Spatial Contextual Classification and Prediction Models for Mining Geospatial Data

Shashi Shekhar* Paul R. Schrater† Ranga R. Vatsavai* Weili Wu*
Sanjay Chawla‡

February 4, 2002

Abstract

Modeling spatial context (e.g., autocorrelation) is a key challenge in classification problems that arise in geospatial domains. Markov Random Fields (MRFs) is a popular model for incorporating spatial context into image segmentation and land-use classification problems. The spatial autoregression model (SAR), which is an extension of the classical regression model for incorporating spatial dependence, is popular for prediction and classification of spatial data in regional economics, natural resources, and ecological studies. There is little literature comparing these alternative approaches to facilitate the exchange of ideas (e.g., solution procedures). We argue that the SAR model makes more restrictive assumptions about the distribution of feature values and class boundaries than MRF. The relationship between SAR and MRF is analogous to the relationship between regression and Bayesian classifiers. This paper provides comparisons between the two models using a probabilistic and an experimental framework.

Keywords: Spatial Context, Spatial Data Mining, Markov Random Fields, Spatial Autoregression.

1 Introduction

Spatial databases (e.g., remote sensing imagery, maps, census data) are an important subclass of multimedia databases due to several reasons. First, the industry-wide Structured Query Language Multimedia standard (SQL/MM) [20] includes spatial data types along with traditional image, audio and video data types. Secondly, spatial concepts and techniques are often crucial in indexing and retrieval of image and video databases. Finally, according to several estimates, spatial data constitutes almost 80% of all digital data including multimedia data.

Widespread use of spatial databases [28], is leading to an increasing interest in mining interesting and useful but implicit spatial patterns[14, 19, 10, 26]. Traditional data mining algorithms[1]

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

†Department of Psychology, University of Minnesota, Minneapolis, MN, 55455. Email: schrater@eye.psych.umn.edu

‡Vignette Corporation, Boston, MA, USA. Email: schawla@vignette.com

often make assumptions (e.g., independent, identical distributions) which violate Tobler's first law of geography: everything is related to everything else but nearby things are more related than distant things[30]. In other words, the values of attributes of nearby spatial objects tend to systematically affect each other. In spatial statistics, an area within statistics devoted to the analysis of spatial data, this is called spatial autocorrelation[7]. Knowledge discovery techniques which ignore spatial autocorrelation typically perform poorly in the presence of spatial data. Often the spatial dependencies arise due to the inherent characteristics of the phenomena under study, but in particular they arise due to the fact that the spatial resolution of imaging sensors are finer than the size of the object being observed. For example, remote sensing satellites have resolutions ranging from 30 meters (e.g., the Enhanced Thematic Mapper of the Landsat 7 satellite of NASA) to one meter (e.g., the IKONOS satellite from SpaceImaging), while the objects under study (e.g., Urban, Forest, Water) are often much larger than 30 meters. As a result, per-pixel-based classifiers, which do not take spatial context into account, often produce classified images with *salt and pepper* noise. These classifiers also suffer in terms of classification accuracy.

There are two major approaches for incorporating spatial dependence into classification/prediction models: spatial autoregression models [2], [15], [16], [17], [23], [24] and Markov Random Field models [5], [6], [9], [13], [18], [29], [31]. Here we want to make a note regarding the terms *spatial dependence* and *spatial context*. These words originated in two different communities. Natural resource analysts and statisticians use *spatial dependence* to refer to *spatial autocorrelation* and the image processing community uses spatial context to mean the same. We use *spatial context*, *spatial dependence*, and *spatial autocorrelation* interchangeably to relate to readers of both communities. We also use *classification* and *prediction* interchangeably. Natural resource scientists, ecologists and economists have incorporated spatial dependence in spatial data analysis by incorporating spatial autocorrelation into logistic regression models (called SAR). The Spatial Autoregressive Regression (SAR) model states that the class label of a location is partially dependent on the class labels of nearby locations and partially dependent on the feature values. SAR tends to provide better models than logistic regression in terms of achieving higher confidence (R^2). Similarly, Markov Random Fields (MRFs) is a popular model for incorporating spatial context into image segmentation and land-use classification problems. Over the last decade, several researchers [29], [13], [31] have exploited spatial context in classification using Markov Random Fields to obtain higher accuracies over their counterparts (i.e., non-contextual classifiers). MRFs provide a uniform framework for integrating spatial context and deriving the probability distribution of interacting objects.

There is little literature comparing alternative models for capturing spatial context, hampering the exchange of ideas across communities. For example, solution procedures [17] for SAR tend to be computationally expensive just like the earlier stochastic relaxation [9] approaches for MRF despite optimizations such as sparse-matrix techniques [23], [24]. Recently, new solution procedures, (e.g., graph cuts [5]), have been proposed for MRF. An understanding of the relationship between MRF and SAR may facilitate the development of new solution procedures for SAR. It may also likely lead to cross fertilization of other advances across the two communities.

We compare the SAR and MRF models in this paper using a common probabilistic framework. SAR and MRF use identical models of spatial contexts for spatial locations. However, SAR makes more restrictive assumptions about the probability distributions of feature values as well as the class boundaries. We show that the SAR assumption of the conditional probability of a feature value given a class label means that SAR belongs to the exponential family of models, (e.g., Gaussian, Binomial). In contrast, MRF models can work with many other probability distributions. SAR also assumes the linear separability of classes in a transformed feature space resulting from a spatial smoothing of feature values based on autocorrelation parameters. MRF can be used with

non-linear class boundaries. Readers familiar with classification models which ignore spatial context may find the following analogy helpful. The relationship between SAR and MRF is similar to the relationship between logistic regression and Bayesian classifiers.

Outline and Scope of the Paper:
The rest of the paper is organized as follows. In Section 1.1 we introduce a motivating example which will be used throughout the paper. In Section 1.2 we formally define the location prediction problem. Section 2 presents a comparison of classical approaches that do not consider spatial context, namely logistic regression and Bayesian classifiers. In Section 3 we present two modern approaches that model spatial context, namely Spatial Autoregressive Regression (SAR) [15], and Markov Random Fields. In Section 4 we compare and contrast the SAR and MRF models in a common probabilistic framework and provide experimental results. Finally, Section 5 provides conclusions and future research directions.

This paper focuses on a comparison of SAR and MRF. Comparison of other models of spatial context, and evaluation and translation of new solution procedures for MRF, (e.g., Graph cuts, to new solution procedures for SAR are beyond the scope of this paper. We plan to address these issues in future work.

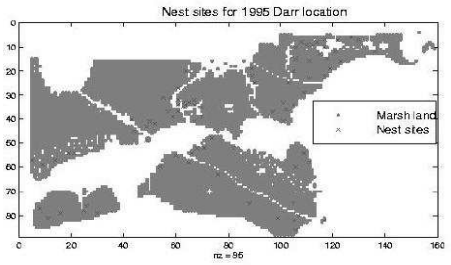
1.1 An Illustrative Application Domain

First we introduce an example which will be used throughout this paper to illustrate the different concepts in spatial data mining. We are given data about two wetlands, named Darr and Stubble, on the shores of Lake Erie in Ohio USA in order to *predict* the spatial distribution of a marsh-breeding bird, the red-winged blackbird (*Agelaius phoeniceus*) [21], [22]. The data was collected from April to June in two successive years, 1995 and 1996.

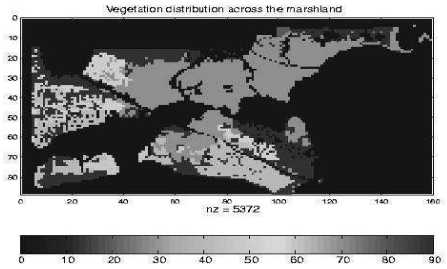
A uniform grid was imposed on the two wetlands and different types of measurements were recorded at each cell or pixel. In total, the values of seven attributes were recorded at each cell. Domain knowledge is crucial in deciding which attributes are important and which are not. For example, *Vegetation Durability* was chosen over *Vegetation Species* because specialized knowledge about the bird-nesting habits of the red-winged blackbird suggested that the choice of nest location is more dependent on plant structure and plant resistance to wind and wave action than on the plant species.

An important goal is to build a model for predicting the location of bird nests in the wetlands. Typically, the model is built using a portion of the data, called the **Learning** or **Training** data, and then tested on the remainder of the data, called the **Testing** data. In this study we build a model using the 1995 Darr wetland data and then tested it 1995 Stubble wetland data. In the learning data, all the attributes are used to build the model and in the training data, one value is *hidden*, in our case the location of the nests. Using knowledge gained from the 1995 Darr data and the value of the independent attributes in the test data, we want to predict the location of the nests in 1995 Stubble data.

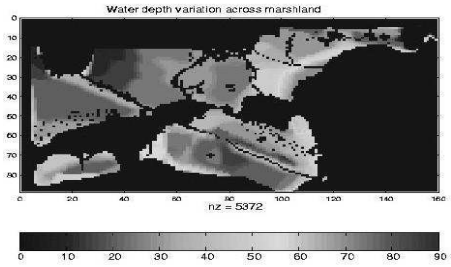
In this paper we focus on three independent attributes, namely *Vegetation Durability*, *Distance to Open Water*, and *Water Depth*. The significance of these three variables was established using classical statistical analysis [22]. The spatial distribution of these variables and the actual nest locations for the Darr wetland in 1995 are shown in Figure 1. These maps illustrate the following two important properties inherent in spatial data. The value of attributes which are referenced by spatial location tend to vary gradually over space. While this may seem obvious, classical data



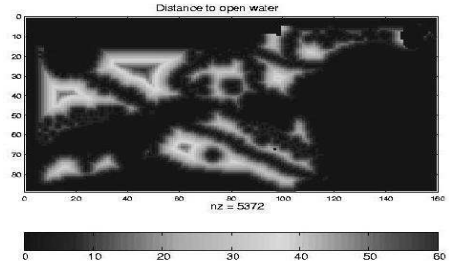
(a) Nest Locations



(b) Vegetation Durability



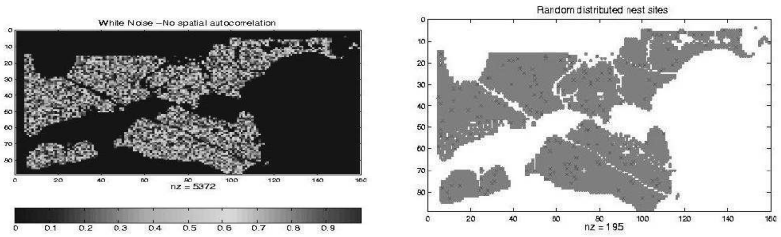
(c) Water Depth



(d) Distance to Open Water

Figure 1: (a) Learning dataset: The geometry of the Darr wetland 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*.

mining techniques, either explicitly or implicitly, assume that the data is *independently* generated. For example, the maps in Figure 2 show the spatial distribution of attributes if they were independently generated. Previous studies have evaluated classical data mining techniques such as logistic regression[22], neural networks[21], decision trees, and classification rules to build prediction models for bird nesting locations. Logistic regression was used because the dependent variable is binary (nest/no-nest) and the logistic function "squashes" the real line onto the unit-interval. The values in the unit-interval can then be interpreted as probabilities. These studies concluded that with the use of logistic regression, the nests could be classified at a rate 24% better than random[21]. In general, logistic regression and neural network models have performed better than decision trees and classification rules on this dataset. The fact that classical data mining techniques ignore spatial autocorrelation and spatial heterogeneity in the model-building process is one reason why these techniques do a poor job. A second, more subtle, but equally important reason is related to the choice of the objective function to measure classification accuracy. For a two-class problem, the standard way to measure classification accuracy is to calculate the percentage of correctly classified objects. This measure may not be the most suitable in a spatial context. *Spatial accuracy*—how far the predictions are from the actuals—is as important in this application domain due to the effects of discretizations of a continuous wetland into discrete pixels, as shown in Figure 3. Figure 3(a) shows the actual locations of nests and 3(b) shows the pixels with actual nests. Note the loss of information during the discretization of continuous space into pixels. Many nest locations barely fall within the pixels labeled 'A' and are quite close to other blank pixels, which represent 'no-nest'. Now consider two predictions shown in Figures 3(c) and 3(d). Domain scientists prefer prediction



(a) Pixel property with independent identical distribution

(b) Random nest locations

Figure 2: Spatial distribution satisfying random distribution assumptions of classical regression

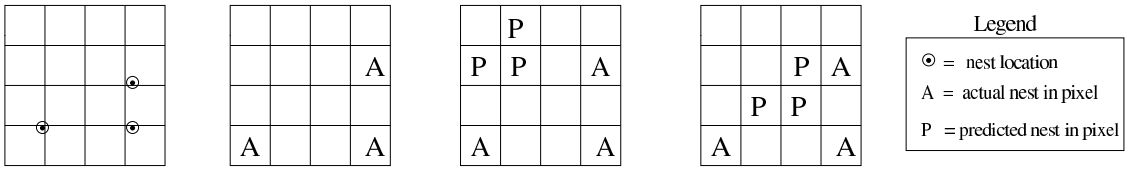


Figure 3: An example showing different predictions: (a)The actual locations of nests, (b)Pixels with actual nests, (c)Locations predicted by one model, (d)Locations predicted by another model. Prediction (d) is spatially more accurate than (c).

3(d) over 3(c), since predicted nest locations are closer on average to some actual nest locations. The classification accuracy measure cannot distinguish between 3(c) and 3(d), and a measure of spatial accuracy is needed to capture this preference.

1.2 Location Prediction: Problem Formulation

The Location Prediction problem is a generalization of the nest location prediction problem. It captures the essential properties of similar problems from other domains including crime prevention and environmental management. The problem is formally defined as follows:

Given:

- A spatial framework S consisting of sites $\{s_1, \ldots, s_n\}$ for an underlying geographic space G.
- A collection X 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. Let $X = [1, X]$, which also includes a constant vector along with explanatory functions.
- A dependent class variable $f_C : S \rightarrow C = \{c_1, \ldots c_M\}$
- An value for parameter α, relative importance of spatial accuracy.

Find: Classification model: $\hat{f}_C : R^1 \times \ldots R^k \rightarrow C$.

Objective: Maximize similarity $(map_{s_i \in S}(\hat{f}_C(f_{X_1}, \ldots, f_{X_k})), map(f_C))$
$= (1 - \alpha)$ classification_accuracy$(\hat{f}_C, f_C) + (\alpha)$spatial_accuracy$((\hat{f}_C, f_C)$

Constraints:

1. Geographic Space S is a multi-dimensional Euclidean Space [1].
2. The values of the explanatory functions, $f_{X_1}, ..., f_{X_k}$ and the dependent class variable, f_C, may not be independent with respect to the corresponding values 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 dependent variable, $C = \{0, 1\}$.

The above formulation highlights two important aspects of location prediction. It explicitly indicates that (i) the data samples may exhibit spatial autocorrelation and, (ii) an objective function (i.e., a map similarity measure), is a combination of classification accuracy and spatial accuracy. The *similarity* between the dependent variable f_C and the predicted variable $\hat{f}_C$ is a combination of the "traditional classification" accuracy and representation-dependent "spatial classification" accuracy. The regularization term α controls the degree of importance of **spatial accuracy** and is typically domain dependent. As $\alpha \to 0$, the map similarity measure approaches the traditional classification accuracy measure. Intuitively, α captures the spatial autocorrelation present in spatial data.

The study of the nesting locations of red-winged black birds[21, 22] is an instance of the location prediction problem. The underlying spatial framework is the collection of $5m \times 5m$ pixels in the grid imposed on the marshes. Examples of the explanatory variables include water depth, vegetation durability index, and distance to open water, and examples of dependent variables include nest locations. The explanatory and dependent variables exhibit spatial autocorrelation, (e.g., gradual variation over space, as shown in Figure 1). Domain scientists prefer spatially accurate predictions which are closer to actual nests, (i.e., $\alpha > 0$).

2 Classification Without Spatial Dependence

In this section we briefly review two major statistical techniques that have been commonly used in the classification problem: logistic regression and Bayesian classifiers. These models do not consider spatial dependence. Readers familiar with these two models will find it easier to understand the comparison between SAR and MRF presented later.

2.1 Logistic Regression Modeling

Logistic regression decomposes $\hat{f}_C$ into two parts, namely linear regression and logistic transformation. Given an $n-$vector y of observations and an $n \times m$ matrix X of explanatory data, classical linear regression models the relationship between y and X as

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

where $\beta = (\beta_0, \ldots, \beta_m)^T$. The standard assumption on the error vector ϵ is that each component is generated from an independent, identical, zero-mean normal distribution (i.e., $\epsilon_i = N(0, \sigma^2)$).

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

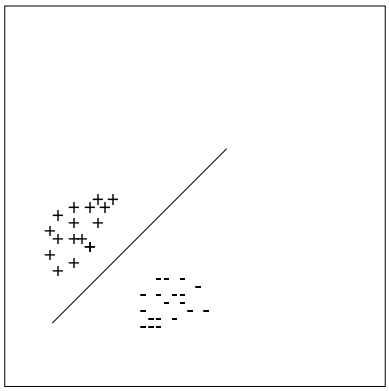


Figure 4: Two-dimensional feature space, with two classes (+:nest, -:no-nest) that can be separated by a linear surface

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, $Pr(c_i|y) = \frac{e^y}{1+e^y}$. This transformed model is referred to as **logistic** regression [2].

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 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 may be a poor fit to the geospatial data. Incidentally, the notion of spatial autocorrelation is similar to that of time autocorrelation in time series analysis but is more difficult to model because of the multi-dimensional nature of space. A statistic that quantifies spatial autocorrelation is introduced in the spatial autoregression model (SAR).

The logistic regression finds a discriminant surface, which is a hyperplane in feature space, as shown in Figure 4. Formally, a logistic-regression-based classifier is equivalent to a perceptron [12], [27], [11], which can only separate linearly separable classes.

2.2 Bayesian Classifiers

Bayesian classifiers estimate $\hat{f}_C$ using Bayes' rule and compute the probability of the class labels c_i given the data X as:

$$Pr(c_i|X) = \frac{Pr(X|c_i)Pr(c_i)}{Pr(X)} \quad (1)$$

In the case of the location prediction problem, where a single class label is predicted for each location, a decision step can assign the most-likely class chosen by Bayes' rule to be the class for a given location. This solution is often referred to as the maximum a posteriori estimate(MAP).

Given a learning dataset, $Pr(c_i)$ can be computed as a ratio of the number of locations s_j with

$f_C(s_j) = c_i$ to the total number of locations in S. $Pr(X|c_i)$ also can be estimated directly from the data using histograms or a kernel density estimate over the counts of locations s_j in S for different values X of features and different class labels c_i. This estimation requires a large training set if the domains of features f_{X_k} allow a large number of distinct values. A possible approach is that when the joint-probability distribution is too complicated to be directly estimated, then a sufficiently large number of samples from the conditional probability distributions can be used to estimate the *statistics* of the full joint probability distribution [2]. $Pr(X)$ need not be estimated separately. It can be derived from estimates of $Pr(X|c_i)$ and $Pr(c_i)$. Alternatively, it may be left as unknown, since for any given dataset, $Pr(X)$ is a constant that does not affect the assignment of class labels.

	Classifier	Classifier
Criteria	Logistic Regression	Bayesian
Input	$f_{x_1}, \ldots, f_{x_k}, f_c$	$f_{x_1}, \ldots, f_{x_k}, f_c$
Intermediate Result	β	$Pr(c_i)$, $Pr(X\|c_i)$ using kernel esti.
Output	$Pr(c_i\|X)$ based on β	$Pr(c_i\|X)$ based on $Pr(c_i)$ and $Pr(X\|c_i)$
Decision	Select most likely class for a given feature value	Select most likely class for a given feature value
Assumptions - $Pr(X\|c_i)$ - class boundaries - autocorrelation in class labels	 Exponential Family linearly separable in feature space none	 - - none

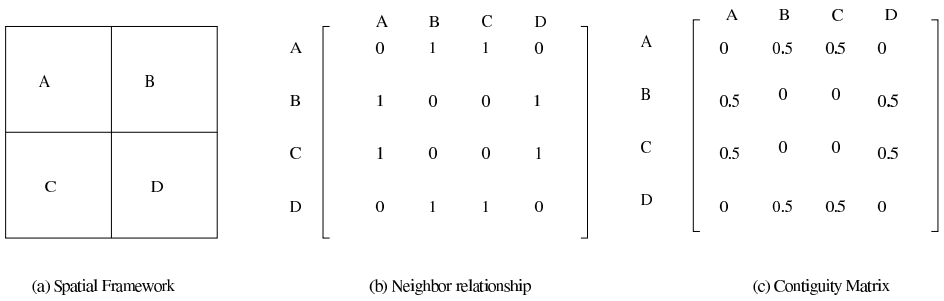
Table 1: Comparison of Logistic Regression and Bayesian Classifiers

Table 1 summarizes key properties of logistic-regression-based classifiers and Bayesian classifiers. Both models are applicable to the location prediction problem if spatial autocorrelation is insignificant. However, they differ in many areas. Logistic regression assumes that the $Pr(X|c_i)$ distribution belongs to an exponential family (e.g., binomial, normal) whereas Bayesian classifiers can work with arbitrary distributions. Logistic regression finds a linear classifier specified by β and Bayesian classifier is most effective when classes are not linearly separable in feature space, since it allows non-linear interaction among features in estimating $Pr(X|c_i)$. Logistic regression can be used with a relatively small training set since it estimates only $(k+1)$ parameters (i.e., β). Bayesian classifiers usually need a larger training set to estimate $Pr(X|c_i)$ due to the potentially large size of the feature space. In many domains, parametric probability distributions (e.g., normal [29], Beta) are used with Bayesian classifiers if large training datasets are not available.

3 Modeling Spatial Dependencies

Several previous studies [13], [29] have shown that modeling of spatial dependency (often called context) during the classification process improves overall classification accuracy. Spatial context can be defined by the relationships between spatially adjacent pixels in a small neighborhood. The spatial relationship among locations in a spatial framework is often modeled via a contiguity matrix. A simple contiguity matrix may represent the neighborhood relationship defined using adjacency, Euclidean distance, etc. Example definitions of neighborhood using adjacency

[2]While this approach is very flexible and the workhorse of Bayesian statistics, it is a computationally expensive process. Furthermore, at least for non-statisticians, it is a non-trivial task to decide what "priors" to choose and what analytic expressions to use for the conditional probability distributions.



(a) Spatial Framework (b) Neighbor relationship (c) Contiguity Matrix

Figure 5: A spatial framework and its four-neighborhood contiguity matrix

include four-neighborhood and eight-neighborhood. Given a gridded spatial framework, the four-neighborhood assumes that a pair of locations influence each other if they share an edge. The eight-neighborhood assumes that a pair of locations influence each other if they share either an edge or a vertex.

Figure 5(a) shows a gridded spatial framework with four locations, namely A, B, C, and D. A binary matrix representation of a four-neighborhood relationship is shown in Figure 5(b). The row normalized representation of this matrix is called a contiguity matrix, as shown in Figure 5(c). Other contiguity matrices can be designed to model neighborhood relationship based on distance. The essential idea is to specify the pairs of locations that influence each other along with the relative intensity of interaction. More general models of spatial relationships using cliques and hypergraphs are available in the literature [31].

3.1 Logistic Spatial Autoregression Model(SAR)

Logistic SAR decomposes $\hat{f}_C$ into two parts, namely Spatial autoregression and logistic transformation. We first show how spatial dependencies are modeled in the framework of logistic regression analysis. In the spatial autoregression model, the spatial dependencies of the error term, or, the dependent variable, are directly modeled in the regression equation[2]. If the dependent values y_i are related to each other, then the regression equation can be modified as

$$y = \rho W y + X\beta + \epsilon. \tag{2}$$

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 the correction term $\rho W y$ is introduced, the components of the residual error vector ϵ are then assumed to be generated from independent and identical standard normal distributions. As in the case of classical regression, the SAR equation has to be transformed via the logistic function for binary dependent variables.

We refer to this equation as the ***Spatial Autoregression Model (SAR).*** Notice that when $\rho = 0$, this equation collapses to the classical regression model. The benefits of modeling spatial autocorrelation are many: The residual error will have much lower spatial autocorrelation (i.e., systematic variation). With the proper choice of W, the residual error should, at least theoretically, have no systematic variation. If the spatial autocorrelation coefficient is statistically significant, then SAR 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. Finally, the model will have a better fit, (i.e., a higher R-squared statistic). We compare SAR with linear regression for predicting nest location in Section 4.

A mixed model extends the general linear model by allowing a more flexible specification of the covariance matrix of ϵ. The SAR model can be extended to a mixed model that allows for explanatory variables from neighboring observations [16]. The new model (MSAR) is given by

$$y = \rho Wy + X\beta + WX\gamma + \epsilon. \tag{3}$$

The marginal impact of the explanatory variables from the neighboring observations on the dependent variable y can be encoded as a $k * 1$ parameter vector γ.

Solution Procedures
The estimates of ρ and β can be derived using maximum likelihood theory or Bayesian statistics. We have carried out preliminary experiments using the spatial econometrics matlab package[3], which implements a Bayesian approach using sampling-based Markov Chain Monte Carlo (MCMC) methods[17]. Without any optimization, likelihood-based estimation would require $O(n^3)$ operations. Recently [23], [24], and [16] have proposed several efficient techniques to solve SAR. The techniques studied include divide and conquer, and sparse matrix algorithms. Improved performance is obtained by using LU decompositions to compute the log-determinant over a grid of values for the parameter ρ by restricting it to $[0, 1]$.

3.2 Markov Random Field based Bayesian Classifiers

Markov random field based Bayesian classifiers estimate classification model $\hat{f}_C$ using MRF and Bayes' rule. A set of random variables whose interdependency relationship is represented by an undirected graph (i.e., a symmetric neighborhood matrix) is called a Markov Random Field [18]. The Markov property specifies that a variable depends only on its neighbors and is independent of all other variables. The location prediction problem can be modeled in this framework by assuming that the class label, $l_i = f_C(s_i)$, of different locations, s_i, constitute an MRF. In other words, random variable l_i is independent of l_j if $W(s_i, s_j) = 0$.

The Bayesian rule can be used to predict l_i from feature value vector X and neighborhood class label vector L_i as follows:

$$Pr(l_i|X, L_i) = \frac{Pr(X|l_i, L_i)Pr(l_i|L_i)}{Pr(X)} \tag{4}$$

The solution procedure can estimate $Pr(l_i|L_i)$ from the training data, where L_i denotes a set of labels in the neighborhood of s_i excluding the label at s_i, by examining the ratios of the frequencies of class labels to the total number of locations in the spatial framework. $Pr(X|l_i, L_i)$ can be estimated using kernel functions from the observed values in the training dataset. For reliable estimates, even larger training datasets are needed relative to those needed for the Bayesian classifiers without spatial context, since we are estimating a more complex distribution. An assumption on $Pr(X|l_i, L_i)$ may be useful if the training dataset available is not large enough. A common assumption is the uniformity of influence from all neighbors of a location. For computational efficiency it can be assumed that only local explanatory data $X(s_i)$ and neighborhood label L_i are

[3]We would like to thank James Lesage (http://www.spatial-econometrics.com/) for making the matlab toolbox available on the web.

relevant in predicting class label $l_i = f_C(s_i)$. It is common to assume that all interaction between neighbors is captured via the interaction in the class label variable. Many domains also use specific parametric probability distribution forms, leading to simpler solution procedures. In addition, it is frequently easier to work with a Gibbs distribution specialized by the locally defined MRF through the Hammersley-Clifford theorem [4].

Solution Procedures
Solution procedures for the MRF Bayesian classifier include stochastic relaxation [9], iterated conditional modes [3], dynamic programming [8], highest confidence first [6] and graph cut [5]. We have used the graph cut method and provided its description in Appendix I.

4 Comparison of SAR and MRF Bayesian Classifiers

Both SAR and MRF Bayesian classifiers model spatial context and have been used by different communities for classification problems related to spatial datasets. We compare these two approaches to modeling spatial context in this section using a probabilistic framework as well as an experimental framework.

4.1 Comparison of SAR and MRF Using a Probabilistic Framework

We use a simple probabilistic framework to compare SAR and MRF in this section. We will assume that classes $l_i \in (c_1, c_2, \ldots, c_M)$ are discrete and that the class label estimate $\hat{f}_C(s_i)$ for location s_i is a random variable. We also assume that feature values (X) are constant since there is no specified generative model. Model parameters for SAR are assumed to be constant, (i.e., β is a constant vector and ρ is a constant number). Finally, we assume that the spatial framework is a regular grid.

We first note that the basic SAR model can be rewritten as follows:

$$y = X\beta + \rho Wy + \epsilon$$

$$(I - \rho W)y = X\beta + \epsilon$$

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\epsilon = (QX)\beta + Q\epsilon \tag{5}$$

where $Q = (I - \rho W)^{-1}$ and β, ρ are constants (because we are modeling a particular problem). The effect of transforming feature vector X to QX can be viewed as a spatial smoothing operation. The SAR model is similar to the linear logistic model in terms of the transformed feature space. In other words, the SAR model assumes the linear separability of classes in transformed feature space.

Figure 6 shows two datasets with a *salt and pepper* spatial distribution of the feature values. There are two classes, c_1 and c_2, defined on this feature. Feature values close to 2 map to class c_2 and feature values close to 1 or 3 will map to c_1. These classes are not linearly separable in the original feature space. Local spatial smoothing can eliminate the *salt and pepper* spatial pattern in the feature values to transform the distribution of the feature values. In the top part of Figure 6, there are few values of 3 and smoothing revises them close to 1 since most neighbors

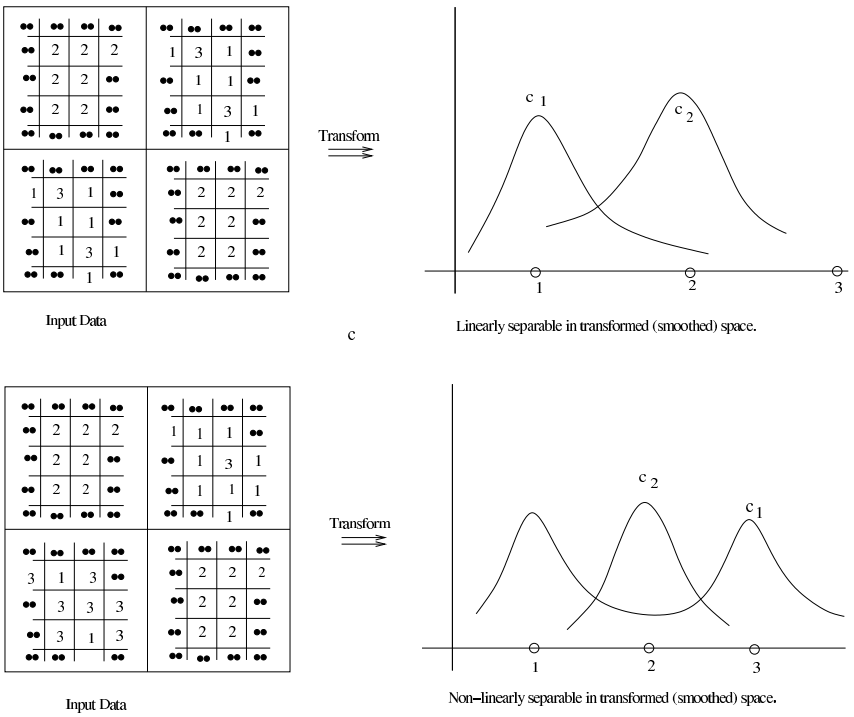


Figure 6: Spatial datasets with *salt and pepper* spatial patterns

have values of 1. SAR can perform well with this dataset since classes are linearly separable in the transformed space. However, the bottom part of Figure 6 shows a different spatial dataset where local smoothing does not make the classes linearly separable. Linear classifiers cannot separate these classes even in the transformed feature space assuming $Q = (I - \rho W)^{-1}$ does not make the classes linearly separable.

Although MRF and SAR classification have different formulations, they share a common goal, estimating the posterior probability distribution: $p(l_i|X)$. However, the posterior for the two models is computed differently with different assumptions. For MRF the posterior is computed using Bayes' rule. On the other hand, in logistic regression, the posterior distribution is directly fit to the data. For logistic regression, the probability of the set of labels L is given by:

$$Pr(L|X) = \prod_{i=1}^{N} p(l_i|X) \tag{6}$$

One important difference between logistic regression and MRF is that logistic regression assumes no dependence on neighboring classes. Given the logistic model, the probability that the binary label takes its first value c_1 at a location s_i is:

$$Pr(l_i|X) = \frac{1}{1 + \exp(-Q_i X\beta)} \tag{7}$$

where the dependence on the neighboring labels exerts itself through the W matrix, and subscript i (in Q_i) denotes the i^{th} row of the matrix Q. Here we have used the fact that y can be rewritten as in equation 5.

To find the local relationship between the MRF formulation and the logistic regression formulation (for the two class case $c_1 = 1$ and $c_2 = 0$), at point s_i

$$\begin{aligned} Pr((l_i = 1)|X, L_i) &= \frac{Pr(X|l_i = 1, L_i)Pr(l_i = 1, L_i)}{Pr(X|l_i = 1, L_i)Pr(l_i = 1, L_i) + Pr(X|l_i = 0, L_i)Pr(l_i = 0, L_i)} \\ &= \frac{1}{1 + \exp(-Q_i X\beta)} \end{aligned} \tag{8}$$

which implies

$$Q_i X\beta = \ln(\frac{Pr(X|l_i = 1, L_i)Pr(l_i = 1, L_i)}{Pr(X|l_i = 0, L_i)Pr(l_i = 0, L_i)}) \tag{9}$$

This last equation shows that the spatial dependence is introduced by the W term through Q_i. More importantly, it also shows that in fitting β we are trying to simultaneously fit the relative importance of the features and the relative frequency ($\frac{Pr(l_i=1,L_i)}{Pr(l_i=0,L_i)}$) of the labels. In contrast, in the MRF formulation, we explicitly *model* the relative frequencies in the class prior term. Finally, the relationship shows that we are making distributional assumptions about the class conditional distributions in logistic regression. Logistic regression and logistic SAR models belong to a more general exponential family. The exponential family is given by

$$Pr(u|v) = e^{A(\theta_v)+B(u,\pi)+\theta_v^T u} \tag{10}$$

where u, v are location and label respectively. This exponential family includes many of the common distributions such as Gaussian, Binomial, Bernoulli, and Poisson as special cases. The parameters θ_v and π control the form of the distribution. Equation 9 implies that the class conditional distributions are from the exponential family. Moreover the distributions $Pr(X|l_i = 1, L_i)$ and $Pr(X|l_i = 0, L_i)$ are matched in all moments higher than the mean (e.g., covariance, skew, kurtosis, etc.), such that in the difference $ln(Pr(X|l_i = 1, L_i)) - ln(Pr(X|l_i = 0, L_i))$, the higher order terms cancel out, leaving the linear term $(\theta_v^T u)$ in equation 10 on the left hand-side of equation 9.

4.2 Experimental Comparison of SAR and MRF

We carried out experiments to compare the classical regression, spatial autoregressive regression and MRF-based Bayesian classifiers. We compared two families of kernel functions, namely the Gaussian Mixture Model (GMM) and Polynomials (P) for MRF-based Bayesian classifiers. We refer to these two families as MRF-GMM and MRF-P respectively.

The goals of the experiments were:

1. To determine whether the real bird habitat datasets follows a Gaussian distribution?
2. To evaluate the effect of including a spatial autoregressive term ρWy in the logistic regression equation.
3. To compare models of spatial context on both real bird habitat datasets and a non-linear simulated synthetic dataset.

The experimental setup is shown in Figure 7. The explanatory variables of bird habitat datasets as described in Section 1.1 were used for the learning portion of the experiments. The dependent

class variable, (i.e., nests), that was used in learning experiments, is of two types, namely real (see Figure 1(a)) and synthetic. Synthetic bird datasets were generated using a non-linear equation 11. All variables in these datasets were defined over a spatial grid of approximately 5000 cells. The 1995 data acquired in the Stubble wetland served as the testing dataset. This data is similar to the learning data except for the spatial locations. We also generated a synthetic dependent class variable Stubble wetlands.

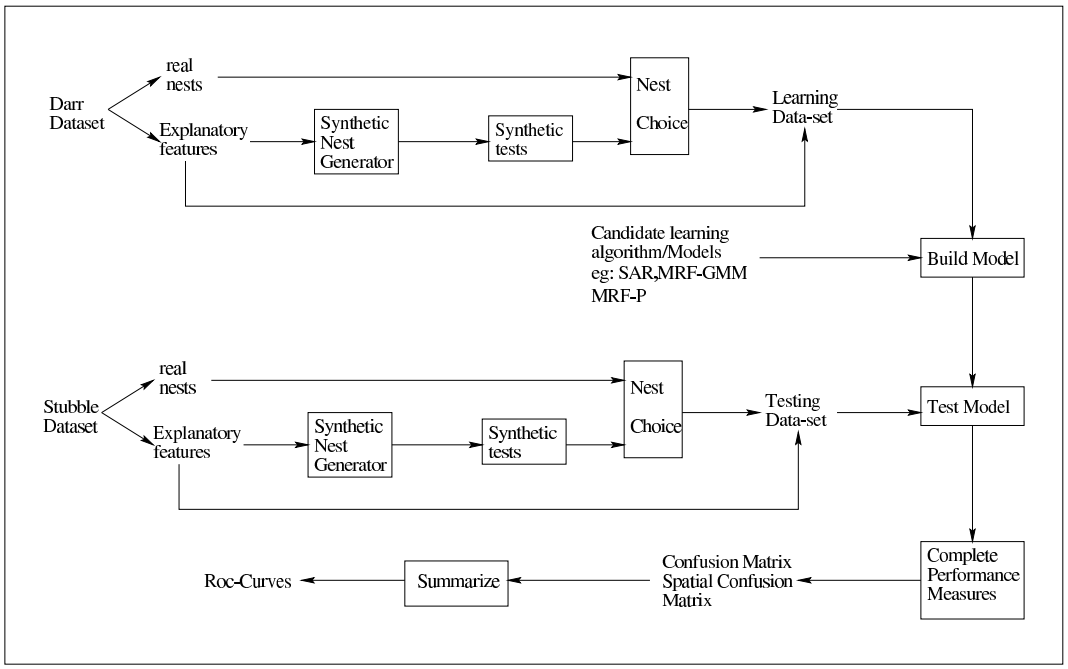


Figure 7: Experimental Method for the Evaluation of SAR and MRF

Metrics of Comparison for Classification Accuracy: Consider boolean vectors, $A_n[i] = f_C[s_i]$ representing actual nest locations, and $P_n[i] = \hat{f}_C(s_i)$ representing predicted nest locations and their inverses, $A_{nn}[i] = 1 - A_n[i]$ and $P_{nn}[i] = 1 - P_n[i]$. The classification accuracy of various measures for such a binary prediction model is summarized in a matrix as shown in Table 2, using the boolean vectors.

	Predicted Nest (Present)	Predicted No-nest (Absence)
Actual Nest (Present)	A_nP_n	A_nP_{nn}
Actual No-nest (Absence)	$A_{nn}P_n$	$A_{nn}P_{nn}$

Table 2: Confusion Matrix

The traditional measure of classification accuracy compares the prediction at location s_i with the actual value at location s_i. This classical measure is not sensitive to the distance between predicted nest and actual nest if the distance is no-zero. We propose new map similarity measures shown in Table 3. The new map similarity measures compare the prediction at location s_i with the actual value at s_i as well as the actual values at neighbors of s_i.

Where A_n is an actual nest, A_{nn} is an actual no-nest, P_n is a predicted nest, P_{nn} is a predicted no-nest, and $M = W + I$ is a matrix addition of a contiguity matrix W and an indentity matrix I. The spatial accuracy measure (SAM) is defined as $SAM = A_nMP_n + A_{nn}MP_{nn}$

	Predicted Nest (Present)	Predicted No-nest (Absence)
Actual Nest (Present)	$A_n MP_n$	$A_n MP_{nn}$
Actual No-nest (Absence)	$A_{nn} MP_n$	$A_{nn} MP_{nn}$

Table 3: Spatial Confusion Matrix

We summarize various accuracy measures in Table 4.

Measure	Definition	Description
ROC Curve	locus of the pair $(TPR(b), FPR(b))$ for each cut-off probability $TPR = \frac{AnPn}{AnPn+AnPnn}$ $FPR = \frac{AnnPn}{AnnPn+AnnPnn}$	The higher the curve above the straight line $TPR = FPR$, the better the accuracy of the model
Total Error (TE)	$TE = AnPnn + AnnPn$	The lower the value of TE,
Classification Acc.(CA)	$CA = \frac{AnPnn+AnnPn}{AnPnn+AnnPn+AnnPn+AnPnn}$	the better the model
Spatial Acc. Measure	$SAM = A_n MP_n + A_{nn} MP_{nn}$	the higher the value of SAM
SAM (Normalized)	$SAMN = \frac{A_n MP_n + A_{nn} MP_{nn}}{A_n MP_n + A_{nn} MP_{nn} + A_n MP_{nn} + A_{nn} MP_n}$	the better the accuracy of the model
ADNP	$ADNP(A,P) = \frac{1}{K} \sum_{k=1}^{K} d(A_k, A_k.nearest(P))$	the lower the value of ADNP, the better the model

Table 4: Definition of Measures

ADNP Measure: An orghogonal measure of spatial accuracy is the Average Distance to Nearest Prediction (ADNP) from the actual nest sites, which is formulated as ADNP(A,P) in Table 4. A_k represents the actual nest locations, P is the map layer of predicted nest locations, and $A_k.nearest(P)$ denotes the nearest predicted nest location to A_k. K is the number of actual nest sites.

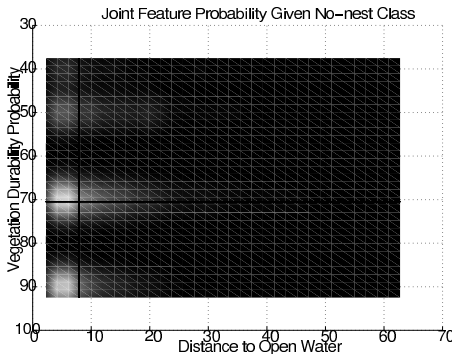
4.3 Experiments with Real Datasets

We used real datasets from Darr and Stubble wetlands for the results presented in this subsection. The explanatory variables and class labels were described in Section 1.1.

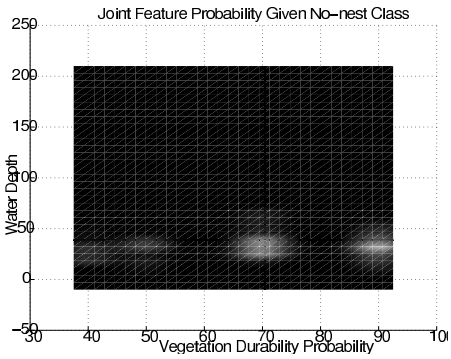
4.3.1 Characterizing the Probability Distribution ($Pr(X|c_i)$)

We analyzed actual wetland datasets to estimate $Pr(X|c_i)$ for the feature values of Vegetation Durability (Veg), Distance to Open Water (DOW) and Water Depth (WD), which were selected as explanatory variables. We explored the statistical probability distribution of each feature given a certain class category (e.g., no-nest class). Figure 8 illustrates the characteristic probability distribution of each feature value given a nest class for the union of real datasets (learning dataset and testing dataset together). We used the "kernel density estimation toolbox" of matlab to fit a smooth function to obtain the observations shown in Figure 8.

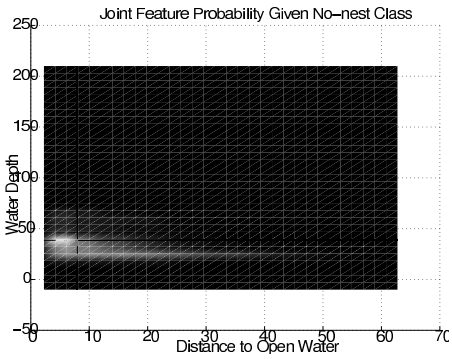
The joint feature probability distribution for a "no-nest" class is displayed in three slices shown in Figure 8(a), (b) and (c). Figure 8(a) shows the slice of the 3-D joint feature probability of Vegetation Durability versus Distance to Open Water given a "no-nest" class when the other feature (Water Depth) is fixed at value 38.6. Figure 8(b) displays the slice of the 3-D joint feature



(a) Joint Feature Probability (Veg vs DOW | No-nest class)



(b) Joint Feature Probability(WD vs Veg | No-nest class)



(c) Probability (WD vs DOW | No-nest class)

Figure 8: Joint feature probability distribution for whole datasets: (a) Pr(Vegetation Durability vs Distance to Open Water | no-nest class), (b) Pr(Water Depth vs Vegetation Durability | no- nest class), (c) Pr(Water Depth vs Distance to Open Water | no-nest class)

probability of Water Depth versus Vegetation Durability given a "no-nest" class when the other feature (DOW) is fixed at value 7.97. The slice of the joint feature probability of Water Depth versus Distance to Open Water given a "no-nest" class when the other feature (Vegetation) is fixed at value 70.45 is shown in Figure 8(c).

It is clear that none of the probability distributions of the real datasets fits a normal distribution, which is a key assumption for regression models (both classical regression and SAR models). However, MRF relaxes this assumption. In the following section, we report some experimental results of a comparison of SAR and MRF on both a real bird habitat dataset and a synthetic bird dataset.

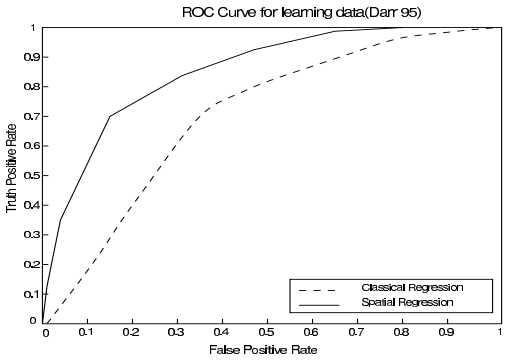
4.3.2 Comparison of Different Models

We built a model using the 1995 Darr wetland data and then tested it on the 1995 Stubble wetland data. In the learning data, all the attributes were used to build the model and in the testing data, one value was hidden, in this case the location of bird nests. Using the knowledge gained from the 1995 Darr data and the value of the independent attributes in the Stubble test data, we predicted the location of the bird nests in Stubble 1995.

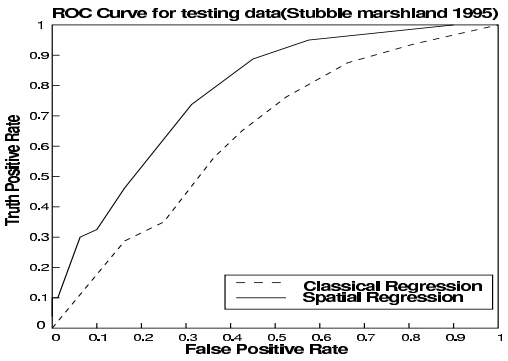
Evaluation of the SAR and Classical Regression Models on Real Datasets: Figure 9(a) illustrates the ROC curves for spatial autoregressive regression (SAR) and classical regression models built using the real 1995 Darr learning data and Figure 9(b) displays the ROC curve for the real 1995 Stubble testing data. It is clear that using spatial regression resulted in better predictions at all cut-off probabilities relative to the classical regression model.

Evaluation of the SAR, MRF-GMM and MRF-P models We also compared several spatial contextual models. Figure 10 illustrates learning and testing results for the comparison between SAR, MRF-GMM, and MRF-P kernel density estimation.

The MRF-P model yields better spatial accuracy and as well as better classification accuracy than MRF-GMM and SAR in both learning and testing experiments. In this real dataset, the



(a) ROC curves for learning



(b) ROC curves for testing

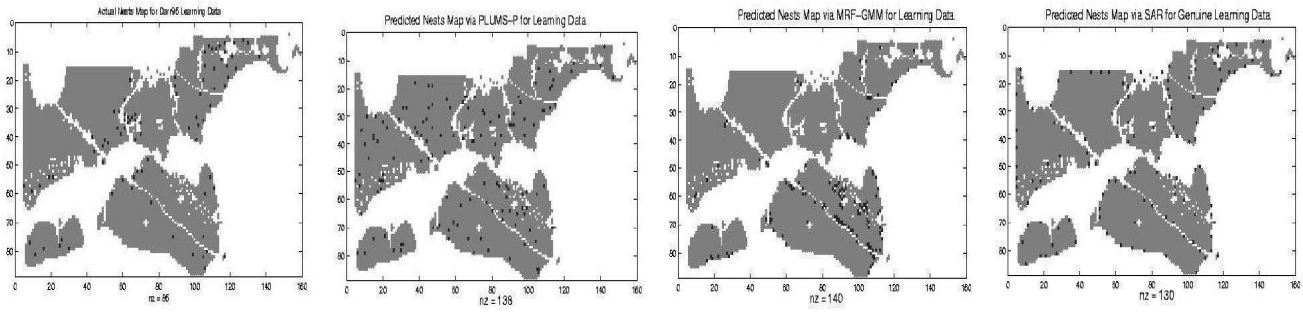
Figure 9: (a) Comparison of the classical regression model with the spatial autoregression model on the Darr learning data. (b) Comparison of the models on the Stubble testing data.

		Learning Data				Test Data			
		Classical Measure		Map Similarity Measure		Classical Measure		Map Similarity Measure	
		Predicted Nest	Predicted No-nest	Predicted Nest	Predicted No-nest	Predicted Nest	Predicted No-nest	Predicted Nest	Predicted No-nest
MRF-P	Actual Nest	42	43	42.78	42.42	9	21	9.36	20.64
	Actual No-nest	96	5191	90.75	5196.2	71	1716	68.36	1718.6
MRF-GMM	Actual Nest	33	52	33.62	51.38	5	25	5.68	24.32
	Actual No-nest	107	5180	97.55	5189.5	73	1714	71.96	1715.0
SAR	Actual Nest	27	58	16.83	68.17	4	26	4.93	25.07
	Actual No-nest	103	5184	107.95	5179.1	76	1711	75.12	1711.8

Figure 10: Error matrix of genuine learning data

prediction accuracies of MRF-GMM and SAR are very compatible.

We also show maps of the predicted nest locations to visualize the results. Figure 11(a) shows the actual nest sites for the genuine learning data (i.e., 1995 Darr bird habitat dataset). Figure 11(b), (c), and (d) shows the predicted nest locations via the MRF-P kernel density estimation, MRF Gaussian mixture model, and the SAR model respectively. From these maps, we can see that MRF-P yields better prediction. The testing maps are shown in Figure 11(e), (f), (g) and (h). The ADNP values for each model prediction were also shown in correponding figure captions. As can be seen, the SAR predictions are extremely localized, missing actual nests over a large part of the Stubble marsh lands. The SAR predictions in Figure 11(d) seem to be concentrated on pixels adjacent to water, (i.e., at a small distance to water). This reliance on a single feature is a problem of linear models such as SAR. This is also reflected in the relatively large (2-3 times larger than those for MRF models) ADNP values for the predictions from SAR model.

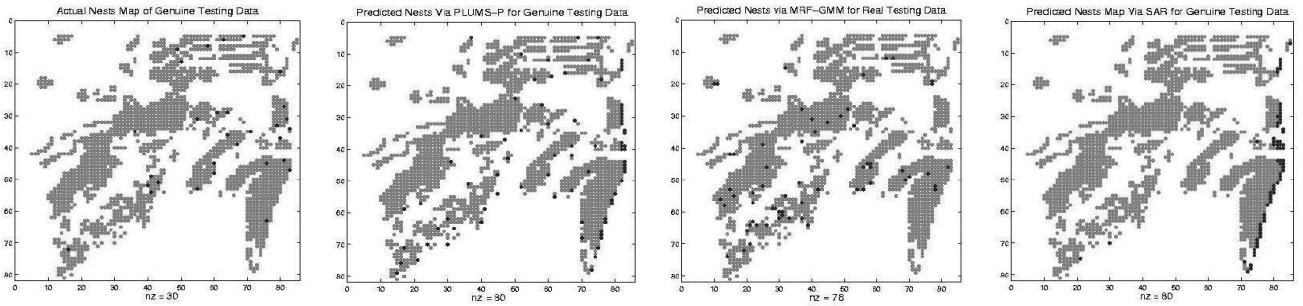


(a) Actual nest sites (Real Learning data)

(b) MRF-R Prediction(ADNP=3.361)

(c) MRF-GMM Prediction (ADNP=5.882)

(d) SAR Prediction (ADNP=9.795)



(e) Nest Locations(Real Testing Data)

(f) MRF-P Prediction (ADNP=2.847)

(g) MRF-GMM (ADNP=3.348)

(h) SAR Prediction (ADNP=8.627)

Figure 11: Predicted nest locations and ADNP values for MRF-P, MRF-GMM, and SAR models

4.4 Non-linear Class Boundary Simulation by Synthetic Bird Datasets

We created a set of synthetic bird datasets based on non-linear generalization. We used the non-linear equation

$$y = (I - \rho W)^{-1} * (\beta * cos(X) + c * random(\epsilon) \tag{11}$$

to generate a set of non-linear class boundaries. X represents the feature values for the independent variables; c is a constant value (we choose 12); $random(\epsilon)$ is a random generated error term; I is the identity matrix; ρ is the spatial co-efficient (we use $\rho = 0.6$ for both the learning and testing synthetic data); and W is the contiguity neighborhood matrix. To generate synthetic non-linear learning data, we used the 1995 Darr wetland feature values for X and the contiguity matrix W, and we made the β values the same as SAR's β value. Similarly, using 1995 Stubble wetlands feature values for X, Stubble 95 contiguity matrix W, and the same β values, we generated a synthetic testing dataset on Stubble 1995. For the non-linear class boundary simulation, we built a model using the non-linear dataset generated using the Darr wetland and then tested it on the non-linear synthetic data generated on the 1995 Stubble wetland data. In the learning stage, all the feature values of the attributes and spatial dependency are used to build the model and in the testing step, one value is hidden, the location of bird nests. Using the knowledge gained from the learning model and the feature values of the explanatory attributes and spatial dependency in the

Stubble test data, we predicted the bird nest locations in the non-linear synthetic data on Stubble 1995.

We carried out experiments on these synthetic bird nesting datasets. Figure 12 presents accuracy results for MRF-P, MRF-GMM and SAR models on the non-linear simulated learning and testing datasets. The confusion matrix shows both classical measure results and map similarity measure results. From Figure 12, we can easily calculate the Total Error (TE) of the classical measure and the spatial accuracy measure (SAM) for the learning model. The total error of MRF-P is $866 + 938 = 1804$, which is significantly less than the total error of MRF-GMM(2151) and SAR (2216). The spatial accuracy measure of MRF-P is $703.74 + 2899.74 = 3603$, which is greater than those of MRF-GMM(3245) and SAR (3162).

		Learning Data				Test Data			
		Classical Measure		Map Similarity Measure		Classical Measure		Map Similarity Measure	
		Predicted Nest	Predicted No-nest	Predicted Nest	Predicted No-nest	Predicted Nest	Predicted No-nest	Predicted Nest	Predicted No-nest
MRF-P	Actual Nest	686	866	703.74	848.26	64	76	69.7	70.3
	Actual No-nest	938	2882	920.26	2899.74	68	1609	62.3	1617.4
MRF-GMM	Actual Nest	522	1030	534.08	1017.92	32	108	37.95	102.05
	Actual No-nest	1121	2699	1108.92	2711.08	81	1596	75.05	1601.55
SAR	Actual Nest	480	1072	485.18	1066.82	21	119	22.74	117.26
	Actual No-nest	1144	2676	1142.7	2677.3	119	1558	117.26	1559.74

Figure 12: Error matrix of the non-linear synthetic learning and testing data generated for Darr95

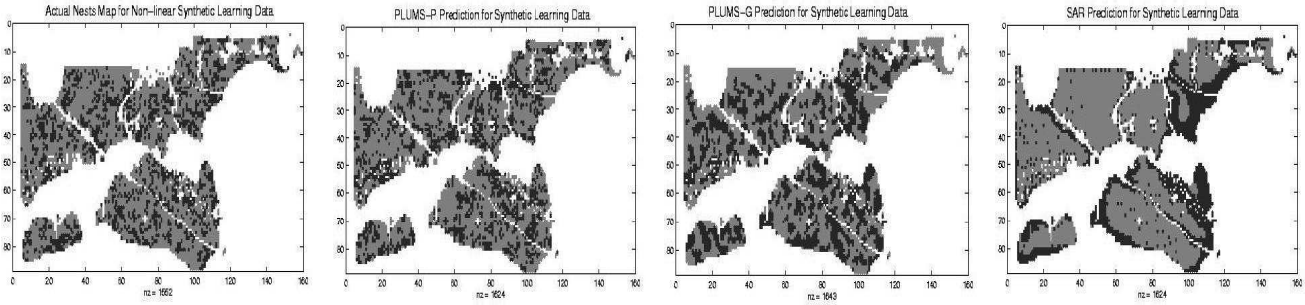
In the non-linear synthetic dataset, MRF-P achieves better spatial accuracy as well as better classification accuracy than MRF-GMM and SAR in both the learning and testing datasets. The prediction accuracy of MRF-GMM is better than that of SAR in both learning and testing.

We also drew maps of the predicted nest locations to visualize the results (see Figure 13). Trends were similar to those observed in Figure 11.

5 Conclusion and Future Work

In this paper we have presented two popular classification approaches that model spatial context in the framework of spatial data mining. We have provided theoretical results using a probabilistic framework and as well as experimental results validating the comparison between SAR and MRF. Our study shows that the SAR model makes more restrictive assumptions about the distribution of features and class shapes (or decision boundaries) than MRF. We also observed an interesting relationship between classical models that do not consider spatial dependence and modern approaches that explicitly model spatial context. The relationship between SAR and MRF is analogous to the relationship between logistic regression and Bayesian Classifiers.

In the future we would like to compare other models that consider spatial context in the classification decision process. We would also like to extend the Graph cut solution procedure for SAR. Finally, we observe that 'precision' and 'recall' [25] for the learning methods were low (i.e., less than 0.5) for nest predictions, even though classification and spatial accuracies are reasonable. We would like to explore techniques to improve 'precision' and/or 'recall'.

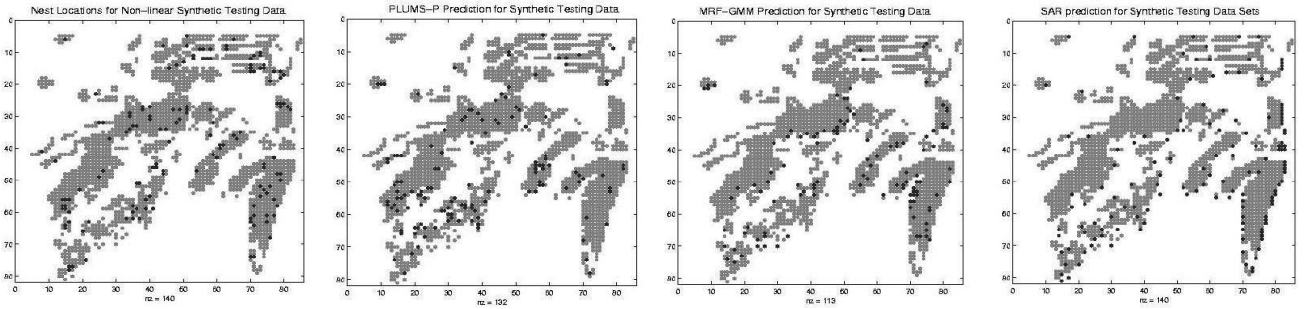


(a) Actual Nest Sites(Non-linear Synthetic Learning Data)

(b) MRF-P Prediction (ADNP=1.353)

(c) MRF-GMM Prediction (ADNP=2.208)

(d) SAR Prediction (ADNP=6.887)



(e) Nest Sites(Synthetic non-linear testing data)

(f) MRF-P Prediction (ADNP=2.400)

(g) MRF-GMM Prediction (ADNP=3.722)

(h) SAR Prediction (ADNP=5.743)

Figure 13: Predicted nest locations and ADNP values for MRF-P, MRF-GMM, and SAR models

5.1 Acknowledgments

We would like to thank our collaborator Uygar Ozesmi for his help and providing the bird habitat datasets. We would like to thank James Lesage for providing matlab toolbox and Chang-Tien Lu, Huang Yan for their useful comments. The comments of Kimberly Koffolt have greatly improved the readability of this paper.

References

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

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

[3] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Statistical Soc.*, (48):259–302, 1986.

[4] J.E. Besag. Spatial Interaction and Statistical Analysis of Latice Systems. *Journal of Royal Statistical Society, Ser. B (Publisher: Blackwell Publishers)*, 36:192–236, 1974.

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *International Conference on Computer Vision*, September 1999.

[6] P.B. Chou, P.R. Cooper, M. J. Swain, C.M. Brown, and L.E. Wixson. Probabilistic network inference for cooperative high and low levell vision. In *In Markov Random Field, Theory and Applicaitons*. Academic Press, New York, 1993.

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

[8] H. Derin and H. Elliott. Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, (9):39–55, 1987.

[9] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[10] C. Greenman. Turning a map into a cake layer of information. *New York Times*, January 20th (http://www.nytimes.com/library/tech/00/01/ circuits/arctiles/20giss.html) 2000.

[11] S. Haykin. *Neural Networks - A Comprehensive Foundation*. MacMilan, ISBN-002-352761-7, 1994.

[12] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, 1989.

[13] Yonhong Jhung and Philip H. Swain. Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 34(1):67–75, 1996.

[14] 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.

[15] J. LeSage. Regression Analysis of Spatial data. *The Journal of Regional Analysis and Policy (Publisher: Mid-Continent Regional Science Association and UNL College of Business Administration)*, 27(2):83–94, 1997.

[16] J. P. LeSage and R.K. Pace. Spatial dependence in data mining. In *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, forthcoming, 2001.

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

[18] S. Li. Markov Random Field Modeling. *Computer Vision (Publisher: Springer Verlag*, 1995.

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

[20] Jim Melton and Andrew Eisenberg. Sql multimedia and application packages (sql/mm). *SIGMOD Record*, 30(4), December 2001.

[21] 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.

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

[23] R. Pace and R. Barry. Quick Computation of Regressions with a Spatially Autoregressive Dependent Variable. *Geographic Analysis*, 1997.

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

[25] B. Ribeiro-Neto R. Baeza-Yates. *Modern Information Retrieval*. ACM Press and Addision-Wesley, 1999.

[26] 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.

[27] W.S. Sarle. Neural Networks and Statistical Models. In *Proceeding of 9th Annual SAS user group conference*. SAS Institue, 1994.

[28] S. Shekhar, S. Chawla, S. Ravada, A.Fetterer, X.Liu, and C.T. Lu. Spatial databases: Accomplishments and Research Needs. *IEEE Transactions on Knowledge and Data Engineering*, 11(1), Jan-Feb 1999.

[29] A. H. Solberg, Torfinn Taxt, and Anil K. Jain. A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transaction on Geoscience and Remote Sensing*, 34(1):100–113, 1996.

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

[31] C. E. Warrender and M. F. Augusteijn. Fusion of image classifications using Bayesian techniques with Markov rand fields. *International Journal of Remote Sensing*, 20(10):1987–2002, 1999.

6 Appendix: Solving Markov Random Fields with Graph Partitioning

Markov Random Fields (MRFs) generalize Markov Chains to multi-dimensional structures. Since there is no natural order in a multi-dimensional space, the notion of a transition probability matrix is absent in MRFs.

MRFs have found applications in image processing and spatial statistics, where they have been used to estimate spatially varying quantities like intensity and texture for noisy measurements. Typical images are characterized by piece-wise smooth quantities, i.e, they vary smoothly but have sharp jumps(discontinuities) at the boundaries of the homogeneous areas. Because of these discontinuities the least-square approach does not provide an adequate framework for the estimation of these quantities. MRFs provide a mathematical framework to model our *a prioir* belief that spatial quantities consist of smooth patches with occasional jumps.

We follow the approach suggested in[5], where it is shown that the maximum a posteriori estimate of a particular configuration of an MRF can be obtained by solving a suitable min-cut multiway graph partitioning problem. We will formally describe this approach later in the appendix but first we illustrate the underlying concept with some examples.

Example 1: A Classification Problem With No Spatial Constraints

Even though MRFs are inherently multi-dimensional we will use a simple one-dimensional example to illustrate the main points. Consider the graph $G = (V, E)$ shown in Figure 14(a). The node-set V itself consists of two disjoint sets, S and C. The members of S are $\{s_1, s_2, s_3\}$ and the members of C are $\{c_1, c_2\}$. Typically the $X(s_i)$'s are the feature values at site s_i and the $c_i's$ are the labels, like *nest* or *no-nest*. There is an edge between each member of the set S and each member of set C. Here we interpret the edge weights as probabilities. For example, $p_1 = Pr(X(s_1) = c_1) = 0.7$ and $p_2 = Pr(X(s_1) = c_2) = 0.3$; $p_1 + p_2 = 1$.

Our goal is to provide a *label* for each location s_i in S using explanatory feature $X(s_i)$. This is done by partitioning the graph into two disjoint sets (not S and C) by removing certain edges such that:

1. There is a many-to-one mapping from the set S to C. Every element of S must be mapped to one and only one element of C.
2. Multiple elements of C cannot belong to a single partition. Thus there are no edges between elements of C and therefore the number of partitions is equal to the cardinality of C, and
3. The sum of the weights of the edges removed (the cut-set) is the minimum of all possible cut-sets.

In this example the cut-set is easily determined. For example, of the two edges connecting each element of S and an element of C, remove the edge with the *smaller* weight. Figure 14(b) shows the graph with the cut-set removed. Thus we have just shown that when the weights of the edges are interpreted as probabilities, the min-cut graph partition induces a maximum *a posterior* (MAP) estimate for the pixel labels. We prefer to say that the *min-cut induces a Bayesian classification* on the underlying pixel set. This is because we will use Bayes' theorem to calculate the edge weights of the graphs.

Example 2: Adding Spatial Constraints In the previous example we did not use any information

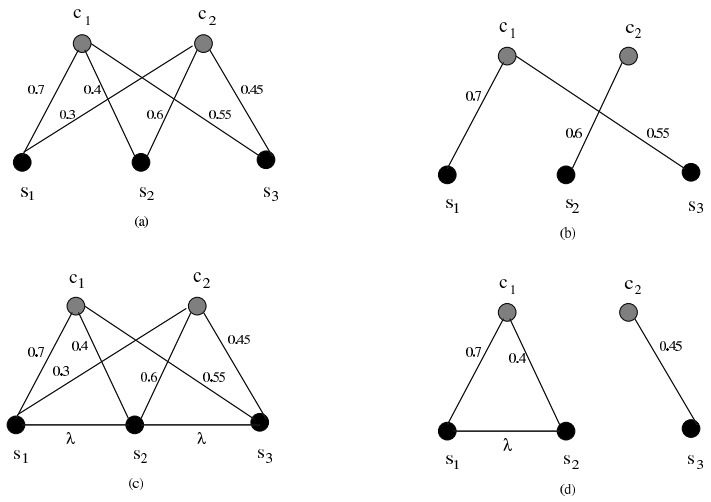


Figure 14: MRF solution with graph-cut method: (a) Initially each pixel is assigned to both labels with different edge weights. The edge weights correspond to probabilities about assigning each pixel to a different label, (b) A min-cut graph partitioning induces a labeling of the pixel set. Labels which correspond to the maximum probabilities are retained, (c) `Spatial autocorrelation` is modeled by introducing edges between pixel nodes, (d) A min-cut graph partitioning does not necessarily induce a labeling where the labeling with maximum probabilities are retained. If two neighboring pixels are assigned different labels, then the edge connecting the pixels is added to the cut-set.

about the spatial proximity of the pixels relative to each other. We do that now by introducing additional edges in the graph structure.

Consider the graph shown in Figure 14(c) in which we have added two extra edges (s_1, s_2) and (s_2, s_3) with a weight λ. In this example we have chosen $\lambda = 0.2$.

Now if we want to retain the same partitions of the graph as in Example 1, then the cut-set has two extra edges, namely (s_1, s_2) and (s_2, s_3). Thus the sum of the weights of the edges in the cut-set,W_{C1}, is

$$W_{C1} = 0.3 + 0.4 + 0.45 + 2\lambda$$

But now, depending upon λ, the cut-set weight may not be minimal. For example, if $\lambda = 0.2$ then the weight of the cut-set, W_{C2}, consisting of the edges $\{(s_1, c_2), (s_2, c_1), (s_3, c_1), (s_1, s_2)\}$ is

$$W_{C2} = 0.3 + 0.4 + 0.55 + 0.2$$

Thus $W_{C2} < W_{C1}$. What is happening is that if two neighboring pixels are assigned to different labels, then the edge between the two neighbors is added to the cut-set. Thus there is a penalty associated with two neighboring nodes being assigned to different labels every time. Thus we can model `spatial autocorrelation` by adding edges between the pixel nodes of the graph. We can also model `spatial heterogeneity` by assigning different *weights*, the λ's to the pixel edges.

Formal Description Using the terminology introduced in [5], we now formalize the observations made in the above two examples. Again, consider a graph $G = (V, E)$ with non-negative edge weights. The set V consists of two types of nodes, *pixels* and *labels*. We will denote the set of pixels as S and the set of labels as C. There are two types of edges too: *n-links* and and *l-link*. An n-link connects two pixels and an l-link connects a pixel with a label. There are no edges between labels. The *l-link* (c_i, s_j) essentially represents the conditional probability $Pr(l_j = c_i | X(s_j))$.

Definition: A set $K \subset E$ is a *multi-way cut* if the label nodes C are completely separated in the graph $G(K) = (V, E - K)$. The sum of the weights of edges in the cut-set K is denoted as $|K|$. A cut-set is a *min cut-set* if its weight is the minimum of all possible cut-sets.

Definition: A cut-set is *feasible* if it induces a many-to-one mapping from S to C and no elements of C can belong to the same set. (From now on we will only consider feasible cut sets)

Lemma 1 If a graph G (as defined above) has no *n-links* and the weights on the *l-links* are the *posteriori* probabilities $Pr(c_i|s_j)$ then the min-cut induces a Bayesian classification on the pixel set S.

Proof: A cut set K induces a graph in which each pixel is assigned to one and only one label. Thus every cut-set induces a classification f on the pixel set S. Now

$$|K| = \sum_{s_j \in S} \sum_{c_i \in C, c_i \neq f(s_j)} Pr(f(s_j) = c_i | X(s_j))$$

Thus

$$\min_f |K| = \min_f \sum_{s_j \in S} \sum_{c_i \in C, c_i \neq f(s_j)} Pr(f(s_j) = c_i | X(s_j)) = \sum_{s_j \in S} \min_f \sum_{c_i \in C, c_i \neq f(s_j)} Pr(f(s_j) = c_i | X(s_j))$$

We can pass the minimum through the first summation because there are no n-links and the cut-sets are feasible. Now for a given $s_j \in S$

$$\sum_{c_i \in C} Pr(f(s_j) = c_i) = 1$$

Therefore

$$\sum_{s_j \in S} \min_f \sum_{c_i \in C, c_i \neq f(s_j)} Pr(f(s_j) = c_i | X(s_j)) = \sum_{s_j \in S} \min_f (1 - Pr(f(s_j) = c_i)$$

The last term is minimized when we choose the maximum probabilities, $Pr(f(s_j) = c_i)$ for each $s_j \in S$. Therefore $min|K|$ induces a classifier f which corresponds to the Bayesian classification of the pixel set S, since Bayes' rule was used to determine the edge weights, $(s_j, c_i) = Pr(f(s_j) = c_i)$. The classification f minimizing $|K|$ is chosen as $(\hat{f}_c)$ solution to location prediction problem.

Definition A *Neighborhood System* N of a multi-way graph G, as defined above, consists of all unordered pixel pairs $\{s_i, s_j\}$ such that there is an *n-link* between s_i and s_j. $N(s_i)$ consists of all pixels in G which are *n-linked* to s_i.

Definition Let f be classifier on the pixel set S of a graph G. Then the energy E associated with f is defined as

$$E(f) = \sum_{s_j \in S} \sum_{c_i \in C, c_i \neq f(s_j)} Pr(f(s_j) = c_i | X(s_j)) + \frac{\lambda}{2} \sum_{s_j \in S} \sum_{s_k \in N(s_j)} (1 - \delta(f(s_j) - f(s_k)))$$

where δ is the impulse function such that

$$\delta(s_j - s_k) = \begin{cases} 1 & \text{if } s_j = s_k \\ 0 & \text{if } s_j \neq s_k \end{cases}$$

Lemma2: Let G be a graph, as defined above, where the weights of the *l-links* are $Pr(f(s_j) = c_i|X(s_j))$ and the weights of the *n-links* are λ. Then a min-cut set of G induces a classifier f on S which minimizes the energy function E.

Proof: By construction of the graph G. The weight of the cut-set is E. A min-cut induces an f which minimizes E.

Minimizing E is equivalent to a MAP estimate of the MRF model [5].

How Are the Edge-weights of the Graph Generated?
We use a training set in conjunction with Bayes theorem to generate the edge weights of the *t-links* of the graph. In general the labels of the pixels are not directly observable (that is what we want to calculate), but we do have an estimate of the "independent" variables, Y. Thus given a label set C and an observation X at s_j, we can compute the required posteriori $Pr(c_i|X(s_j))$ using Bayes' formulae.