An Improved Model for Spatially Correlated Binary Responses

Jennifer A. Hoeting        Molly Leecaster        David Bowden*
Colorado State University

September 1, 1999 (revision)

Technical Report 9719 (revised)
Department of Statistics
Colorado State University
To appear in Journal of Agriculture, Biological, and Environmental Statistics

Abstract

In this paper we use covariates and an indication of sampling effort in an autologistic model (Besag, 1974) to improve predictions of probability of presence for lattice data. The model is applied to sampled data where only a small proportion of the available sites have been observed. We adopt a Bayesian set-up and develop a Gibbs sampling estimation procedure. In four examples based on simulated data, we show that the autologistic model with covariates improves predictions as compared to the simple logistic regression model and the basic autologistic model (without covariates). Software to implement the methodology is available at no cost from StatLib.

Keywords: Autologistic model, Bayesian estimation, Gibbs sampling, Markov random field.

*Jennifer Hoeting is Assistant Professor, Molly Leecaster is a Doctoral Candidate, and David Bowden is Professor at Department of Statistics, Colorado State University. Address correspondence to Jennifer Hoeting, Colorado State University, Fort Collins, CO 80523, email: jah@stat.colostate.edu, webpage: www.stat.colostate.edu/~jah
1 Introduction

It is now common practice to use modern techniques to collect spatial data including using satellites to collect data and global positioning systems (GPS) to determine the location of each sample. One choice for modeling the data collected using these methods is the autologistic model, a spatial model for binary response data. The model improves predictions for data measured with error, over a lattice of spatially correlated responses. Applications of this type of model have extended to agricultural research (Gumpertz et al., 1997), forestry (Preisler, 1993), archaeology (Besag et al., 1991), and biological range mapping (Heikkinen and Högnander, 1994; Högnander and Møller, 1995). In many of these applications, there is more information available than that utilized by the basic autologistic model. In addition, there is often only partial information available about the response of interest. This is the case when a sample of sites is collected over an area of interest.

In this paper we apply an autologistic model with covariates which allows the user to fully utilize data from a sparsely sampled area of interest for prediction and inference over the complete area. To improve predictions, the model incorporates other covariates measured over the area of interest. We adopt a Bayesian setup and present a Gibbs sampling method for estimation.

Our current work is motivated by a project for the United States Forest Service (USFS) to develop maps of likelihood of presence/absence for plant species. In an effort to conserve diversity, the USFS is developing management strategies for lands which have known habitat for rare species. The goal is a map of predicted probabilities of presence (as in the bottom row of plots in Figure 1). These maps, used by Besag et al. (1991) and in other later applications of the autologistic model (see references below), can be used for policy or they can be used to develop further sampling plans. For example, the results shown in Figure 1 could be used to develop a probability sample.

The species of interest to the USFS for the current project, mostly mosses, lichens, and fungi, are assumed to occur in clusters. The available information includes habitat data from GIS and ground surveys at each site in the area of interest (i.e., every pixel in the image), sighting information at some sites, and search intensity at each site. Note that observed presence or absence of species is not known for every site. Search intensity is used to discriminate between sites with observed absence and unobserved sites. We describe methodology to predict presence/absence of one species based on local spatial information, habitat data, search intensity, and observed presence or absence at a sample of sites. The proposed methodology produces an estimate for each pixel in the region of interest as opposed
to kriging or spline-based methods which produce a continuous prediction surface. We use
the language of species presence/absence as our work to date has focused on the project
for the USFS; however, the models proposed here are fully applicable to image analysis and
other spatial problems.

The basic autologistic model was developed by Besag (1972; 1974) to model binary two-
dimensional rectangular lattice scenes. This model uses presence/absence in neighboring
sites to predict probability of presence/absence at the given site. Besag (1974) modeled data
on plant presence/absence. An extension to the autologistic model incorporating search
intensity was developed by Heikkinen and Högmander (1994), and Högmander and Møller

The likelihood for the autologistic model is analytically intractable, except in trivial
cases, so alternative estimation methods are necessary. Besag continued work using esti-
mation techniques such as a coding method and pseudolikelihood methods (1974; 1986).
Augustin et al. (1996) and Preisler (1993) use pseudolikelihood approaches for parameter
estimation. Wu and Huffer (1997) use a Markov chain Monte Carlo (MCMC) to approximate
the maximum likelihood estimates of the parameters. Bayesian approaches to estimate the
parameters of the basic autologistic model have been developed by several authors (Besag,

There are a broad range of applications for the autologistic model (Besag, 1974; Besag,
1986; Preisler, 1993; Heikkinen and Högmander, 1994; Högmander and Møller, 1995;
Augustin et al., 1996; Gumpertz et al., 1997; Wu and Huffer, 1997). Besag et al. (1991) applied
the basic autologistic model to predict missing observations at nine out of 256 sites. Augustin
et al. (1996; 1998) apply a model that is similar to the model used here, but they do
not use a Bayesian approach for estimation. Other recent work on autologistic models with
covariates (Wu and Huffer, 1997; Gumpertz et al., 1997) focuses on parameter estimation
and interpretation. Instead of focusing on parameter estimation, our work focuses more on
the accuracy of the image reconstruction and in this sense is more similar to image analysis.

The autologistic model with covariates assumes site dependence within neighborhoods
and uses covariates to improve prediction. We compare the autologistic model with covariates
to models which utilize only one of these sources of information. The logistic regression model
uses covariates but assumes the observations to be independent. The basic autologistic model
assumes neighborhood dependence but does not accommodate covariate information. In the
examples we consider we show that the autologistic model with covariates makes improved
predictions as compared to both of these models. Thus the autologistic model with covariates
will allow for improved monitoring of rare species.
In Section 2 we describe the autologistic model with covariates and our Bayesian setup. In Section 3 we develop the estimation procedure. Section 4 contains several examples based on simulated data. In Section 5 we conclude and discuss future directions.

2 Autologistic model with covariates

We begin by introducing some notation. There are $n$ sites, or pixels, in a finite lattice. For site $i$, true presence or absence is denoted by $x_i = 1$ or $0$, respectively. Also, for site $i$, $y_i$ denotes the observed response, where $y_i = 0$ indicates observed absence or incomplete information and $y_i = 1$ indicates observed presence, and $a_i$ denotes the level of search intensity. In our examples below there are two search intensities, $a_i = 1$ for sites that were searched and $a_i = 0$ for sites not searched. In many applications sampling is done at more than two levels of intensity. For instance, line transect samples assume high intensity on the line and decreasing intensity of search as distance from the line increases. For these sampling scenarios, search intensity, $a_i$, would take on more than two values. Throughout the paper we use $p()$ to denote a probability distribution function and $\pi()$ to denote a prior distribution.

2.1 Assumptions

Several assumptions underlie our analysis. For a complete discussion see Besag (1974) or Cressie (Section 6.4, 1993). First, we assume observations are mutually independent given true presence or absence at each site. So the likelihood ($\ell$), the joint distribution of observed presence/absence given true presence absence over all sites, is defined

$$\ell(y \mid \underline{x}) = \prod_{i=1}^{n} p(y_i \mid x_i) \tag{1}$$

where $p(y_i \mid x_i)$ is the distribution of observed presence/absence given true presence absence at site $i$. Second, we use a locally dependent Markov random field (LDMRF) to model the true image, $\underline{x}$. The probability of true presence or absence at each site given all other sites is the distribution of $x_i$ given true presence or absence at neighboring sites,

$$\Pr(x_i \mid \underline{x}_{-i}) \equiv p(x_i \mid \underline{x}_{\delta_i}),$$

where $\underline{x}_{-i}$ is the vector of all sites except site $i$ and $\delta_i$ denotes some neighborhood of $i$, defined below.

The final assumption is called the positivity condition. For this application, the positivity condition requires that all possible configurations of presence/absence must be theoretically
possible. This assumption is reasonable for the types of problems considered in this paper. It would not be a reasonable assumption if, say a particular species was territorial, so observing presence in one site would make it (virtually) impossible to observe other individuals in neighboring sites. These territorial species would not be the “clumping” type species that we are seeking to model in this paper.

2.2 Model Statement

The autologistic model with covariates is defined

$$
Pr(x_i = 1 | \mathbf{x}_i, \theta, \beta) = \frac{\exp\{z_i^T \theta + \beta s(x_i)\}}{1 + \exp\{z_i^T \theta + \beta s(x_i)\}},
$$

where $z_i$ is a vector of covariates for the $i$th site with the first element equal to 1, $\theta$ is a vector of parameters for the covariates, and $\beta$ is the parameter associated with the spatial covariate, $s(x_i)$. The spatial covariate for site $i$, $s(x_i)$, is equal to the total number of sites where the species was present in the neighborhood of site $i$. Since $s(x_i)$ is the autocovariate, we have chosen to treat this term separately from other covariates for ease of understanding. A neighborhood could be defined as first order, which is the set of pixels directly north, south, east and west of the pixel of interest, or second order, which also includes the sites diagonal from the site of interest. A proper neighborhood must meet the condition that if site $i$ is a neighbor of site $j$, then $j$ must be a neighbor of $i$. To avoid difficulties with edge effects, only non-edge sites are used to estimate the parameters.

2.3 Bayesian Setup

We adopt a Bayesian approach to estimate the parameters of the model. Our approach builds upon the work of Heikkinen and Högmander (1994) and Högmander and Møller (1995), who use the spatial autocovariate and search intensity but no other covariates. We work with incomplete observations; sampled responses taken over the area of interest. Our data consist of sites that have observed presence, sites which have observed absence, and sites which were not sampled. We also incorporate habitat covariates, measured at every site. We assume that sampled sites are observed without error, so if the species is observed at site $i$, $y_i = 1$, we predict presence at that site with probability 1. In our two level search intensity scenario with fixed observations, if we observe absence then $y_i = 0$, $a_i = 1$, and we predict absence at site $i$ with probability 1.
The likelihood is defined in equation 1 where
\[
p(y_i = 0 \mid x_i, a_i) = \begin{cases} 
1, & (a_i = 0) \text{ or } (x_i = 0 \text{ and } a_i = 1), \\
0, & (x_i = 1 \text{ and } a_i = 1). 
\end{cases}
\]
We necessarily observe \( y_i = 0 \) if the species is truly absent or if we do not search the site. The likelihood defined in this way, limits possible true scenes to those consistent with our observed sample. For cases where observations are made with error, the likelihood could involve a parameter for each level of search. For sampled sites with level \( e \) of search intensity, \( \gamma_e \) would denote the likelihood of not detecting a species which is actually present. This approach can readily be extended to more than two levels of search. If detectability is believed to be related to some other measurable covariates, this could be modeled in the likelihood. The logit of the likelihood would be a linear function of these covariates.

The other components of the hierarchical model are defined as follows. The image prior, based on the LDMRF is
\[
\pi (\boldsymbol{x} \mid \theta, \beta) \propto \exp \left\{ \theta^T \sum_{i=1}^{n} x_i + \beta \sum_{i<j} x_i x_j \right\}, \tag{3}
\]
where \( I_{i[j]} = 1 \) when pixel \( i \) and \( j \) are neighbors and 0 otherwise. The normalizing constant for equation 3 is the sum over all possible configurations of presence and absence. See Huffer and Wu (1998) and Besag (1974) for additional details. The hyperprior for \( \theta \) is assumed to follow a normal distribution, \( \pi (\theta) \sim N(0, \Sigma) \), where \( \Sigma \) is a diagonal covariance matrix with \( (\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_p) \) on the diagonal, which are hyperparameters to be chosen. The covariate coefficients, \( \theta \), are assumed to be independent a priori. When the predictors include a set of indicator (dummy) variables for a categorical variable with \( c \) categories, the posterior distribution should be invariant to the choice of \( c - 1 \) indicator variables. Raftery, Madigan, and Hoeting (1997) suggest an appropriate prior distribution set-up for dummy variables. Finally, the hyperprior for the spatial parameter is defined \( \pi (\beta) \sim \text{Gamma} (\psi, \alpha) \), where \( \psi \) and \( \alpha \) are hyperparameters to be chosen. Thus the spatial parameter is constrained to be greater than 0. In the examples we have considered to date, a negative spatial parameter was not reasonable since we assume that species are clustered.

Under this set-up the posterior distribution is defined as
\[
p (\boldsymbol{x}, \theta, \beta \mid \boldsymbol{y}) \propto \ell (\boldsymbol{y} \mid \boldsymbol{x}) \pi (\boldsymbol{x} \mid \theta, \beta) \pi (\theta) \pi (\beta).
\]
This posterior is analytically intractable due to the normalizing constant in \( \pi (\boldsymbol{x} \mid \theta, \beta) \).
3 Estimation

We use a Gibbs sampler (Geman and Geman, 1984) to carry out the estimation. There are two main steps in this Gibbs sampler. First we update \( x_i \). At the \( t \)-th iteration, we generate \( x_i^t \mid x_{-i}^{t-1}, \theta^{t-1}, \beta^{t-1}, a_i \sim \text{Bernoulli}(p_{i}^{t-1}) \) where

\[
p_{i}^{t-1} = \left( \frac{\exp\{z_i^T \theta^{t-1} + \beta^{t-1}s(x_i^{t-1})\}}{1 + \exp\{z_i^T \theta^{t-1} + \beta^{t-1}s(x_i^{t-1})\} } \right)^{(1-a_i)} \left(x_i^{t-1}\right)^{a_i},
\]

and values superscripted with \( t-1 \) are values from the previous iteration. We update the \( x_i \) in independent blocks, so that the \( x \) values within each set are conditionally independent. The blocking structure is dependent upon the chosen neighborhood structure. This approach of using “coding sets” (Besag, 1974) speeds up the updating process. To speed up convergence we compute the spatial covariate using the predicted probability of presence for the unobserved sites, so \( s \left( p_{i}^{t-1} \right) \) replaces \( s \left( x_i^{t-1} \right) \) in equation 4. Augustin et al. (1996; 1998) use a similar approach in their estimation algorithm.

In the second step we update the parameters for the covariates. The full conditional distribution for each parameter is analytically intractable, so we use a Hastings-Metropolis step similar to Heikkinen and Högmander (1994) to approximate the distributions of the parameters. Following their approach, we update the parameters using a pseudolikelihood approximation instead of the full conditional likelihood for the parameters. The pseudolikelihood approximation is defined over \( S \), the area of interest, as

\[
pl(\beta \mid x, y, \theta) \propto \prod_{i \in S} \pi(x_i \mid x_{-i}, \theta, \beta) \pi(\beta),
\]

and

\[
pl(\theta_j \mid x, y, \theta_{-j}, \beta) \propto \prod_{i \in S} \pi(x_i \mid x_{-i}, \theta_{-j}, \beta) \pi(\theta_j).
\]

Under regularity conditions given by Cox and Hinkley (1974, Sec 10.6, p 399), the log pseudolikelihoods are asymptotically Gaussian with means \( \hat{\beta} \) and \( \hat{\theta} \), the maximum pseudolikelihood estimates, and variances from the Fisher Information of the pseudolikelihoods. These results hold under “increasing-domain asymptotics” (Cressie, Section 7.3.1, 1993), i.e., increasing the number of pixels of the same dimension in an infinite sample space. We use these Gaussian proposal distributions for the Hastings-Metropolis steps. Nonlinear maximization is required to obtain the pseudolikelihood estimates for the proposal distributions at each iteration. Thus, in the Hastings-Metropolis step, the maximum pseudolikelihood estimates are obtained, then a value is sampled from the proper proposal distribution,

\[
\beta' \sim N \left( \hat{\beta}^{t-1}, \sigma_{\beta}^{2t-1} \right),
\]

and
\[ \theta' \sim N \left( \hat{\theta}_{t-1}^{\iota}, \sigma^2_{t-1} I \right). \]

and finally that value is accepted or rejected based on the ratio of densities. To improve the efficiency of the algorithm, we multiply the variance of the proposal distribution by \(2.38^2\) which was recommended by Gelman et al. (1996). This increases the mobility of the sampler in the parameter space. Finally, the parameters for the covariates are updated in random order at each iteration to avoid order bias in the parameter estimates.

4 Examples

4.1 Simulation Setup

We use simulated data to evaluate the performance of our methodology. These simulated data are treated as the truth, which is usually unknown. The model is evaluated on its ability to reproduce the truth based on presence/absence information at a sample of sites and covariate information over the area of interest. The simulations are constructed to test performance under various conditions. We present four examples which demonstrate the model’s performance under different covariates, population densities, and sampling patterns. The predictions from the autologistic model with covariates are compared with the predictions from the basic autologistic model and the logistic regression model.

The simulations were performed on \(50 \times 50\) grids, for a total of 2500 sites in the area of interest. The true scenes and covariates were generated using various simulation procedures described below.

We specified inputs to the Gibbs estimation procedure for all examples as follows. Hyperparameters were chosen based on pilot runs. For \(\beta\), the spatial parameter, we chose a broad gamma distribution \((\psi, \alpha) = (3, 2)\), with most of its mass between 0 and 15. For each covariate parameter, \(\theta_j\), we used a vague prior, \(N(0, 20)\). Sensitivity analysis demonstrated the model to be robust to these inputs. We used a second order neighborhood for the LDMRF, so interior sites have eight neighbors. We found that the first order neighborhood results were less consistent especially for small sample sizes. We used 20,000 iterations for the Gibbs sampler with a burn-in of 1000. The maps of predicted probability of presence are reported as the mean probability of presence at each site after the burn-in.

Four examples follow. Several images are provided for each example: the true scene, the observed scene, the covariates, and the prediction scenes from the logistic regression model, from the autologistic model, and from the autologistic model with covariates. In the true scene black pixels indicate presence. In the observed scenes nonsampled sites are
light grey, sampled sites where species are not present are medium grey, and sampled sites where species are present are black. For the prediction scenes, darker shades of grey indicate that the posterior probability of presence is closer to one, or the species is more likely to be present.

For each example, we compared the performance of the three models using receiver operating characteristic (ROC) curves (Egan, 1975). ROC curves show the true positive rate (TPR) versus the false positive rate (FPR) for different thresholds of prediction. For our problem, TPR and FPR are defined

\[
TPR = \frac{\sum I_{[x_i > b]} I_{[y_i = 1]}}{\sum_{i=1}^{n} x_i},
\]

\[
FPR = \frac{\sum I_{[x_i > b]} I_{[y_i = 0]}}{\sum_{i=1}^{n} (1 - x_i)},
\]

where \(0 \leq b \leq 1\). So, TPR is equal to the number of sites where the estimated probability of presence is greater than the cut-off value, \(b\), for the sites where the species are truly present \((x = 1)\) divided by the total number of sites where the species are truly present. FPR is interpreted similarly for sites where the species are truly absent. The greater the height of the ROC curve above the FPR=TPR line, the better the method discriminates between true presence and true absence.

### 4.2 Simulation Results

The first example (Figure 1) was simulated from an autologistic model with two covariates. The first covariate \((Z1)\), generated from a uniform(0,1) distribution, is random noise. The second covariate \((Z2)\) has four distinct clumps (generated from a uniform(5,10) distribution) on a background generated from a uniform (0,20) distribution. To simulate the true scene \((X)\), a random pattern generated via a Bernoulli \((p=.5)\) trial at each site was used as the starting image and then we applied the autologistic model with these two covariates using 20,000 Monte Carlo simulations to produce a ‘true’ scene of species presence and absence. The parameter values in the simulation were \(\theta_0 = -1, \theta_1 = 0, \theta_2 = -3,\) and \(\beta = 1\). We then sampled this true scene using a systematic grid sample of clusters over the entire area of interest \((Y)\). One hundred and forty four sites were sampled, a 5.8% sample, with 20 observed presence sites.

The predictions from the three models are given in the second row of Figure 1. The autologistic model with covariates reproduces the true image well, while the basic autologistic model has trouble with the small amount of information on species presence. The logistic
regression model mirrors the second covariate, as expected when only one informative covariate is included in the model. The ROC curve shows that the logistic regression model can achieve a high true positive rate but only at the expense of an increased false positive rate. Compared to the other two models the autologistic model with covariates has superior discriminatory performance with a smaller tradeoff between accurate versus false prediction of presence.

In the second example (Figure 2), the sampling plan, the covariates, and the parameter values were equivalent to those in Example 1 except that in the Monte Carlo simulation the value of the spatial parameter, $\beta$, was decreased to 0.7. This resulted in a true scene where there are fewer sites where the species is present. The number of sites where the species was present in the true image decreased to 158 as compared to 320 sites in example 1. The sample size remained at 144 with 4 observed presence sites. The basic autologistic model had
a difficult time with so few observed presence sites and had the worst predictive performance of the three models. (In the prediction image for the basic autologistic model for example 2, the greyscale values of the pixels for observed absence sites were increased so that the pattern of the predictions can be seen. Otherwise the difference between the greyscale values for the observed absence sites and the non-sampled sites is too great and the non-sampled sites appear to be completely black.) The logistic regression model and the autologistic model with covariates reproduced the true image fairly well, with the logistic regression model under-predicting presence and the autologistic model with covariates over-predicting presence. This trend is reflected in the ROC plot.

The third example again mirrors example 1, but this time the value of the spatial parameter, $\beta$, was increased to 1.175 resulting in more sites where the species is present. The number of sites where the species was present in the true image increased to 934 as compared
to 320 true presence sites in example 1. There were 56 observed presence sites out of 144 sites sampled. With the additional observed sites, the basic autologistic model has much improved performance, but the autologistic model with covariates still produced the most accurate predictions of the three models. In the ROC curve for this example, the autologistic model with covariates was superior to the other two models for all possible cut-off values. For these covariates and this sample set-up, $\beta$ values greater than 1.2 produced simulations with a dark mass in the center of the image.

The parameter estimates from the autologistic model with covariates for examples 1–3 are given in Table 1. Overall, the parameter estimates for examples 1 and 3 are closer to the true values than the parameter estimates for example 2. Example 2 is the most challenging the three examples as there are only 4 sites where the species was observed. For all three examples the estimate for $\theta_0$ is not significantly different from 0 when the true
value is equal to -1. The parameter estimates for $\theta_1$ and $\theta_2$ in examples 1 and 3 had smaller standard deviations than the parameter estimates for example 2. The estimates for the spatial parameter, $\beta$, was within 1 standard deviation of the true value for examples 1 and 3 and within 1.5 standard deviation of the true value for example 2.

In our fourth example we used the same true image and the same informative covariate (Z2) as in example 1. A new covariate (Z1) was included which gives conflicting information about the quality of habitat; the range of good habitat is larger than the actual range of the species on the top half of the image but the habitat is poor in the bottom half of the image where there are some pixels with true presence. Finally, the sampling plan is based on a smaller systematic grid sample of clusters over the top portion of the area of interest. The sample size is 96, a 3.84% sample, with 10 observed presence sites.

The prediction scenes from the three models are given in the second row of Figure 4. As
<table>
<thead>
<tr>
<th>Example 1</th>
<th>truth</th>
<th>mean</th>
<th>s.d.</th>
<th>Example 2</th>
<th>truth</th>
<th>mean</th>
<th>s.d.</th>
<th>Example 3</th>
<th>truth</th>
<th>mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta_0$</td>
<td>$\theta_1$</td>
<td>$\theta_2$</td>
<td>$\beta$</td>
<td>$\theta_0$</td>
<td>$\theta_1$</td>
<td>$\theta_2$</td>
<td>$\beta$</td>
<td>$\theta_0$</td>
<td>$\theta_1$</td>
<td>$\theta_2$</td>
</tr>
<tr>
<td>truth</td>
<td>$-1.00$</td>
<td>$0.00$</td>
<td>$-3.00$</td>
<td>$1.00$</td>
<td>$-1.00$</td>
<td>$0.00$</td>
<td>$-3.00$</td>
<td>$0.70$</td>
<td>$-1.00$</td>
<td>$0.00$</td>
<td>$-3.00$</td>
</tr>
<tr>
<td>mean</td>
<td>$0.79$</td>
<td>$0.72$</td>
<td>$-4.64$</td>
<td>$0.84$</td>
<td>$1.20$</td>
<td>$-6.45$</td>
<td>$-4.72$</td>
<td>$1.34$</td>
<td>$-1.75$</td>
<td>$1.28$</td>
<td>$-2.64$</td>
</tr>
<tr>
<td>s.d.</td>
<td>$1.27$</td>
<td>$1.28$</td>
<td>$1.49$</td>
<td>$0.17$</td>
<td>$2.33$</td>
<td>$3.49$</td>
<td>$2.51$</td>
<td>$0.43$</td>
<td>$1.16$</td>
<td>$1.37$</td>
<td>$0.96$</td>
</tr>
</tbody>
</table>

one would expect, the logistic regression model produces results that average over the two covariates. The basic autologistic model again produces poor predictions and the autologistic model with covariates does a fairly good job of identifying the areas of true presence. The ROC curve shows that the autologistic model with covariates is superior.

In the examples presented here the autologistic model with covariates produces superior predictions as compared to the logistic regression and basic autologistic model. In examples 1–3, the autologistic model with covariates was the true model. For examples where the true model is a "covariates only" model, one would expect that the logistic regression model would have superior performance. Similarly, for examples were the true model is a "spatial only" model, one would expect that the basic autologistic model would have superior performance. In the examples considered here, the parameter estimates from autologistic model with covariates for the "noise" covariate (Z1 in examples 1–3) were not significantly different from zero demonstrating that the autologistic model with covariates can be used to identify non-informative covariates. In real applications it generally is not clear a priori whether a "covariates only" or "spatial only" model would be superior. The autologistic model with covariates is a general model which offers significant flexibility in the face of an unknown true model.
5 Conclusions and Future Research

Software to implement the methodology described here is available on the internet. The software, which is written in S-Plus©, can be obtained free of charge via the Web address http://lib.stat.cmu.edu/S/autologit.

While the autologistic model with covariates was developed to assist the USFS with monitoring rare species, we have yet to apply these models to USFS data which is still being collected. The autologistic model with covariates has a broad range of other applications as well.

The autologistic model gives an important base for many extensions in spatial modeling. We have shown here that including two other sources of explanatory information, covariates and search intensity, can improve predictions. These simulations show that the autologistic model with covariates can give improved estimates of species presence/absence as compared to simple logistic regression and the basic autologistic models. Where there is good habitat, the model predicts presence and also “builds clusters” of likely areas around observed presence. The autologistic model with covariates is more discriminating than the logistic regression model and yet is sensitive to covariates while the basic autologistic model does not incorporate the covariate information. Including the two levels of search intensity improves prediction by forcing sampled sites to remain as observed.

A further extension is to use alternative specifications of search intensity. Heikkinen and Högnander (1994) use more than two levels of search intensity for the basic autologistic model. This technique can also be applied here for the autologistic model with covariates. Extending the likelihood to be a continuous function of search is accomplished by the addition of this covariate into the model. This added parameter would have a prior distribution and affect the full conditional distributions used in the estimation procedure.

Just as the inclusion of covariates improved predictions, we expect that using two related species in a bivariate response autologistic model will improve prediction accuracy for both species. We are currently examining a bivariate response autologistic model. The resulting prediction map would indicate areas of probable presence for each species and for both species. This should increase efficiency in sampling to detect rare species.

Acknowledgments

The authors would like to thank the editor, three anonymous referees, and Nicole Augustin for their suggestions which lead to improvements in this manuscript, Raymond Czaplewski
for introducing the USFS application problem, and Matt Calder for the development of and his assistance with his compiler for S-Plus©. This research was partially supported by USDA-USFS contract number PNW 95-0766 and National Science Foundation grant DMS-9806243.

References


