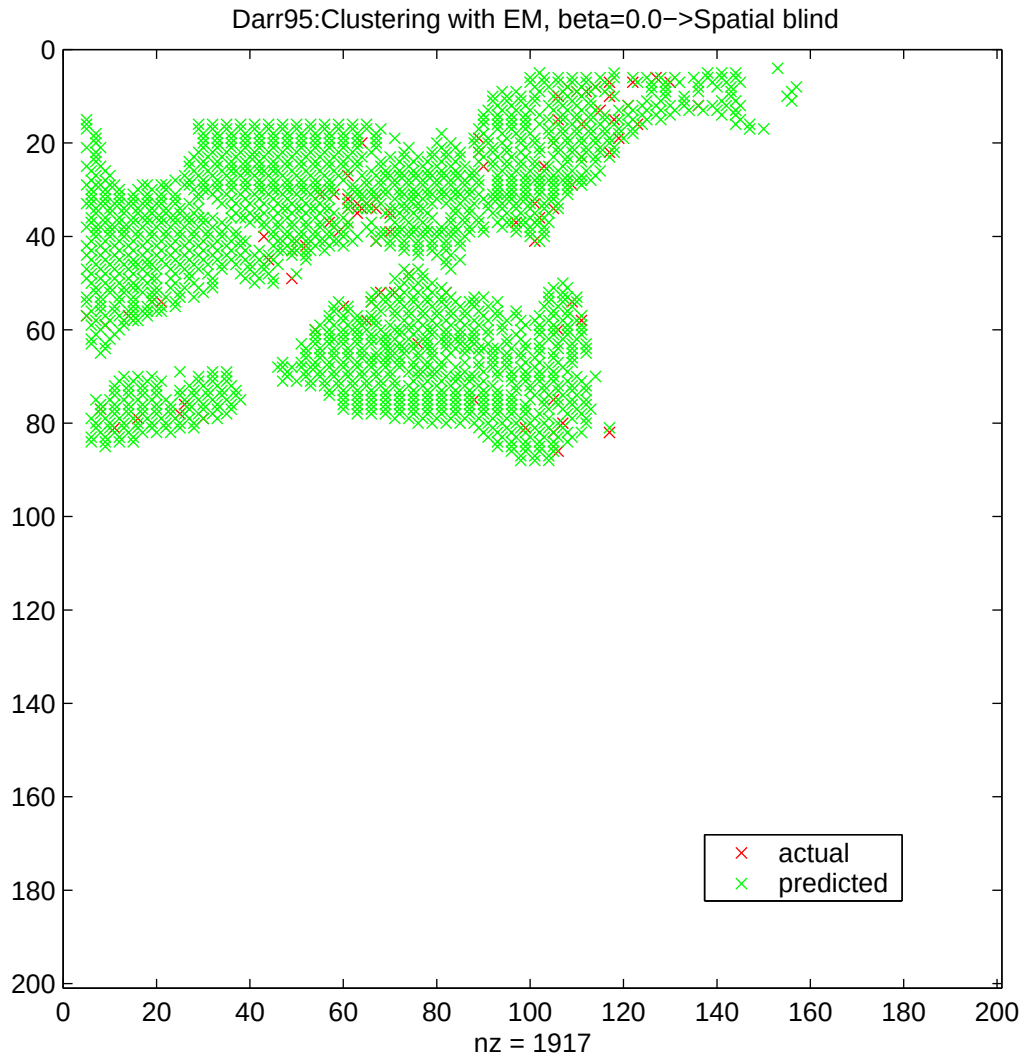


Figure 4: The classical EM algorithm to detect the location of nests. As expected it performs poorly because spatial locational information has been excluded. The name of the marshland is Darr and the data was collected in 1995(Darr95).

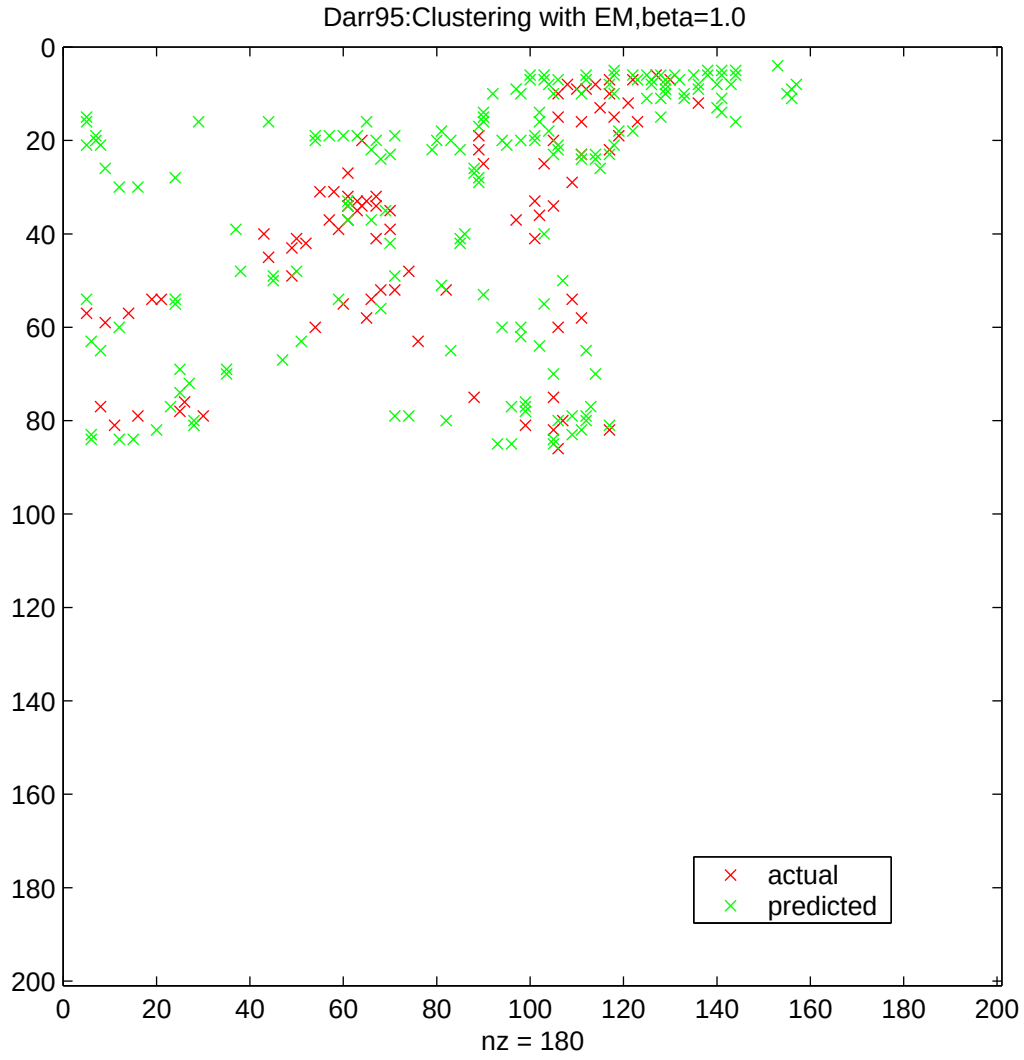


Figure 5: By including spatial information we get a dramatic improvement of results, even better that than the SAR model. Remember the points have been projected onto the 2-dimensional plane. The clusters occur in a conceptual multi-dimensional space. We have chosen the parameter $\beta = 1$. This seems to be a good compromise between the attribute space and the geographic space.

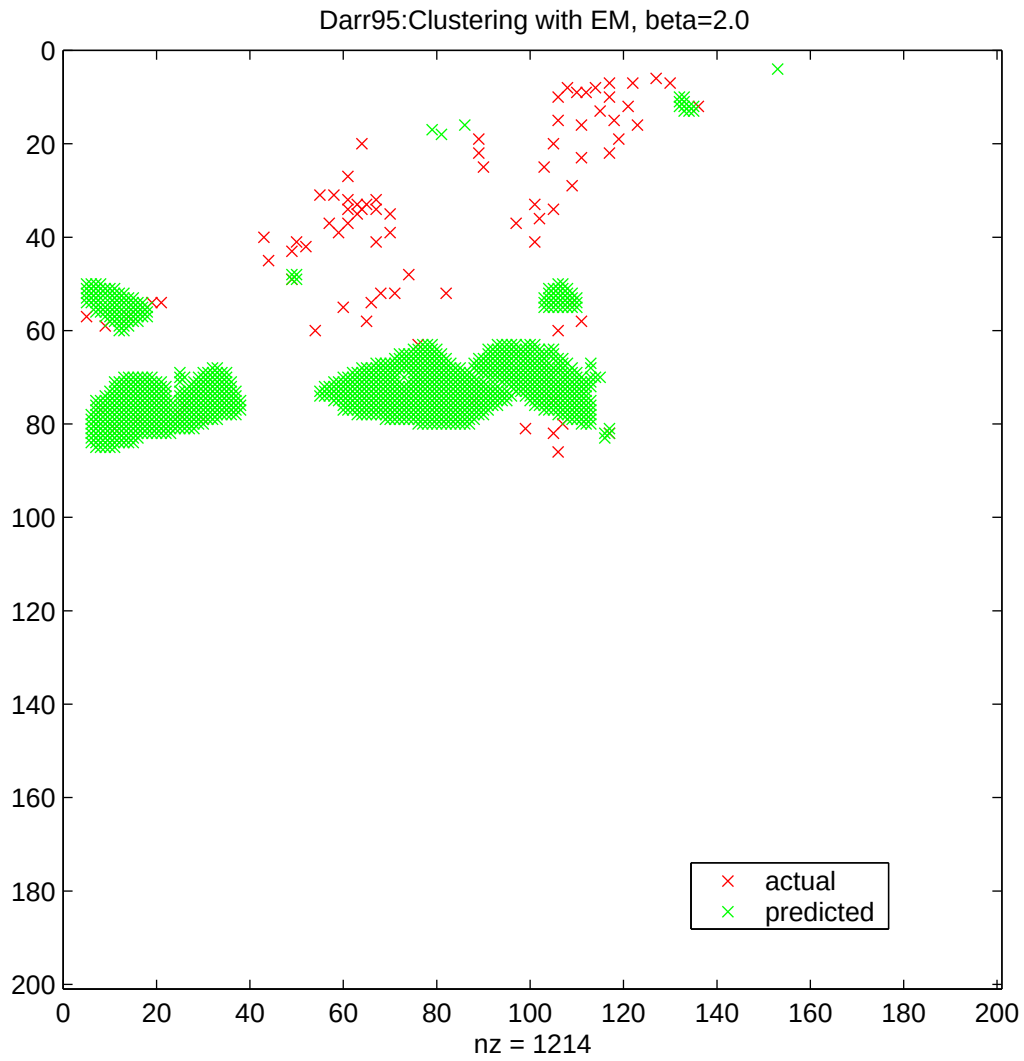


Figure 6: Increasing β is equivalent to increasing the spatial homogeneity effect. This clumps the clusters in the spatial plane and reduces accuracy.