Extending Data Mining for Spatial Applications: A Case Study in Predicting Nest Locations

Sanjay Chawla* Shashi Shekhar* Weili Wu* Uygar Ozesmi†

Abstract

Spatial data mining is a process to discover interesting and potentially useful spatial patterns embedded in spatial databases. Efficient tools for extracting information from spatial data sets can be of importance to organizations which own, generate and manage large 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 auto-correlation. In this paper we will review spatial statistical techniques which can effectively model the notion of spatial-autocorrelation and apply it to the problem of predicting bird nest locations in a wetland.

Keywords: Spatial data mining, spatial autocorrelation, spatial autoregression.

1 Introduction

Widespread use of spatial databases [9, 19] is leading to an increasing interest in mining interesting and useful but implicit spatial patterns[11, 14, 17]. Efficient tools for extracting information from spatial data, the focus of this work, are crucial to organizations which make decisions based on the analysis of large spatial data sets. These organizations are spread across many domains including ecology, environment management, public safety, transportation, public health, business logistics, travel and tourism. Classical data mining algorithms [7] often make assumptions(e.g. independent distributions) which violates the first law of Geography: everything is related to everything else but nearby things are more related than distant things [18]. In other words, the values of attributes of nearby spatial objects tend to systematically affect each other. In spatial statistics, an area within statistics devoted to the analysis of spatial data, this property is called spatial autocorrelation [4]. Knowledge discovery models which ignore spatial autocorrelation typically perform poorly in the presence of spatial data. In this paper we will review techniques from spatial econometrics which take the special properties of spatial data into account. In particular we will show how logistic regression can be generalized to model spatial autocorrelation. We will also make a case for the need for a new measure of spatial classification accuracy.

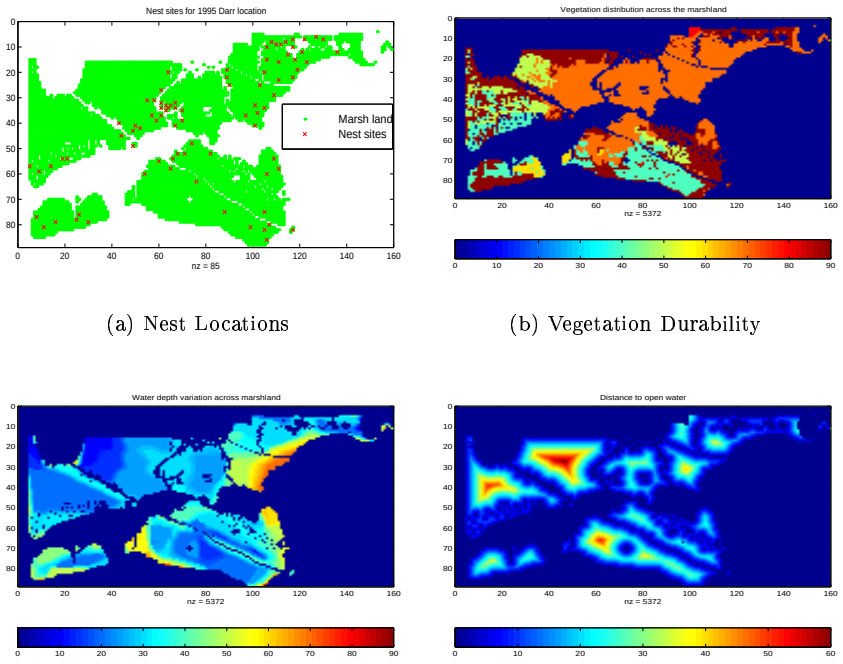
1.1 An Illustrative Application Domain

The availability of accurate spatial habitat models is an important tool for wildlife management, protection of critical habitat and endangered species. Since the underlying process governing the interaction between wildlife and environmental factors is complex, statistical techniques are used to gain insight on the basis of data collected during field work. One of the authors has been involved in the development of spatial model for the nesting locations of a marsh-nesting bird species [15, 16]. We will use this application, and the accompanying data, to explain how logistic regression can be extended to incorporate spatial autocorrelation.

The learning and test datasets were collected in 1995 and 1996 from two wetlands(Darr and Stubble) located on the shores of Lake Erie in Ohio. A uniform grid was imposed on the wetlands 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

*Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, USA. Support in part by the Army High Performance Computing Research Center under the auspices of Department of the Army, Army Research Laboratory Cooperative agreement number DAAH04-95-2-0003/contract number DAAH04-95-C-0008, and by the National Science Foundation under grant 9631539. Email: {chawla,shekhar,wuw}@cs.umn.edu

† Department of Environmental Sciences,Ericyes University,Kayseri, Turkey.Email:uozesmi@erciyes.edu.tr



(a) Nest Locations

(b) Vegetation Durability

(c) Water Depth

(d) Distance to Open Water

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

whether a bird-nest was present or not. The presence of the nest played the role of dependent variable. The geometry of the Darr wetland, locations of the nests and spatial distribution of the explanatory variables are shown in Figure 1. Classical data mining techniques like logistic regression[16] and neural networks[15] were applied to build spatial habitat models. Using logistic regression the nests could be classified 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.

Detailed discussions among authors reveal an important reason why, despite extensive domain knowledge, the results of classical data mining are not "satisfactory". Classical techniques make assumption about identical independent distribution(i.i.d.) for the properties of each pixel, ignoring spatial autocorrelation. Figure 2(a) shows a spatial distribution consistent with the assumptions of classical regression. It looks like "white noise" as properties of pixel are generated from independent identical distributions. Note that the maps of explanatory variable in Figure 1 have much more gradual variation indicating high spatial autocorrelation. Figure 3(b) shows a random distribution of nest locations which is quite different from the distribution of actual nests shown in Figure 1(a).

1.2 Spatial Data Mining: Problem Formulation

Predicting nest locations is a special case of a two-class spatial classification problem which we formally define as follows:

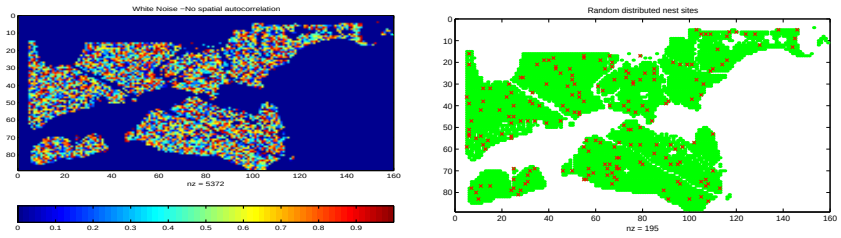
Given :

- A spatial framework S consisting of sites $\{s_1, \ldots, s_n\}$ for an underlying geographic space G.
- A collection of explanatory functions $f_{X_k} : S \rightarrow R^k, k = 1, \ldots K$. R^k is the range of possible values for the explanatory functions.
- A dependent function $f_Y : S \rightarrow R^Y$
- A family $\mathcal{F}$ of learning model functions mapping $R^1 \times \ldots R^K \rightarrow R^Y$.

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

Objective : maximize classification_accuracy$(\hat{f}^Y, f_Y)$

Constraints :



(a) pixel property with i.i.d

(b) Random nest sites

Figure 2: Spatial distribution satisfying distribution assumptions of classical regression

1. Geographic Space S is a multi-dimensional Euclidean Space [1].
2. The values of the explanatory functions, the f_{X_k}'s and the response function f_Y may not be independent w.r.t those of nearby spatial sites, i.e. spatial autocorrelation exists.
3. The domain R^k of the explanatory functions is the one-dimensional domain of real numbers.
4. The domain of the dependent variable, $R^Y = \{0, 1\}$.

1.3 Related Work

Related work includes spatial statistics and spatial data mining.

Spatial Statistics: The goal of spatial statistics is to model the special properties of spatial data. The primary distinguishing property of spatial data is that neighboring data samples tend to systematically affect each other. Thus the classical assumption that data samples are generated from independent and identical distributions is not valid. Current research in Spatial Econometrics, Geo-statistics and Ecological modeling [2, 13, 8] has focused on extending classical statistical techniques in order to capture the unique characteristics inherent in spatial data.

Spatial Data Mining: Spatial data mining [6, 10, 11, 12, 17], a subfield of data mining [1, 7], is concerned with discovery of interesting and useful but implicit knowledge in spatial databases. Challenges in Spatial Data Mining arise from the following issues. *First,* classical data mining[1] deals with numbers and categories. In contrast, spatial data is more *complex* and includes extended objects such as points, lines, and polygons. *Second,* classical data mining works with explicit inputs, whereas spatial predicates (e.g. overlap) are often *implicit*. *Third,* classical data mining treats each input to be independent of other inputs, whereas spatial patterns often exhibit continuity and *high autocorrelation among nearby features.* For example, population density of nearby locations are often related. In the presence of spatial data the standard approach in the data mining community is to materialize spatial relationships as attributes and rebuild the model with these "new" spatial attributes [12, 11].

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

1.4 Scope of Paper and Outline

The primary focus of this paper is to review techniques which generalize logistic regression to model the special properties of spatial data, namely spatial autocorrelation. Using the "bird-nesting" example introduced in Section 1.1 we will show that models which take spatial autocorrelation into account perform uniformly better than classical models. We will also make a case for a new measure for spatial classification accuracy which we believe is more suited to capture the special semantics of spatial data. The rest of the paper is as follows. In Section 2 we briefly describe the logistic regression model and highlight its key limitation for modeling spatial data. We will also introduce a statistic which quantifies the notion of spatial autocorrelation. In section 3 we will show how regression techniques can be generalized to model spatial aucorrelation. We also list some of the key advantages of doing so. In Section 4 we carry out experiements on the bird data set to compare the learning and predictive power of classical and spatial logistic regression. We conclude in Section 5 by making an argument for a new measure of spatial classification accuracy.

2 Basic Concepts: Modeling Spatial Dependencies

2.1 Logistic Regression Modeling

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

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

Here $\mathbf{X} = [1, \underline{X}]$ and $\beta = (\beta_0, \ldots, \beta_m)^t$. The standard assumption on the error vector ϵ is that each

component is generated from an independent and and identical and normal distribution, i.e, $\epsilon_i = N(0, \sigma^2)$.

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

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

2.2 Spatial Autocorrelation and Examples

There are many measures available for quantifying spatial autocorrelation. Each have their own strengths and weaknesses. Here we will briefly describe the Moran I measure.

In most cases the Moran's I measure (henceforth MI) ranges between -1 and +1 and thus is similar to the classical measure of correlation. Intuitively, a higher positive value is indicative of high spatial autocorrelation. This implies that like values tend to cluster together or attract each other. A low negative value is an indication that high and low values are interspersed. Thus like values are de-clustered and tend to repel each other. A value close to zero is an indication that no spatial trend (random distribution) is discernible using the given measure. The exact definition of MI is given in the Appendix.

All spatial autocorrelation measures are crucially dependent on the choice and design of the contiguity matrix W. The design of the matrix itself is predicated on determining "what constitutes a neighborhood of influence?" Two common choices are the four and the eight neighborhood. Thus given a lattice structure and a point S in the lattice, a four-neighborhood assumes that S influences all cells which share an edge with S. In an eight-neighborhood it is assumed that S influences all cells which either share an edge or a vertex. An eight neighborhood contiguity matrix is shown in Figure 3. The contiguity matrix of the uneven lattice(left) is shown on the right hand side. The contiguity matrix plays a pivotal role in the spatial extension of the regression model.

3 Spatial Regression Models

We now show how spatial dependencies are modeled in the framework of regression analysis. This may serve as a template for modeling spatial dependencies in other data mining techniques.

3.1 Spatial Autoregressive Model(SAM)

In spatial regression the spatial dependencies of the error term or the dependent variable are directly modeled in the regression equation [2]. Assume that the dependent values y_i' are related to each other, i.e. $y_i = f(y_j)$ $i \neq j$. Then the regression equation can be modified as

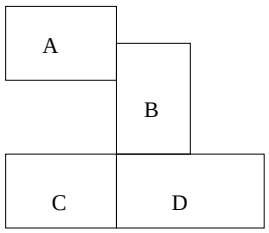
$$\mathbf{y} = \rho W\mathbf{y} + \mathbf{X}\beta + \epsilon.$$

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

We will refer to this equation as the Spatial Autoregressive Model(SAM). Notice when $\rho = 0$, this equation collapses to the classical regression model. The benefits of modeling spatial autocorrelation are many: (1) The residual error will have much lower spatial autocorrelation, i.e., systematic variation. With proper choice of W, the residual error should, at least theoretically, have no systematic variation. (2) If the spatial autocorrelation coefficient is statistically significant then it will quantify the presence of spatial autocorrelation. It will indicate the extent to which variations in the dependent variable (y) are explained by the average of neighboring observation values. (3) Finally, the model will have a better fit, i.e., higher R-squared statistic(See the Appendix for a dramatic example).

As in the case of classical regression, the SAM equation has to be transformed via the logistic function for binary dependent variables. The estimates of ρ and β can be derived using maximum likelihood theory or Bayesian statistics. We have carried out preliminary experiments using the spatial econometrics matlab package [2] which implements a Bayesian approach via Gibbs Sampling [13].

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



(a) Map

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

b) Contiguity Matrix W

Figure 3: A spatial neighborhood and its contiguity matrix

4 Experiment Evaluation of Spatial Autoregression

4.1 Experiment Design

Goals: The goal of the experiments is to evaluate the effects of including the spatial autoregressive term, $\rho W\mathbf{y}$, in the logistic regression model. The experimental setup is shown in Figure 4. The 1995 Darr wetland data was used as the learning set to build the two models. The parameters of the classical and spatial models were dervied using maximum likelihood estimation and Gibbs Sampling respectively. The two models were evaluated based on their ability to predict the nest locations on the test data. Classification accuracy, which we describe next, was used to evalute the two models.

Metric of Comparison: Classification accuracy achieved by classical and spatial logistic regression are compared on the test data. We use the Receiver Operating Characteristic(ROC) [5] curves to compare classification accuracy. ROC curves plot the relationship between the true positive rate(TPR) and the false positive rate(FPR). For each cut-off probability b, $TPR(b)$ measures the ratio of the number of sites where the nest is actually located and was predicted divided by the number of actual nest sites. The FPR measures the ratio of the number of sites where the nest was absent but predicted divided by the number of sites where the nests were absent. The ROC curve is the locus of the pair $(TPR(b), FPR(b))$ for each cut-off probability. The higher the curve above the straight line $TPR = FPR$ the better the accuracy of the model.

Comparison in Space: We use the 1995 Stubble wetland 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 wetland. In this case there is virtually no significant improvement between the classical and spatial models. This is not entirely surprising because in 1996 the nests of two bird species were counted in the Darr wetland. Also some environmental factors(e.g. water depth) have changed significantly in one year [15, 16].

5 Discussion, Conclusion and Future Work

The standard measure for classification accuracy may not be the most appropriate for making spatial predictions. What is needed is a measure of spatial classification accuracy. Spatial accuracy is important because of the effects of discretizations of continuous marsh into discrete pixels, as shown in Figure 7. Figure 7(a) shows the actual locations of nests and 7(b) shows the pixels with actual nests. Note the loss of information during the discretization of continuous space into pixels. Many nest location barely fell within the pixels labeled 'A' and were quite close to other pixels with label of no-nest. Now consider two predictions shown in Figure 7(c) and 7(d). Domain scientists prefer prediction 7(d) over 7(c), since predicted nest locations are closer on average to some actual nest locations. Classification accuracy measure cannot distinguish between 7(c) and 7(d), and a measure of spatial accuracy is needed to capture this preference [3].

We have shown that augmenting classical regression models with a spatial autoregressive term leads to substantial improvements in the predictive power of the models. In order to include the spatial autoregressive term, a contiguity matrix which captures the spatial relationship between the locations of data samples has to be constructed. We have also shown that the classical measures of classification accuracy may not be appropriate to measure the predictive power of spatial regression models.

References

[1] R. Agrawal. Tutorial on database mining. In *Thirteenth ACM Symposium on Principles of Databases Systems*, pages 75–76, Minneapolis, MN, 1994.

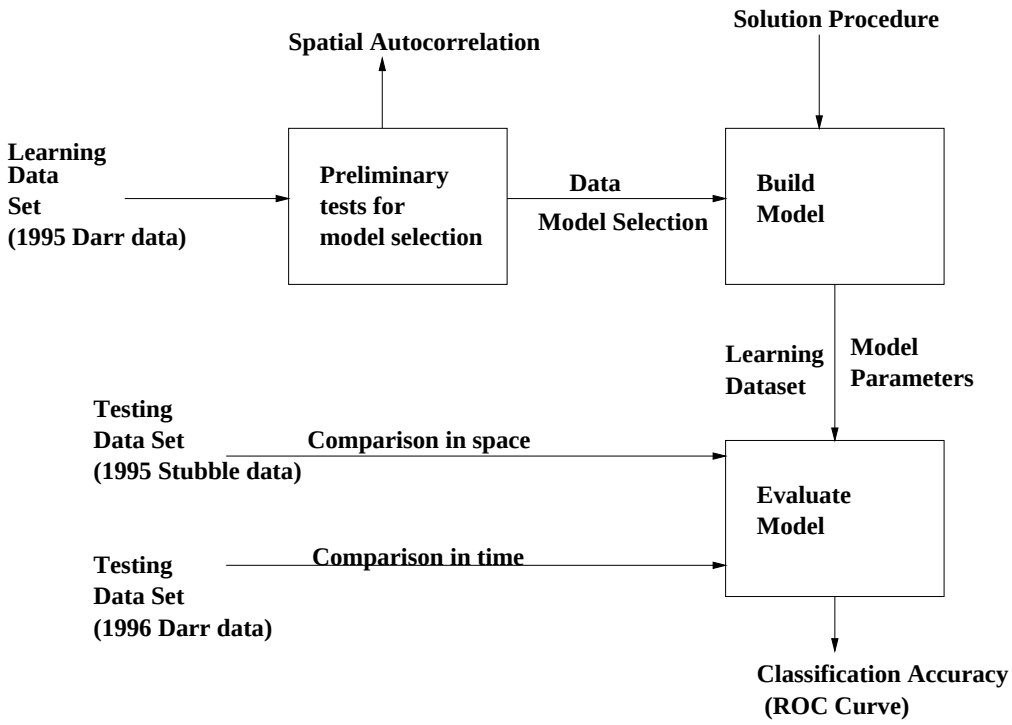


Figure 4: Experimental Method for evaluation spatial autoregression

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

[3] S. Chawla, S. Shekhar, W-L Wu, and U. Ozesmi. Modeling spatial dependencies for mining geospatial data: An introduction. In *Geographic data mining and Knowledge Discovery(GKD) (Ed. Harvey Miller and Jiawei Han), under contract with Taylor and Francis, URL: http://www.geog.utah.edu/ ˜hmiller/gkd_text.*

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

[5] J.P. Egan. *Signal Detection Theory and ROC analysis*. Academic Press, New York, 1975.

[6] M. Ester, H-P Kriegel, and J. Sander. Knowledge discovery in spatial databases. In *Advances in Artificial Intelligence, 23rd Annual German Conference on Artificial Intelligence*, pages 61–74, Bonn, Germany, September 1999.

[7] U. M. Fayyad. Knowledge discovery in databases: An overview. In *Inductive Logic Programming, 7th International Workshop, ILP-97, Lecture Notes in Computer Scienc*, volume 1297, pages 3–16. Springer, September 1997.

[8] D. Griffith. Statistical and mathematical sources of regional science theory: Map pattern analysis as an example. *Papers in Regional Science (Publisher: Springer)*, (78):21–45, 1999.

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

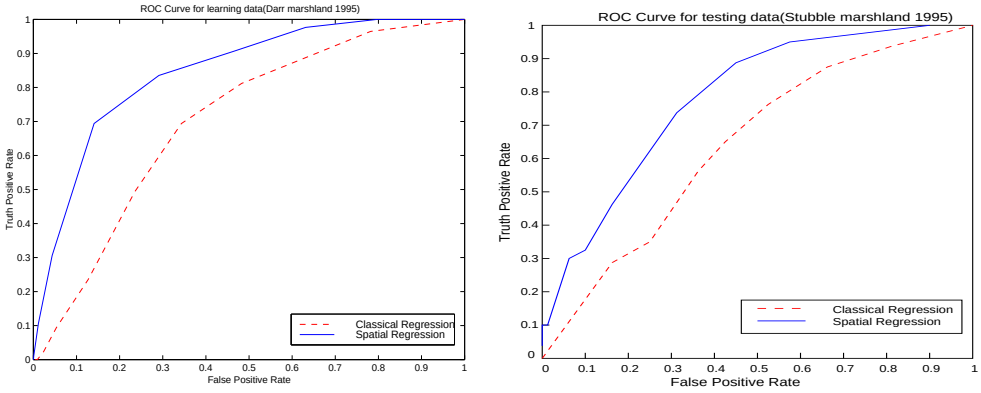
[10] E. Knorr and R. Ng. Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining. *IEEE TKDE*, 8(6):884–897, 1996.

[11] K. Koperski, J. Adhikary, and J. Han. Spatial data mining: Progress and challenges. In *Workshop on Research Issues on Data Mining and Knowledge Discovery(DMKD'96)*, pages 1–10, Montreal, Canada, 1996.

[12] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Advances in Spatial Databases, Proc. of 4th International Symposium, SSD'95*, pages 47–66, Portland, Maine, USA, 1995.

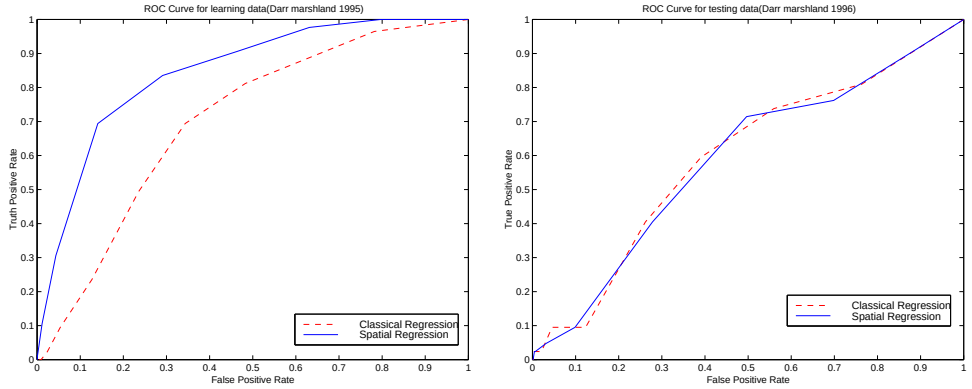
[13] J.P. LeSage. Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, (20):113–129, 1997.

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



(a) Learning Data (b) Test Data

Figure 5: (a) Comparison of the probit and probit with spatial autocorrelation on the 1995 Darr wetland learning data. (b) Comparision of the two models on the 1995 Stubble wetland testing data.



(a) Learning Data (b) Test Data

Figure 6: (a) Comparison of the probit and probit with spatial autocorrelation on the 1995 Darr wetland learning data. (b) Comparision of the two models on the 1996 Darr wetland testing data.

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

[16] U. Ozesmi and W. Mitsch. A spatial habitat model for the Marsh-breeding red-winged blackbird(agelaius phoeniceus l.) In coastal lake Erie wetlands. *Ecological Modelling (Publisher: Elsevier Science B. V.)*, (101):139–152, 1997.

[17] John F. Roddick and Myra Spiliopoulou. A bibliography of temporal, spatial and spatio-temporal data mining research. *ACM Special Interest Group on Knowledge Discovery in Data Mining(SIGKDD) Explorations*, 1999.

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

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

6 Appendix:Spatial Autocorrelation

6.1 Moran's I measure

There are many measures available for quantifying spatial autocorrelation. Each have their own str engths and weaknesses. The two most well known measures are

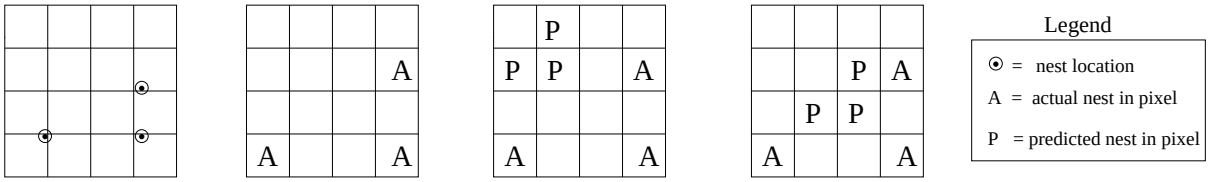


Figure 7: (a)The actual locations of nest, (b)Pixels with actual nests, (c)Location predicted by a model, (d)Location predicted by another mode. Prediction(d) is spatially more accurate than (c). Classical measures of classification accuracy will not capture this distincition.

Moran's I and Geary's C measure. Here we will briefly describe the Moran I measure.

In most cases the Moran's I measure (henceforth MI) ranges between -1 and +1 and thus is similar to the classical measure of correlation. Intuitively, a higher positive value is indicative of high spatial autocorrelation. This implies that like values tend to cluster together or attract each other. A low negative value is an indication that high and low values are interspersed. Thus like values are de-clustered and tend to repel each other. A smooth surface will have a high spatial autocorrelation and a ches sboard-like surface a high negative spatial autocorrelation. A value close to zero is an indicatio n that no spatial trend (random distribution) is discernible using the given measure.

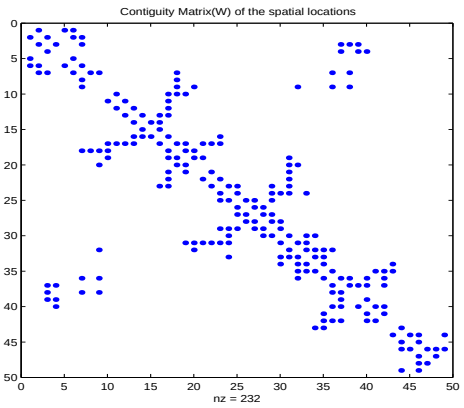
The formula for MI is

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

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

6.2 Example of including the spatial autoregressive term

Figure 8 shows an example of how by adding a spatial autoregressive term leads to improvement in the accuracy of a linear regression model when applied to spatial data. The data was extracted from a crime data set in 49 neighborhoods in Columbus, Ohio [2]. The dependent variable is the *number of crime incidents* and the independent variables are *mean income* and *mean house value*.



(a) The contiguity matrix of locations

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

(b) R-square and Moran I of residual

Figure 8: (a)The contiguity matrix of the 49 neighborhoods in Columbus, Ohio. (b) Including the spatial autoregressive term reduces the systematic variation in the residual error term(lower Moran I) and consequently is a better fit(higher R^2).