Spatial Data Mining: An Emerging Tool for Policy Makers

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

April 19, 2000

1 Introduction

Widespread use of spatial databases(Shekhar, 1999) is leading to an increasing interest in *mining* 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 are being sucessfully used as tools in decision-making and planning both in the public and private sector. 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(Mark 1999). A recent article in the New York Times(January 20th, 2000) on spatial data mining(SDM) is an indicator that interest in this technology has permeated into the wider public domain.

Data mining is the process of discovering potentially interesting and useful *patterns* of information embedded in large databases. A pattern can be a summary statistic, like the mean, median standard deviation of a probability distribution or a correlation rule. A well publicized pattern, which has now become part of data mining lore, was discovered in the transaction database of national retailer: *People who buy diapers in the afternoon also tend to buy beer.* This was an unexpected and interesting finding which the company put to profitable use by rearranging the store. Thus data mining encompasses a set of techniques to generate hypothesis followed by their validation and verification via standard statistical tools. For example, if the store has a modest 100 items then to check for which two items are correlated or "go together" will require 4950 correlation tests. The promise of data mining is the ability to rapidly and automatically search for *local* and potentially high *utility* patterns using computer algorithms.

The difference between classical and spatial data mining parallels the difference between classical and spatial statistics. One of the fundamental assumptions that guide statistical analysis is that the data samples are independently generated: like successive tosses of coin, or the rolling of a die. When it comes to the analysis of spatial data the assumption about the independence of samples is generally false. Infact spatial data tends to be highly self correlated. For example, people with similar characteristics, occupation and background tend to cluster together in the same neighborhoods. The economies of a region tend to be similiar. Changes in natural resources, wildlife and temperature vary gradually over space. Infact this property of like things to cluster in space is so fundamental that Geographers have elevated it to the status of the first law of geography: *Everything is related to everything else but nearby things are more related than distant things*(Tobler, 1979). In spatial statistics, an area within statistics devoted to the analysis of spatial data, this is called spatial autocorrelation.

In this brief article 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 marshlands based on envioronmental factors. We will also describe how a statistical problem can be transformed into a data mining framework where it can be solved using fast computer algorithms.

2 Spatial Data Mining

Spatial data mining is the search for patterns embedded in large spatial databases. Well known examples of spatial databases are maps, repositories of remote-sensing images and the dicennial census. Infact any dataset which has a spatial, locational or geographic component is an example of a spatial database. Over the years the size of the databases have grown so large that it is important to automate the search for potentially useful patterns. For example, an interesting problem in crime analysis is the detection, explanation and prediction of "hot spots", which are local bursts of high crime activity in a community or city. The current approach towards detection of such spots is for an expert to use a GIS to correlate different map layers of attribute data which are available for that city. The promise of data mining is that it allows the problem to be recast as a search problem in a high dimensional parameter space. Using high speed computers and smart algorithms it is now possible to search this large parameter space for parameters which characterize potential hot spots. Thus the domain expert who earlier searched for hot spots with the aid of a GIS is now involved in setting up the correct problem and interpreting the output from a data mining algorithm.

In fact some very well known examples, of what we now called spatial data mining, occured well before the the invention of computers. For example(Griffith ,1999)

1. In 1855 when the Asiatic cholera was sweeping through London, an epidemiologist marked all locations on a map where the disease had struck and discovered that the locations formed a cluster whose centroid turned out to be a water-pump. When the government authorities turned-off the water pump the cholera began to subside.

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 into one-piece - 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 flouride in the local drinking water supply. Now all municipalities in the United States ensure that all drinking water supply is fortified with flouride.

Spatial data mining holds the promise of discovering similar patterns within existing spatial databases with minimal human intervention.

Spatial data mining can be powerful aid in policy decision making. Infact three diverse areas where spatial data mining is playing an important role were showcased in a recent New York Times article.

Monitoring lending patterns of institutions Consumer advocacy groups are using spatial data mining to map the lending practices of banks and other lending institutions. By relating location of the banks with the demographics of surrounding neighborhood, SDM techniques can be used to determine whether poor neighborhoods are being denied fair access to credit.

Protecting the environment Spatial data mining is being used to design optimal habitat environments for birds listed on the endangered species act. For example, by determining factors which influence a bird's choice of nesting location, conservation managers can ensure that these factors are preserved. We will elaborate on this example in the next section.

Crime mapping and hot-spot analysis Techniques from SDM can be used to detect local patterns in crime databases and examine related databases to search for an explanation. For example, a sudden spurt in crime in a given neihborhood may be the result of an ex-convict moving into the neighborhood.

3 The Spatial Autoregression Model

We now briefly describe how standard statistical techniques can be extended to effectively handle the special properties of spatial data. In order to do that we go back to the first principles of *model building.*

In simplistic terms a *model* is an abstraction of reality. In general there are two approaches to build a model. The first is to understand the "physics" of the phenomenon by determining the relationships between the various components of the observed phenomenon. Often these relationships are cast in strict mathematical terms. For example in transporation applications the gravity model equation quantifies the relationships between two cities based on their distance. Once the model has been ascertained, the data from the phenomenon under observation is used to determine the parameters of the model.

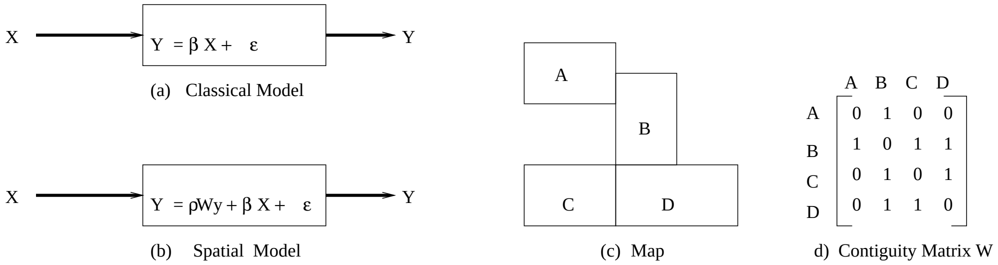
In statistics, data is the primary focus of observation and analysis. A statistical model is supposed to capture *interesting* patterns in the data generated from observing a phenomenon. For example, statisticians often use the assumption that the data is generated from a *normal* distribution irrespective of the underlying behaviour of the phenomenon.

A ubiquitous technique in statistics is regression analysis which is used to predict the value of an independent variable Y on the basis of a dependent variable(s) X. This scenario is shown in Figure 1(a) where the relationship between Y and X is modeled as a linear equation and the error term ϵ is assumed to be independently and identically distributed. This essentially means that errors associated with one sample observation are not dependent on errors associated with other observations and are generated from the same probability distribution.

When classical linear regression is used for modeling phenomenon where the data samples have a spatial location it is often observed that the error term ϵ varies systematically over space. This is because the independent(and dependent variables) themselves are are related in the spatial neighborhood. One way to account for this interdependence is to use a correction term in the model equation as shown in Figure 1(b). The $\rho W y$ term corrects for the spatial autocorrelation present in the dependent variable y. The contiguity matrix W captures the spatial relationship between location where the data samples were collected. For example, an often used assumption in modeling spatial autocorrelation is that y is only dependent on its neighbors. Or in other words the value of the variable y at a given location is conditionally independent given its neighbors. An example of a contiguity matrix for the "map" in Figure 1(c) is shown in Figure 1(d). It is important to notice that the contiguity matrix W is independent of the sample observation values and is only dependent on the spatial relationships between the location of where the data samples were collected.

Once the model has been agreed upon the next step is to determine the parameters of the model. For example in the classical linear regression this amounts to the determination of the parameter vector β. For the spatial model the parameter ρ and β have to be estimated. The classical approach is to use the least square error estimate to derive these parameters.

There are certain limitations of the least square error approach when applied to spatial phenomenon. Consider the example shown in Figure 2. Here the goal is to predict the locations marked A(Figure 2 using regression analysis. A least-square approach will fail to distinguish between the model which predicts the locations shown in Figure 2(c) and another model which predicts locations shown in Figure 2(d). This despite the fact that the predictions in Figure 2(d) are closer to the actual locations than those predicted by Figure 2(c). We have used this observation to design a new framework to solve the two-class spatial classification problem(Chawla,2000).



$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & 1 & 0 & 0 \\ B & 1 & 0 & 1 & 1 \\ C & 0 & 1 & 0 & 1 \\ D & 0 & 1 & 1 & 0 \end{array}$$

(a) Classical Model

(b) Spatial Model

(c) Map

d) Contiguity Matrix W

Figure 1: (a) The classical regression model to determine the relationship between variables Y and X. (b) The model is modified to account for spatial autocorrelation in the dependent variable. (c) An example map with boundaries. (d) The contiguity matrix W of the map shown in (c). A non-zero entry in the matrix records the fact that the corresponding spatial entities on the map are neighbors.

4 An Illustrative Application Domain

The availability of accurate spatial habitat models is an important tool for wildlife management, protection of critical habitat and endangered species. Since the underlying process governing the

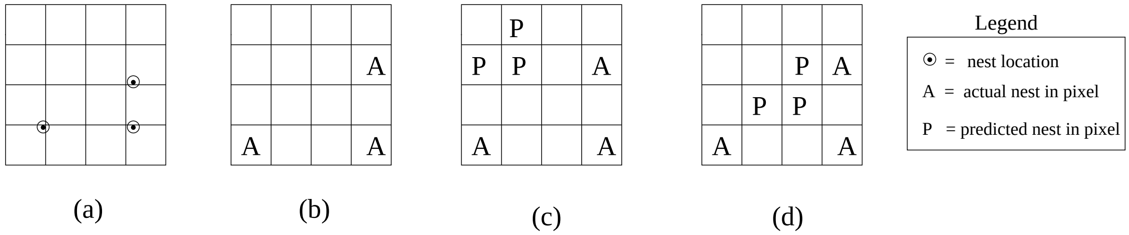


Figure 2: (a)The actual locations of nest, (b)Pixels with actual nests, (c)Location predicted by a model, (d)Location predicted by another mode. Prediction(d) is spatially more accurate than (c). Classical measures of classification accuracy will not capture this distincition.

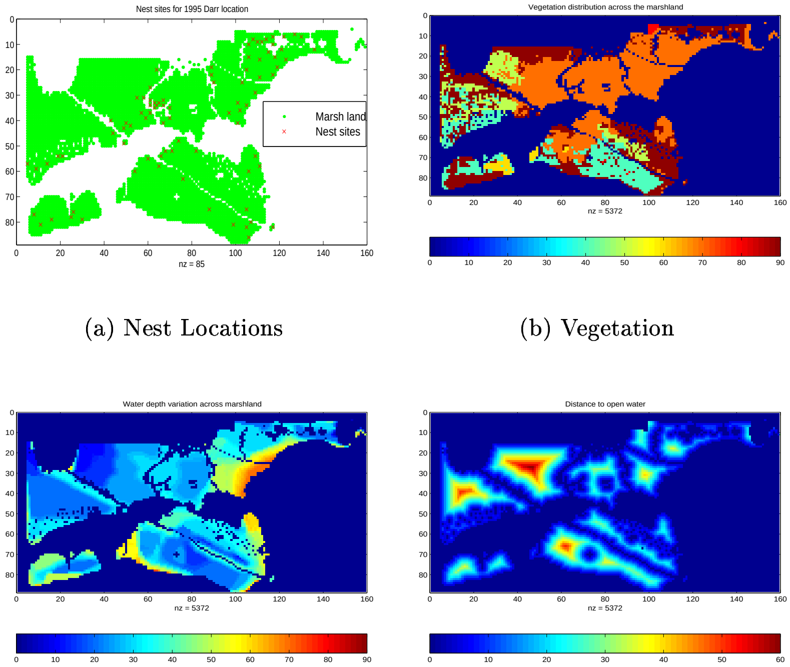
interaction between wildlife and environmental factors is complex, statistical techniques are used to gain insight on the basis of data collected during field work. One of our colleagues(Ozesmi, 1998) has been involved in the development of spatial model for the nesting locations of a marsh-nesting bird species. We will use this application, and the accompanying data, to illustrate the benefits of extending regression analysis with the spatial autocorrelation term.

The data was collected in 1995 and 1996 from two marshlands(Darr and Stubble) located on the shores of Lake Erie in Ohio. A uniform grid was imposed on the marshlands and in each cell the values of several structural and environmental factors were recorded, including *water depth, dominant vegetation durability index and distance to open water.* These three factors play the role of most significant explanatory variables. At each cell was also recorded the fact whether a bird-nest was present or not. The presence of the nest played the role of dependent variable. The geometry of the Darr marshland, locations of the nests and spatial distribution of the explanatory variables are shown in Figure 3. Our colleagues had already applied classical data mining techniques like linear regression and neural networks(non-linear regression) to build spatial habitat models. The model based on linear regression could predict the location of new nests at a 24% rate better than random. The use of neural networks actually decreased the classification accuracy but led to a better understanding of the interaction between the explanatory and the dependent variable.

When the data is mapped it is apparent that both the dependent and independent variables show a moderate to high degree of spatial autocorrelation. For example, Figure 4(a) shows a hypothetical spatial distribution if there was no spatial autocorrelation. It looks like “white noise” as properties of pixel are generated from independent and identical distributions. Note that the maps of explanatory variable in Figure 3 have much more gradual variation indicating high spatial autocorrelation. Figure 3(b) shows a random distribution of nest locations which is quite different from the distribution of actual nests shown in Figure 1(a).

4.1 Experiments Design and Evaluation

Goals: The goal of the experiments was to evaluate the effects of including the spatial autoregressive term, $\rho W\mathbf{y}$, in the regression model. The 1995 Darr marshland data was used as the *learning* set to build the classical and spatial models. Since the dependent variable is binary(nest/no-nest) we used a modified version of the regression equation which predicts the *probability* of a nest being present. This is the standard approach in statistics to deal with binary variables and is called *logistic regression.*


Nest sites for 1995 Darr location
Marsh land
Nest sites
nz = 85
Vegetation distribution across the marshland
nz = 5372
Water depth variation across marshland
nz = 5372
Distance to open water
nz = 5372
0
20
40
60
80
100
120
140
160
10
30
50
70
90

(a) Nest Locations

(b) Vegetation

(c) Water Depth

(d) Distance to Open Water

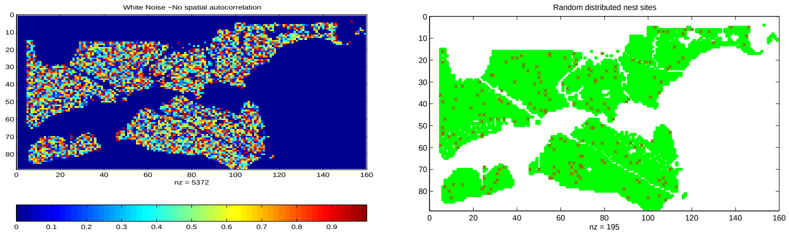
Figure 3: (a) Learning dataset: The geometry of the marshland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the marshland, (c) The spatial distribution of *water depth*, and (d) The spatial distribution of *distance to open water*.

The two models were evaluated based on their ability to predict the nest locations on the *test* data. We now describe a measure which can be used to compare the performance of the two models.

Metric of Comparison: We use Receiver Operating Characteristic(ROC) curves to compare the effectiveness of the two models. ROC curves plot the relationship between the true positive rate(TPR) and the false positive rate(FPR). For each cut-off probability b, $TPR(b)$ measures the ratio of the number of sites where the nest is actually located and was predicted divided by the number of actual nest sites. The FPR measures the ratio of the number of sites where the nest was absent but predicted divided by the number of sites where the nests were absent. The ROC curve is the locus of the pair $(TPR(b), FPR(b))$ for each cut-off probability. The higher the curve above the straight line $TPR = FPR$ the better the accuracy of the model.

Comparison in Space: We use the 1995 Stubble marshland data to make comparison in space. The result is shown in Figure 5. Clearly, by including a spatial autocorrelation term, there is substantial and systematic improvement for all levels of cut-off probability on both the learning data(1995 Darr) and test data(1995 Stubble).

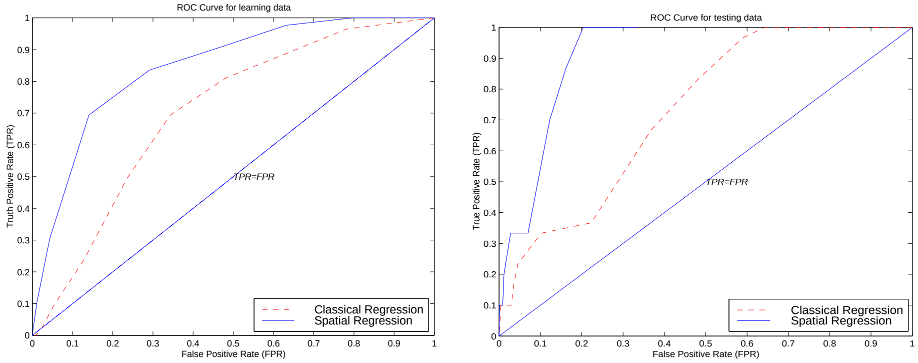
Comparison in Time: We also carried out experiments for making comparison in time. For this we used the 1996 data acquired in the Darr marshland. In this case there is virtually no significant improvement between the classical and spatial models. This is not entirely surprising because in



(a) pixel property with independent identical distribution

(b) Random nest locations

Figure 4: Spatial distribution satisfying random distribution assumptions of classical regression



(a) Learning Data

(b) Test Data

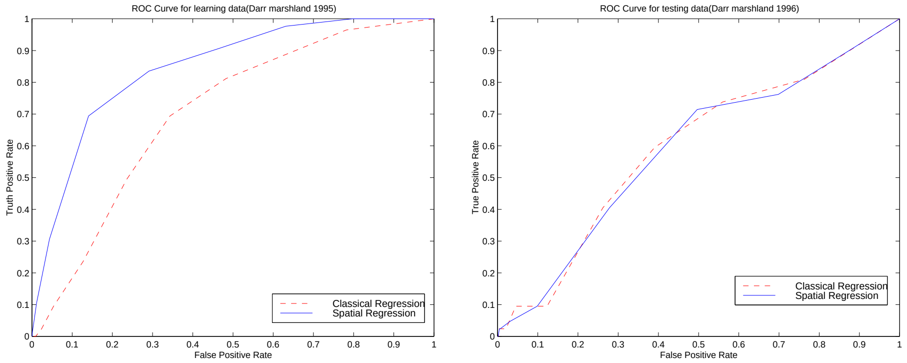
Figure 5: (a) Comparison of the classical and spatial regression model on the 1995 Darr marshland learning data. (b) Comparision of the two models on the 1995 Stubble marshland testing data.

1996 the nests of two bird species were counted in the Darr marshland. Also some environmental factors(e.g. water depth) have changed significantly in one year.

5 Applications of Spatial Data Mining for policy makers

Spatial data mining can be a helpful aid for managers involved in various levels of public decision making. For example, the bird habitat model that was discussed above can be used for environmental conservation efforts as it clearly describes the form of the relationships between different features in a marshland. The model can also serve as a basis for similar models in other areas of interest including fisheries management.

One of the greatest challenges facing city managers is to decide a way to balance the growth of cities while retaining unique environmental and wildlife zones within the city. Spatial data mining tools can play a crucial role here because they can quickly generate "What if" scenarios based on data that is being continously collected. Spatial data mining can be used to combine data from



(a) Learning Data

(b) Test Data

Figure 6: (a) Comparison of the classical and spatial regression model onthe 1995 Darr marshland learning data. (b) Comparision of the two models on the 1996 Darr marshland testing data.

remote sensing, cartographic maps, traffic sensors and the census to arrive at high level succint policy recommendations which can be debated, revised and implemented at different levels of the government hierarchy.

6 Discussion, Conclusion and Acknowledgements

Data mining is a relatively new term which unifies a set of problems which are common in many disciplines including econometrics, environmental management, regional science, geography, epedemiomology and remote sensing. The goal of data mining is to rapidly generate interesting and potentially useful hypothesis which can then be verified, modified and refined by using statistical techniques. Data mining is not a substitute for statistics but a tool which researchers and scientists can use to manage and deal with very large datasets which are fast becoming a norm rather than an exception.

In many areas including geography, ecology and regional science, the data generated usually has a strong spatial component. Like the area of spatial statistics, which has attained a distinct identity within statistics we believe that spatial data mining needs to carve out its own niche within the general framework of classical data mining.

In this article we have overviewed techniques from spatial statistics to extend regression analysis to explictly account for the special properties of spatial data, in particular spatial autocorrelation. We have also shown, with the help of an example, that current measures of classification accuracy may not be suitable when it comes to evaluation of spatial-based statistical models. Elsewhere we have exploited this fact and proposed a new measure of spatial accuracy which is similar to the notion of map-similarity. Based on this new measure of spatial accuracy we have proposed a framework to solve the two-class spatial classification problem. Initial experiments show that we can achieve considerable performance improvements(upto two-orders of magnitude) compared with spatial statistical techniques without comprimising on the accuracy of the results.

We would like to thank the Center for Urban and Regional Affairs(CURA) and in particular

William Craig for sponsoring this research effort. The work was also 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. We would also like to than James Lesage for making the spatial econometrics package available on the web: http://www.econo.utoledo.edu/l̃esage