

Modeling the Distribution of Plant Species Using The Autologistic Regression Model

Hulin Wu¹

and

Fred W. Huffer

Department of Mathematical Sciences

Department of Statistics

The University of Memphis

The Florida State University

Memphis, TN 38152

Tallahassee, FL 32306-3033

Tel. (901) 678-4147

Tel. (904) 644-8731

Email: wuh@hermes.msci.memst.edu

Email: huffer@stat.fsu.edu

Biographical Sketch: Hulin Wu is currently visiting Assistant Professor of Statistics at University of Memphis. He is a recent graduate of the Department of Statistics at Florida State University. This paper is part of his Ph.D. dissertation under the direction of Dr. Fred W. Huffer. He won first prize at the student paper award given by the Section on Statistics and the Environment of ASA at the 1993 Joint Statistical Meetings. He also received a student paper award at the 1994 ENAR Spring Meetings. His dissertation, which is about modeling the distribution of plant species, won the 1994 R.A. Bradley Award at Florida State University.

Fred W. Huffer is currently Associate Professor of Statistics at Florida State University. He obtained his Ph.D. from Stanford University in 1982 and since then has done research in various areas including geometrical probability, multivariate probability inequalities and survival analysis. This work is part of an ongoing project devoted to the modeling the relationship between species distribution and climate variables.

Acknowledgements: We thank Dr. D.W. Crumpacker of the University of Colorado at Boulder for providing us with species and climate data. This research was partially supported by EPA Grant #CR818052-01-0 to Dr. David W. Crumpacker.

¹corresponding author

Modeling the Distribution of Plant Species Using The Autologistic Regression Model

Hulin Wu* and Fred W. Huffer‡

*Department of Mathematical Sciences, University of Memphis,
Memphis, TN 38152, USA

‡Department of Statistics, The Florida State University, Tallahassee, FL 32306, USA

Abstract

For modeling the distribution of plant species in terms of climate covariates, we consider an autologistic regression model for spatial binary data on a regularly spaced lattice. This model generalizes Besag's (1974) autologistic model by including covariates in the model. Three estimation methods, the coding method, maximum pseudolikelihood method and Markov chain Monte Carlo method are studied and compared via simulation and real data examples. As examples, we use the proposed methodology to model the distributions of two plant species in the state of Florida.

Keywords: binary data, coding method, ecological data, environmental statistics, Markov chain Monte Carlo, plant species, pseudolikelihood, spatial data

1 Introduction

In this paper, we attempt to model the distribution of plant species in terms of climate variables like temperature, moisture and rainfall. In particular, we were given data on the distribution of about 180 plant species in the state of Florida, and data on the values of nine climate variables which were expected to be important factors in determining the distribution of the plant species. This data was assembled by Dr. D.W. Crumpacker of the University of Colorado at Boulder. The plant distribution data was obtained by digitizing maps from Little (1978). The long-term climatic data was compiled from 106 meteorological stations.

In Figure 1, we present four examples of the distribution maps. These four species are *Chamaecyparis Thyoides* (No. 1), *Pinus Clausa* (No. 4), *Castanea Pumila* (No. 38) and *Zanthoxylum Clava-herculis* (No. 158).

Place of Figure 1

The following is a list of climate variables we are using in this paper.

TMM = mean minimum temperature in degrees centigrade of coldest month
(usually January).

TM = mean temperature in degrees centigrade of coldest month (usually January).

TAV = mean annual temperature in degrees centigrade.

LT = lowest temperature recorded from 1931 to 1990 in degrees centigrade.

FZF = median freeze-free period in days.

PRCP = mean total annual precipitation in millimeters.

MI = moisture index = $(PRCP)/(TAV \times 58.93)$,

where $TAV \times 58.93$ = estimate of mean annual potential evapotranspiration
by the Holdridge method.

PMIN = mean total precipitation of driest month in millimeters.

ELV = elevation in feet.

As noticed by Huffer and Wu (1995), these distribution maps reveal a strong degree of spatial correlation in the data, that is, whether a species is present or absent at a given site is strongly related to its presence or absence at neighboring sites. Some spatial correlation may be explained simply by the climate covariates which are themselves spatially correlated and display definite spatial patterns. However, for many species the degree of spatial correlation is much larger than can be explained in terms of the covariates alone.

Ecologists have analyzed similar data by a variety of means. Box et al. (1993) built a climatic-envelope model for many woody Florida plant species. Their work is particularly relevant to ours as they use very similar climate variables and many of the same plant species that we do. Their model is not a formal statistical model, that is, they do not propose a probability model and then proceed to estimate the parameters in this

model. Rather, they give a reasonable method for deducing, for each species, upper and lower limits (an envelope) for each climate variable which describe the habitable range for that species. Bartlein et al. (1986) used ecological response surfaces from pollen data to describe the way in which the abundances of some eastern North American taxa depend on the joint effects of several environmental variables. They remap the abundance patterns from geographic space into climate space and then explain the distributions in terms of biological processes. However, the spatial pattern in the relative sensitivity of different taxa and the spatial variation in the climate variables are ignored in their models. Huntley et al. (1989) also constructed response surfaces for some beech trees in Europe and North America by a locally-weighted averaging technique. These response surfaces characterize the relationship between the pollen percentages and some climate variables. They predicted the patterns of beech pollen abundance in the respective continents. Austin et al. (1990) used logistic regression to model environmental niches of five eucalyptus species. Their results provide evidence for asymmetric responses of species to environmental variables contrary to the symmetrical responses commonly assumed in ecological theory. And also their models suggest that the environmental variables were inappropriate or insufficient for modeling some species with complex response shapes. These authors also did not consider spatial interaction/correlation in their models; they used logistic regression models which assumed independent responses. In this paper, and the papers by Wu (1994) and Huffer and Wu (1995), we try to model the binary data (species present or absent) in ways which properly account for both the spatial correlation and the dependence on the climate covariates.

The model we shall use in this paper is the autologistic regression model which is explained in the next section (Section 2). The autologistic regression model is a straightforward generalization of the autologistic model introduced by Besag (1974, 1975) and studied by many authors (see Zhao and Prentice (1990), Geyer and Thompson (1992), Wu (1994), Huffer and Wu (1995)). In order to solve the fitting difficulties caused by the spatial dependence for autologistic regression models, three methods have been proposed (Besag, 1974, 1975; Wu, 1994; Huffer and Wu, 1995). The three

methods are reviewed in Section 3. We compare the three methods through a simulation study in Section 4. In Section 5, we apply the three estimation methods to fit the distribution data for two plant species, and we compare the results in terms of the fitted errors defined in this section. A summary of conclusions and some discussion are given in Section 6.

2 The Autologistic Regression Model

Following the notation of Huffer and Wu (1995), we assume that our data is recorded at m locations (sites) forming a subset \mathcal{S} of a rectangular lattice. Each site in \mathcal{S} is described by giving coordinates (k, ℓ) specifying the row and column of the lattice at which it is located. The sites in \mathcal{S} are numbered from 1 to m in some arbitrary fashion. At each site i we observe a binary response y_i and a $p \times 1$ vector of covariates \mathbf{x}_i . Taken altogether, the m binary responses constitute a map $Y = (y_i)$. The autologistic regression model specifies the conditional probability p_i that $y_i = 1$ given all the other values y_j ($j \neq i$) as follows:

$$\begin{aligned} p_i &= P(y_i = 1 | \text{all other values}) = P(y_i = 1 | \text{nearest neighbors}) \\ &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \quad \text{where} \quad \eta_i = \beta_0 + \mathbf{x}_i' \boldsymbol{\beta}_1 + \gamma y_i^*. \end{aligned} \quad (1)$$

Here y_i^* denotes the neighborhood sum for site i , that is,

$$y_i^* = \sum_{j=1}^m y_j I(i \sim j) = \sum_{j: i \sim j} y_j$$

where $i \sim j$ indicates that sites i and j are “neighbors”. For a site i with coordinates (k, ℓ) , the neighbors j are any members of \mathcal{S} occupying the four nearest locations $(k - 1, \ell)$, $(k + 1, \ell)$, $(k, \ell - 1)$, $(k, \ell + 1)$. A site in the interior of \mathcal{S} will have 4 neighbors. Sites on the boundary of \mathcal{S} will have fewer neighbors.

The parameters in this model are the intercept β_0 , a $p \times 1$ vector $\boldsymbol{\beta}_1$ which specifies the covariate effects, and a parameter γ which determines the degree of spatial interaction or correlation in the data. When $\gamma = 0$, the model (1) reduces to the ordinary logistic regression model which is appropriate for independent binary responses. When $\boldsymbol{\beta}_1 = 0$, the model becomes the autologistic model of Besag (1974, 1975).

3 Estimation Methods

Two estimation methods, the coding method (COD) and maximum pseudolikelihood (MPL) method have been proposed for the autologistic model by Besag (1974, 1975). Recently a new statistical technique, Markov chain Monte Carlo (MCMC), has been developed to approximate intractable distributions. This technique can also be used to approximate the exact maximum likelihood estimate (MLE). We briefly review these three methods in this section.

3.1 Coding method

The coding method was first proposed by Besag (1972) and then generalized in his famous paper in 1974. We sketch the method as follows in Besag's notation.

In order to fit a first-order scheme, we first label the interior sites of the lattice, alternately by \times and \bullet , as shown in Figure 2.

Place of Figure 2

According to the first-order Markov assumption (1), the random variables located at the \times sites, given the observed values at all other sites, are mutually independent. Thus we can get the conditional likelihood function for all \times sites:

$$\prod_{\text{all } \times \text{ sites}} P(Y_i = y_i | \text{all other values}). \quad (2)$$

The maximum-likelihood estimates of the unknown parameters can now be obtained in the usual way. Similarly, alternative estimates may be obtained using the \bullet sites. We combine the two sets of estimates (usually by averaging) to obtain our final estimates.

In order to estimate the parameters of a second or higher order scheme, we may code the interior sites of the lattice in an appropriate way, then combine the results.

See Besag (1974) for details.

Using the coding method, we may easily construct likelihood-ratio tests to examine the goodness of fit of particular schemes. Its simplicity and flexibility are also great advantages, but its drawbacks cannot be neglected. The estimates from the different coded sets may be far apart; just averaging them may not be a good method of combination. A weighted average may be better, but we need to determine appropriate weights. Another disadvantage is that there are many coding schemes, and the results from the different schemes may not be consistent. Generally, the coding method is not efficient. The efficiency of this method was studied in the context of Gaussian lattice processes by Besag and Moran (1975). The coding method was soon replaced by the pseudolikelihood method, which is claimed to be more efficient and better by several authors (Besag and Moran 1975). This is probably the reason why there is no more literature on the coding method after this point.

3.2 Maximum pseudolikelihood method

The method of maximum pseudolikelihood estimation (MPLE) also originates with Besag (1975). Ripley (1988) states a general version of the pseudolikelihood. Jensen and Møller (1991) derive the pseudolikelihood by a direct argument for general spatial point processes. Since this method is intuitively plausible, computationally convenient, easily implemented (some statistical packages, such as SAS, S-PLUS, and GLIM can be directly used to obtain the MPLEs), and more efficient than the COD method, it has drawn many authors' attention. Besag (1977) studied the efficiency of pseudolikelihood estimation for simple Gaussian fields. The first proof of consistency of MPL estimators (for fully observed data) was established by Geman and Graffine (1987), and Gidas (1986) gives an alternative proof. Comets (1992) proves strong consistency of a class of maximum objective estimators for exponential parametric families of Markov random fields on \mathbf{Z}^d , including both the MLE and MPLE, using large deviation estimates. Gidas (1991) claims that the MPL estimators are, under appropriate conditions, asymptotically normal (but not efficient), but the proof is unpublished. More work on the asymptotic normality is needed. Some modifications of the MPL method have

been introduced by Chalmond (1986). Some applications of the MPLE are presented by Strauss and Ikeda (1990), Arnold and Strauss (1991), and Preisler (1991).

The following is a simple description of the MPL method. We describe it in terms of our autologistic regression model.

The maximum pseudolikelihood estimates (MPLEs) of the unknown parameters are those parameter values which maximize the quantity

$$\prod_i p_i(\beta_0, \boldsymbol{\beta}_1, \gamma) = \prod_i P(Y_i = y_i | \text{all other values}) \quad (3)$$

with respect to the unknown parameters.

These estimates are closely related to those obtained by the coding method. Indeed, the MPLEs can be thought of as a weighted average of the coding method estimates (CODEs). The disadvantage of the MPL method is its inefficiency, especially in cases where the spatial interaction is strong (Besag 1975).

3.3 Markov chain Monte Carlo method

Markov chain Monte Carlo methods can be used to approximate the MLEs for any family of distributions having probability densities known up to a constant of proportionality (Moyeed and Baddeley, 1991; Geyer, 1991, 1992; Geyer and Thompson, 1992; Gelman and Rubin, 1992). Huffer and Wu (1995) studied the MCMC method in the context of autologistic regression models. We shall follow their notation to give a brief review of this method. See Wu (1994), Huffer and Wu (1995), or the papers of Geyer for further details.

Let $\theta = (\beta_0, \boldsymbol{\beta}_1, \gamma)$ denote the vector of parameters in our autologistic regression model, and let P_θ be the probability measure of a random map $Y = (y_i)$ generated from this model. The measures P_θ form an exponential family; we can write

$$P_\theta(Y = y) = c(\theta)^{-1} \exp\{\theta^t t(y)\} \quad (4)$$

where

$$t(y) = \left(\sum_{i=1}^m y_i, \sum_{i=1}^m \mathbf{x}_i y_i, \frac{1}{2} \sum_{i=1}^m y_i y_i^* \right) \quad (5)$$

is the vector of sufficient of sufficient statistics, and $c(\theta)$ is an intractable normalizing constant. Given an observed data map Y_{obs} , we would like to compute the maximum likelihood estimate (MLE) of θ . However, because of the intractable normalizing constant, we cannot compute the likelihood function directly and must resort to Monte Carlo methods.

For any given parameter vector ψ , we can use a Gibbs sampler, based on the conditional formulation of the autologistic model in (1), to generate a sample Y_1, Y_2, \dots, Y_n from P_ψ . Using this sample we can construct a Monte Carlo approximation to the likelihood function valid for θ sufficiently close ψ . We can then find the value $\hat{\theta}$ which maximizes this approximate likelihood. This value $\hat{\theta}$ (when it exists) is the Monte Carlo approximant of the MLE; we refer to it as the MCMC estimate. The value $\hat{\theta}$ can be obtained by solving (using Newton-Raphson or variants thereof) the Monte Carlo score equation given by

$$\frac{\sum_{j=1}^n T_j e^{(\theta-\psi)'T_j}}{\sum_{j=1}^n e^{(\theta-\psi)'T_j}} = T_{\text{obs}} \quad (6)$$

where Y_{obs} denotes the observed data, and $T_{\text{obs}} = t(Y_{\text{obs}})$ and $T_j = t(Y_j)$ for $j = 1, 2, \dots, n$.

The success of this approach depends on the choice of ψ . If ψ is too far from the exact MLE, the score equation (6) may fail to have a solution, or may have a solution $\hat{\theta}$ which is far from the exact MLE. In our simulation work, we use the maximum pseudolikelihood estimate (MPLE) for ψ . The simulation studies by Huffer and Wu (1995) show that the MCMC estimates in our autologistic regression models are approximately normally distributed and that the MCMC estimates of Fisher information may be used to estimate the variance of the MCMC estimates and to construct confidence intervals.

4 Simulation Comparison of Estimation Methods

In this section, we compare the COD, MPL and MCMC estimation methods via simulation. In the following simulation, we shall use the autologistic model on a 40×40

lattice. We shall restrict consideration to a single covariate ($p = 1$) so that η_i in (1) can be written simply as $\eta_i = \beta_0 + \beta_1 x_i + \gamma y_i^*$. This covariate has the form of a diagonal sine wave; for a site i with coordinates (k, ℓ) , the covariate takes on the value

$$x_i = 2.5 \times \sin(0.1 \times (k + \ell)).$$

In our simulations we choose values of the spatial interaction parameter γ ranging from 0.0 to 1.0, we keep the covariate coefficient $\beta_1 = 2.0$ in all cases, and we change the value of β_0 from -1.0 to 1.0 in such a way as to balance the spatial interaction and avoid situations where the value “1” or “0” dominates the entire lattice. For each set of “true” parameter values, 500 sets of the pseudo-observed data were generated by a Gibbs sampler (Geman and Geman 1984) using the model (1), i.e., the 500 data sets are obtained from 500 independent Markov chains. For each pseudo-observed data set, we use each of the three estimation methods (COD, MPL and MCMC) to obtain estimates of the parameters $(\beta_0, \beta_1, \gamma)$.

In the COD method, we take the average of the two estimates from the two coded parts as the final estimate. In the MCMC method, the “one long run” scheme (Geyer, 1992a) is used, the MPLEs are used as the initial guess for ψ , and the burn-in or warm-up period of 100, spacing of 1 and sample size of 2000 are chosen by experience and the suggestions of Geyer and Thompson (1992). The computational work is done using an S-PLUS interface to FORTRAN on SUN SPARC Stations.

For each parameter case, we obtained 500 independent estimates for each estimation method. The means, standard errors and mean squared errors (MSEs) of the estimates from the three methods are summarized in Tables 1 and 2 (the standard errors are in the parenthesis of Table 1).

Place of Tables 1 and 2

From these tables, we can see that the means of the MCMC MLEs, MPLEs and

CODEs are very close, although a weak trend can be identified, i.e., the bias of the CODEs is smallest, and the bias of the MCMC estimates is biggest. However, the standard errors and mean square errors (MSEs) of the MCMC method are consistently smaller than those of the MPLEs and CODEs (see Tables 1 and 2). Especially for greater spatial interaction cases, the estimation errors (variance or MSE) of the MCMC estimates are much smaller. From here, we can conclude that the MCMC method really does improve upon the MPL and COD, although it costs more in computation time.

The simulation results also display a tendency in the three methods for the error of the estimates to increase as the spatial interaction increases. This phenomenon can be explained heuristically as follows. In standard statistical settings which involve i.i.d. random variables, the variance of a parameter estimate is proportional to $1/n$ where n is the sample size. With spatially correlated data, one might expect the variance to be proportional to $1/n$ where now n is some sort of “effective sample size”. Increasing the value of γ will increase the degree of correlation between nearby sites on the lattice and thus decrease the effective sample size.

5 Applications to Plant Species

In Section 1, we introduced the distribution data for about 180 plant species in the state of Florida with nine climate covariates. Here we take the species, *Castanea pumila* (No. 38) and *Zanthoxylum Clava-herculis* (No. 158) as examples to model and analyze. The other species can be modeled in the same way.

For this analysis, we “digitize” the map of Florida so that it is represented as a collection of $m = 1845$ sites forming a subset of a rectangular 68×80 lattice. These 1845 sites form the set \mathcal{S} of Section 2. We also digitize the distribution maps of species No. 38 and No. 158 from Little (1978), that is, at each site i we record $y_i = 1$ or 0 according to whether the species was present or absent at this site. Then we are ready to apply the autologistic regression models to model the distribution of the plant species. Based on the AIC criterion, Wu (1994) suggested that the autologistic regression model with covariates (TAV, PRCP, MI) and an intercept is plausible for

species No. 38; and for species No. 158, the model with covariates (TM, TMM, ELV) and an intercept is a plausible choice. We fitted the two models for these two species using the three estimation methods reviewed in Section 3. In this section we present the fitting results from the three methods and compare them in terms of the fitted errors which will be defined later.

5.1 Parameter Estimation

Tables 3 and 4 present the parameter estimation results from the COD, MPL and MCMC methods for species No. 38 and No. 158. These tables give both parameter estimates and approximate standard errors.

Place of Tables 3 and 4

Both the COD and MPL estimates are obtained using standard logistic regression routines in S-PLUS; we just need to include the neighborhood sum y_i^* as an additional covariate. To obtain the MPL estimates we simply fit a logistic regression model which includes the desired climate variables and the neighborhood sum y_i^* as covariates. For the COD method, we fit two such logistic regression models, one for each of the coded parts of the data, and then take the average of the two sets of estimates as our final estimate, i.e., $\hat{\theta} = (\hat{\theta}_1 + \hat{\theta}_2)/2$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimates from the first and second coded parts of the data.

The “approximate standard errors” reported for the COD and MPL estimates are obtained from the output of the logistic regression software. For the MPL estimates, we give the standard errors exactly as listed in the logistic regression output. For the COD estimates, we report those standard errors which would be valid if the two coded halves of our data were statistically independent, that is, if we assume that $\hat{\theta}_1$ and $\hat{\theta}_2$

are independent, then

$$\text{Appr. S.E.} = \sqrt{\text{Var}(\hat{\theta})} = \sqrt{\frac{1}{4}(\text{Var}(\hat{\theta}_1) + \text{Var}(\hat{\theta}_2))}. \quad (7)$$

Here $\text{Var}(\hat{\theta}_1)$ and $\text{Var}(\hat{\theta}_2)$ are the standard errors (based on the Fisher information) given by the logistic regression software for the two sets of estimates. These standard errors reported for the COD and MPL estimates are based on naive acceptance of the output from the logistic regression software. Since the logistic regression assumptions are not valid here, they have no theoretical justification, but we are hoping they are still roughly correct.

The standard errors for the MCMC method are based upon the Fisher information for the autologistic regression model as discussed in Wu (1994) and Huffer and Wu (1995). The simulations in Huffer and Wu (1995) have shown that these values are fairly accurate when the data is generated from an autologistic regression model. In implementing the MCMC method, we take a burn-in or warm-up period of 100, spacing of 1, and sample size of 2000. The MPLEs are used as the initial “guess” of ψ to generate MCMC samples.

Tables 3 and 4 shows that the MPLEs and CODEs are relatively close, but the MCMC estimates are far away from the other two estimates. From Tables 3 and 4, we can also see that the MCMC estimates have the smallest approximate standard errors, and the approximate standard errors of MPLEs are a little smaller than that of CODEs. We compare the three estimates in terms of their fitted errors in the next subsection.

5.2 Unconditional Probability Fitting and Fitted Errors

After estimating the parameters in the autologistic regression model, Monte Carlo samples from this model can be generated using the Gibbs sampler (Geman and Geman, 1984; Gelman and Rubin, 1992; Geyer, 1992). We can then compare these generated maps with the actual data map.

We shall now describe an informal approach to studying the goodness-of-fit of our models. Let n be the Monte Carlo sample size, and k_i be the number of times (out of

n) that the species was present at lattice site i . Then $\hat{p}_i = k_i/n$ is an estimate of the unconditional probability that the species is present at this lattice site. Let y_i denote the observed value (1 or 0) at site i ; we define the “fitted errors” as the differences between the fitted unconditional probabilities and the observations. The following two quantities are used to summarize the fitted errors: the sum of absolute errors (SAE) defined by

$$\text{SAE} = \sum_{i=1}^m |y_i - \hat{p}_i|, \quad (8)$$

and the sum of squares of errors (SSE) defined by

$$\text{SSE} = \sum_{i=1}^m (y_i - \hat{p}_i)^2. \quad (9)$$

To compute the \hat{p}_i , we used a single long run of the Gibbs sampler to generate 1000 samples. We chose a spacing of 10 and burn-in period of 1000 for our Gibbs sampler. The starting state is 0 at all sites and the boundary values are 0. The fitted errors from the three sets of estimates, CODEs, MPLEs and MCMC MLEs, are reported in Table 5. We have used “digit plots” to display the fitted unconditional probabilities from the three methods in Figures 3-8. In the “digit plot”, we round the values of \hat{p}_i to the nearest tenth, and display the resulting single digit values on a map.

Place of Table 5

Table 5 shows that the MCMC method has smaller fitted errors than the other two methods as we expect. For Species No. 38 we see that the MCMC method does much better than MPL, which in turn does much better than COD. This is true whether the comparisons are done using SAE or SSE. For Species No. 158 we again have MCMC doing better than MPL, which does (slightly) better than COD. This ordering holds for both measures SAE and SSE. However, for Species No. 158 the differences between the methods is much less dramatic, and, in fact, none of the methods does very well. From

Figures 3-8, we can see that, for both species, the digit plot of the fitted unconditional probabilities from the MCMC method is closest in appearance to the observed map. In terms of computational effort, the MPL method costs least, the COD method comes second, and the MCMC method needs much more computation time than the other two methods.

Place of Figures 3-8

We note that it is very difficult to get a good fit for species No. 158. We feel that adding quadratic terms in the climate variables to our model might improve the fit, and we were able to fit such models (that is, obtain parameter estimates) using COD and MPL. However, when we attempted to generate observations from the fitted models using the Gibbs sampler, we encountered the problem of “phase transition”: our simulated maps tended to consist entirely of ‘1’s or entirely of ‘0’s. For the MCMC method, we ran into numerical problems for models with quadratic terms, and we had difficulty obtaining parameter estimates. We think that Species No. 158 may not be well explained by our covariates; perhaps other covariates (for example, soil type) should be added to the model. Another possibility is that the distribution of Species No. 158 has changed due to historical factors (natural calamities or human causes) unrelated to the covariates.

6 Conclusion and Discussion

In this paper we proposed using an autologistic regression model for the modeling of spatial binary data with covariates. We described three estimation methods, the coding method (COD), maximum pseudo-likelihood estimation (MPL), and Markov chain Monte Carlo (MCMC), and showed how they can be implemented for the autologistic regression model. Simulation studies were conducted to compare the three estimation

methods. These studies led to the following conclusions: The MCMC method gives the best estimates, but requires much more computational time than the other two methods. The MPL method requires the least computation and has estimation errors similar to the COD method. The accuracy advantage of the MCMC method is substantial when the spatial interaction is strong, that is, when γ is large. However, when γ is small, the MPL estimates should be adequate for most purposes. We applied our methodology to the distribution of two plant species in the state of Florida. We obtained reasonable results.

The methodology proposed in this paper may be used in many fields such as geology, ecology, agriculture, medical science, epidemiology, meteorology, and environmental science. Modeling the distribution of plant species is a typical example. We could also use our methods to predict the chance (probability) it will rain at some location based on current or past climate conditions. Or we may estimate the probability of the existence of some mineral in a given strata of the earth based on geological structure.

The analysis of spatial binary data is an important topic. We hope that our research will motivate more research on this topic. Some subjects such as residual diagnostics, goodness-of-fit tests, and covariate selection must be addressed if the autologistic regression model is to become widely used. We hope that most of the machinery developed for the ordinary logistic regression model will be usable in our setting with only minor changes. In this paper we have had to rely on simulation results to reach some conclusions; there needs to be more theoretical research on the properties of the MPL and MCMC estimates. Also, the algorithm we currently use for computing the MCMC estimates has difficulties when the model contains many covariates or the spatial interaction parameter γ is very large. Improvements to this algorithm are needed.

Finally, we wish to mention some possible extensions of the autologistic regression model. In this paper we have considered only first-order models, that is, models with dependence only on the four nearest neighbors. It is easy to formulate higher-order autologistic models. However, it is difficult to predict how useful such models would be in practice or what complications they may introduce in the modeling process. One can also formulate spatial-temporal versions of the autologistic regression model. Such

extensions of the model would be very useful in areas where the time evolution is crucial such as weather forecasting and epidemiology.

References

- Arnold, B. C., and Strauss, D. (1991), "Pseudolikelihood Estimation: Some Examples," *Sankhya: The Indian Journal of Statistics*, Vol.53, Series B, 233-243.
- Austin, M. P., Nicholls, A. O., and Margules, C. R. (1990), "Measurement of the Realized Qualitative Niche: Environmental Niches of Five Eucalyptus Species," *Ecological Monographs*, 60(2), 161-177.
- Bartlein, P. J., Prentice, I. C., and Webb, T. (1986), "Climatic Response Surfaces from Pollen Data for Some Eastern North American Taxa," *Journal of Biogeography*, 13, 35-57.
- Besag, J. (1972), "Nearest-neighbor Systems and the Auto-logistic model for Binary Data (with Discussion)," *Journal of the Royal Statistical Society*, Series B, 34, 75-83.
- (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society (with Discussion)*, Series B, 36, 192-236.
- (1975), "Statistical Analysis of Non-lattice Data," *The Statistician*, 24, 179-195.
- (1977), "Efficiency of Pseudolikelihood Estimators for Simple Gaussian Fields," *Biometrika*, 64, 616-8.
- Box, E. O., Crumpacker, D. W., and Hardin E. D. (1993), "A Climatic Model for Location of Plant Species in Florida, U.S.A.," *Journal of Biogeography*, 20, 629-644.
- Chalmond, B. (1986), "Image Restoration Using an Estimated Markov Model," preprint,

Mathematics Dept., University of Paris, Orsay.

Comets, F. (1992), "On Consistency of a Class of Estimators for Exponential Families of Markov Random Fields on the Lattice," *The Annals of Statistics*, Vol.20, No.1, 455-568.

Cressie, N. (1993), *Statistics for Spatial Data (Revised Edition)*, New York: John Wiley.

Gelman, A., and Rubin, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences (with discussion)", *Statistical Science*, Vol.7, No.4, 457-472.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Geman, S., and Graffine, C. (1987), "Markov Random Field Image Models and Their Applications to Computer Vision," *Proceedings of the 1986 International Congress of Mathematicians*, (A. M. Gleason, ed.) 2 1496-1517. American Mathematical Society, Providence, R.I.

Geyer, C. J. (1991), "Markov Chain Monte Carlo Maximum Likelihood," *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E.M. Keramides, ed.), 156-163.

——— (1992), "Practical Markov Chain Monte Carlo (with discussion)", *Statistical Science*, Vol.7, No.4, 473-511.

Geyer, C. J. and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data (with discussion)," *Journal of the Royal Statistical Society*, Ser.B 54 657-699.

Gidas, B. (1986), "Consistency of Maximum Likelihood and Pseudolikelihood Estimators for Gibbs Distributions," *Proceedings of the Workshop on Stochastic Dif-*

- ferential Systems with Applications in Electrical/Computer Engineering, Control Theory, and Operations Research*, IMA, University of Minnesota. 129-145.
- (1991), “Parameter Estimation for Gibbs Distributions from Fully observed Data”, *Markov Random Fields: Theory and Applications*, (R. Chellappa and A. Jain, eds), Academic, New York. 471-498.
- Huffer, F.W. and Wu, H. (1995), “Markov Chain Monte Carlo for Autologistic Regression Models with Application to the Distribution of Plant Species,” submitted to *Biometrics*.
- Huntley, B., Bartlein, P. J., and Prentice, I. C. (1989), “Climatic Control of the Distribution and Abundance of Beech (*Fagus L.*) in Europe and North America,” *Journal of Biogeography*, 16, 551-560.
- Jensen, J.L. and Møller, J. (1991), “Pseudolikelihood for Exponential Family Models of Spatial Point Processes,” *The Annals of Applied Probability*, Vol. 1, No. 3, 445-461.
- Little, Jr., E. L. (1978), *Atlas of United State Trees*, Volume 5. Florida. Misc. Publ. No. 1361, USDA Forest Service. Washington, D. C.: U.S. Government Printing Office. 256 maps, with indices of common and scientific names.
- Moyeed, R.A., and Baddeley, A. J. (1991), “Stochastic Approximation of the MLE for a Spatial Point Pattern,” *Scand. J. Statist.*, 18, 39-50.
- Preisler, H. K. (1991), “Spatial Patterns of Trees Attacked by Beetles: Pseudolikelihood Estimation and Iterative Simulations,” *8 Proc. of Computer Sci. and Stat.: 8th Annual Symp. on the Interface*, 23, 491-494.
- Ripley, B. D. (1988), *Statistical Inference for Spatial Processes*, Cambridge University Press.

- Schwartz, M. W. (1988), "Species Diversity Patterns in Woody Flora on Three North American Peninsulas," *Journal of Biogeography*, 15, 759-774.
- Strauss, D., and Ikeda, M. (1990), "Pseudolikelihood Estimation for Social Networks," *Journal of the American Statistical Association*, 85, 204-212.
- Wu, H. (1994), "Regression Models for Spatial Binary Data with Application to the Distribution of Plant Species," Ph.D. Dissertation, Department of Statistics, The Florida State University.
- Wu, H. and Huffer, F.W. (1995), "Markov Chain Monte Carlo for Autologistic Regression Models with Application to the Distribution of Plant Species," submitted to *Biometrics*.
- Zhao, L. P., and Prentice, R. L. (1990). Correlated Binary Regression Using a Quadratic Exponential Model. *Biometrika*, 77, 642-648.

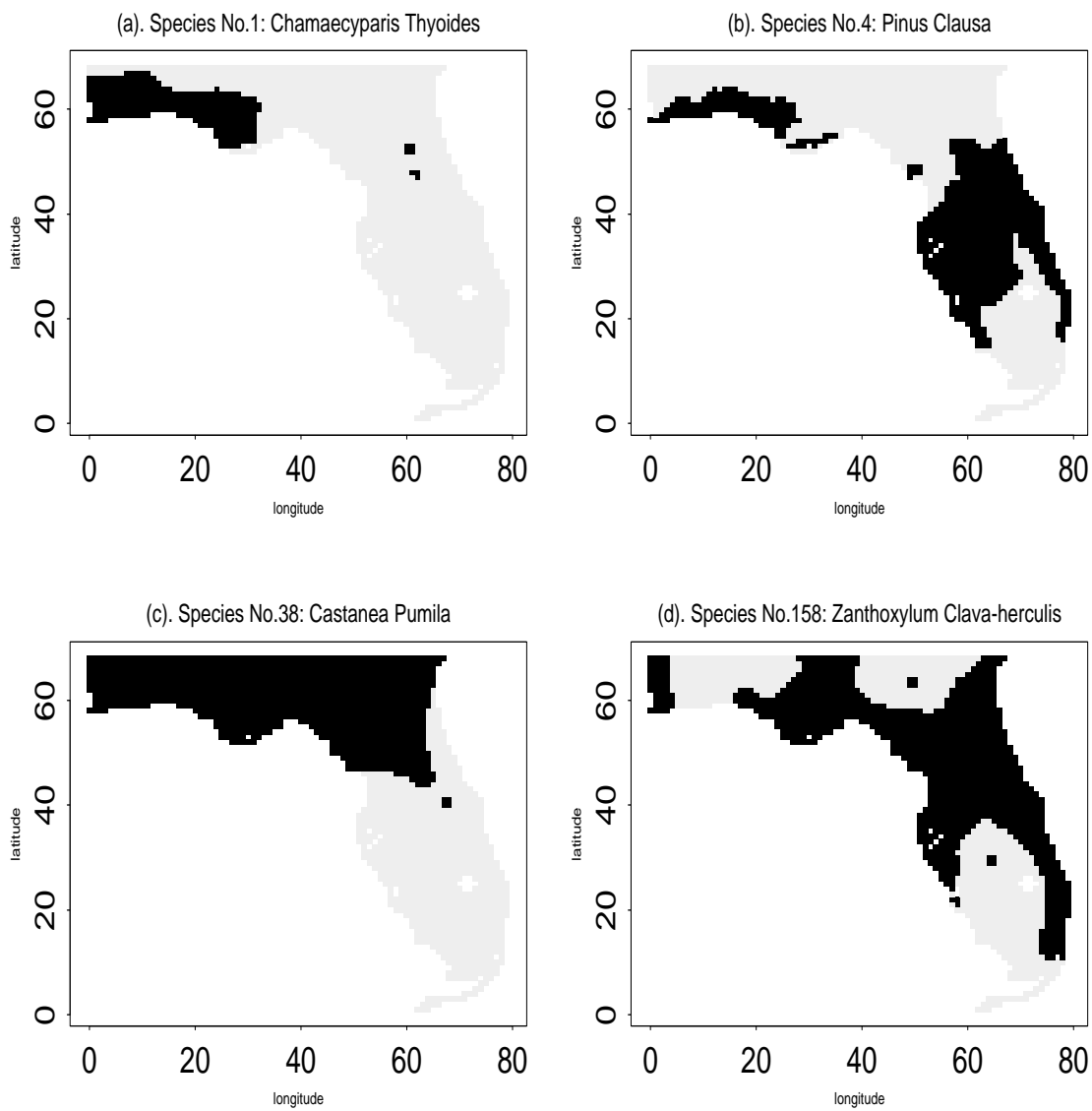


Figure 1: Examples of the distribution maps of plant species, the dark area indicates the presence of the species.

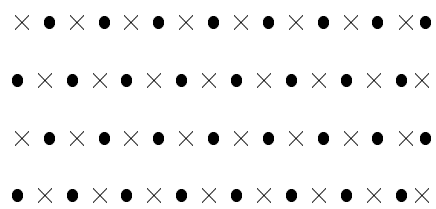


Figure 2: Coding pattern for a first-order scheme.

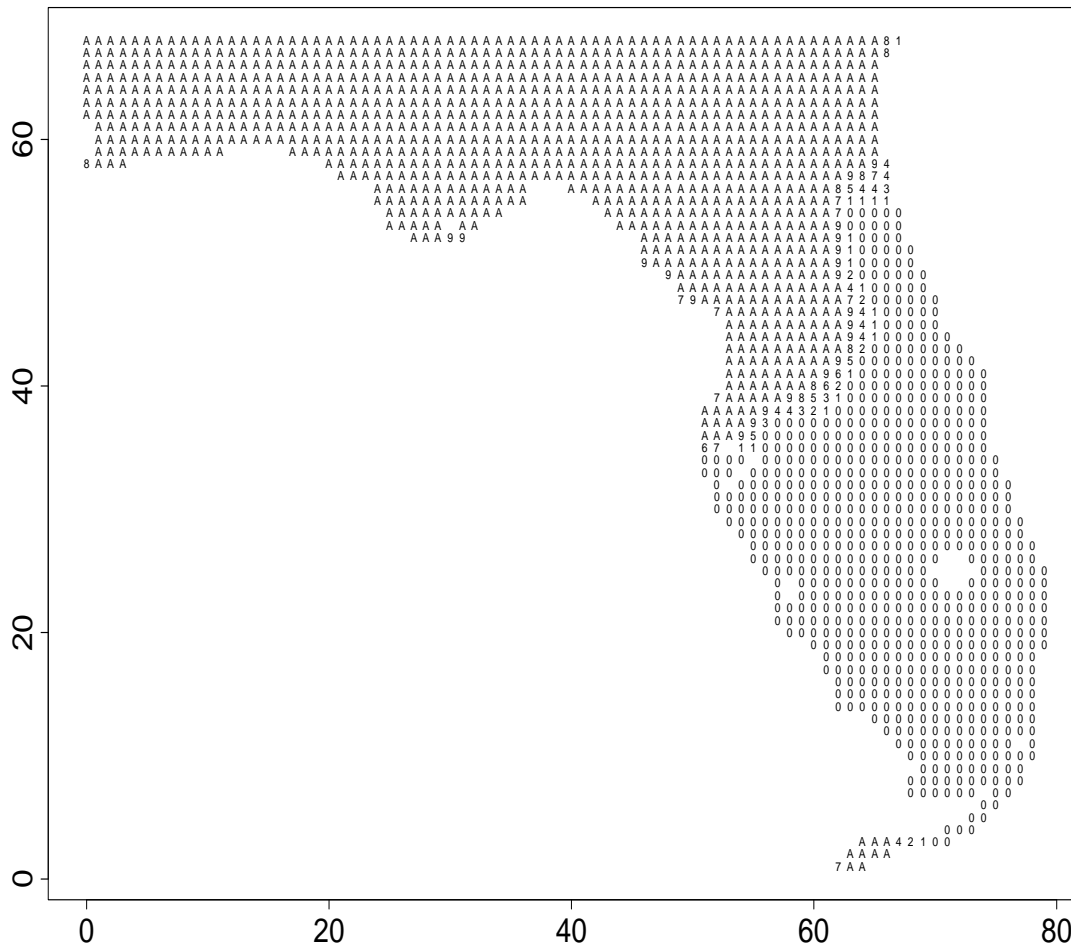


Figure 4: Digit plot of fitted unconditional probabilities for species No. 38: MPLE method ('0'= 0, '1'=0.1, '2'= 0.2, ..., '9'= 0.9, 'A'= 1.0).

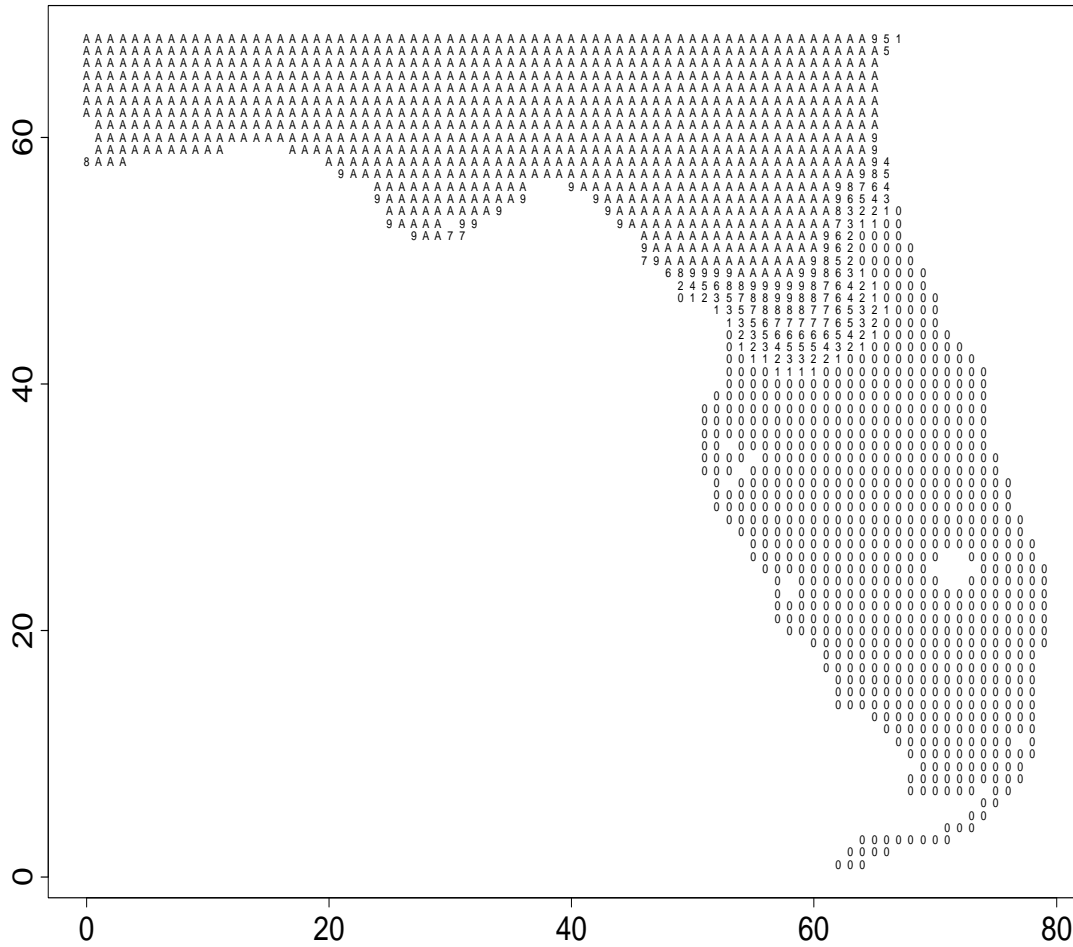


Figure 5: Digit plot of fitted unconditional probabilities for species No. 38: MCMC method ('0'= 0, '1'=0.1, '2'= 0.2, ..., '9'= 0.9, 'A'= 1.0).

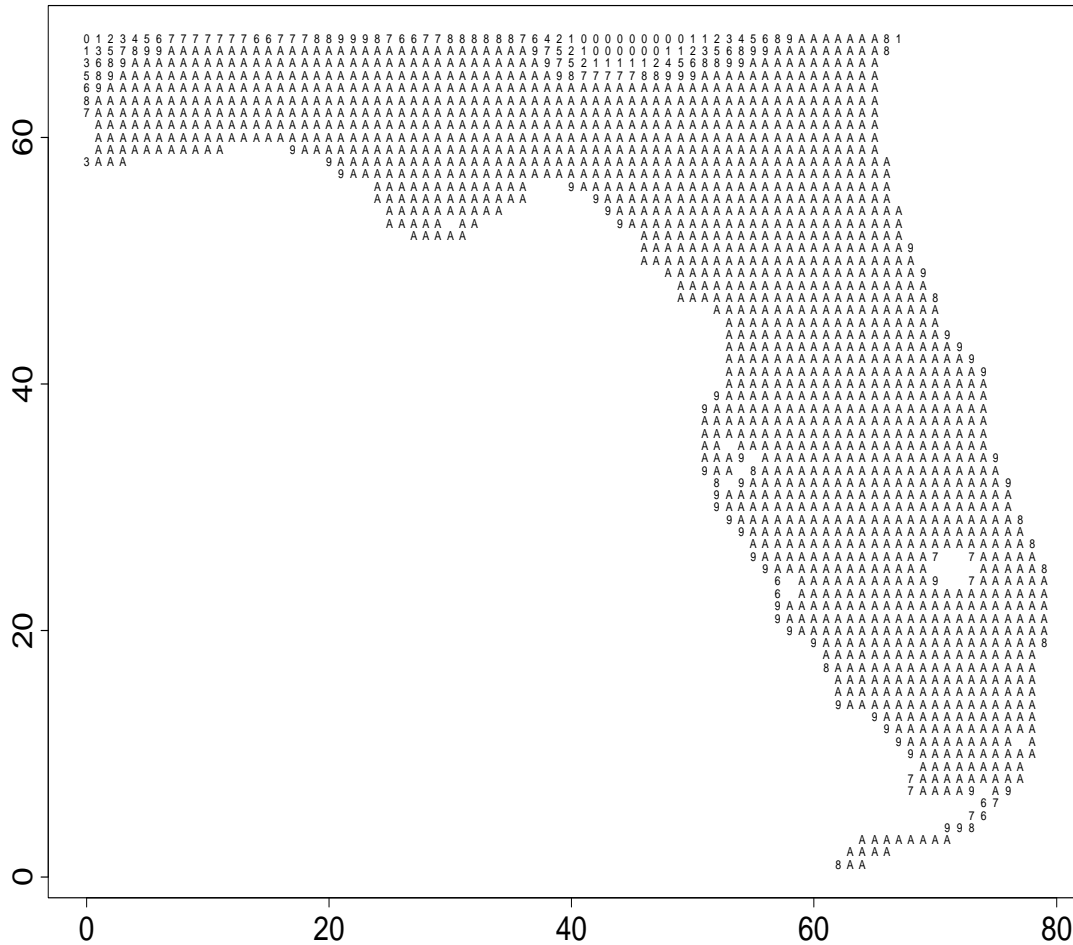


Figure 6: Digit plot of fitted unconditional probabilities for species No. 158: CODE method ('0'= 0, '1'=0.1, '2'= 0.2, ..., '9'= 0.9, 'A'= 1.0).

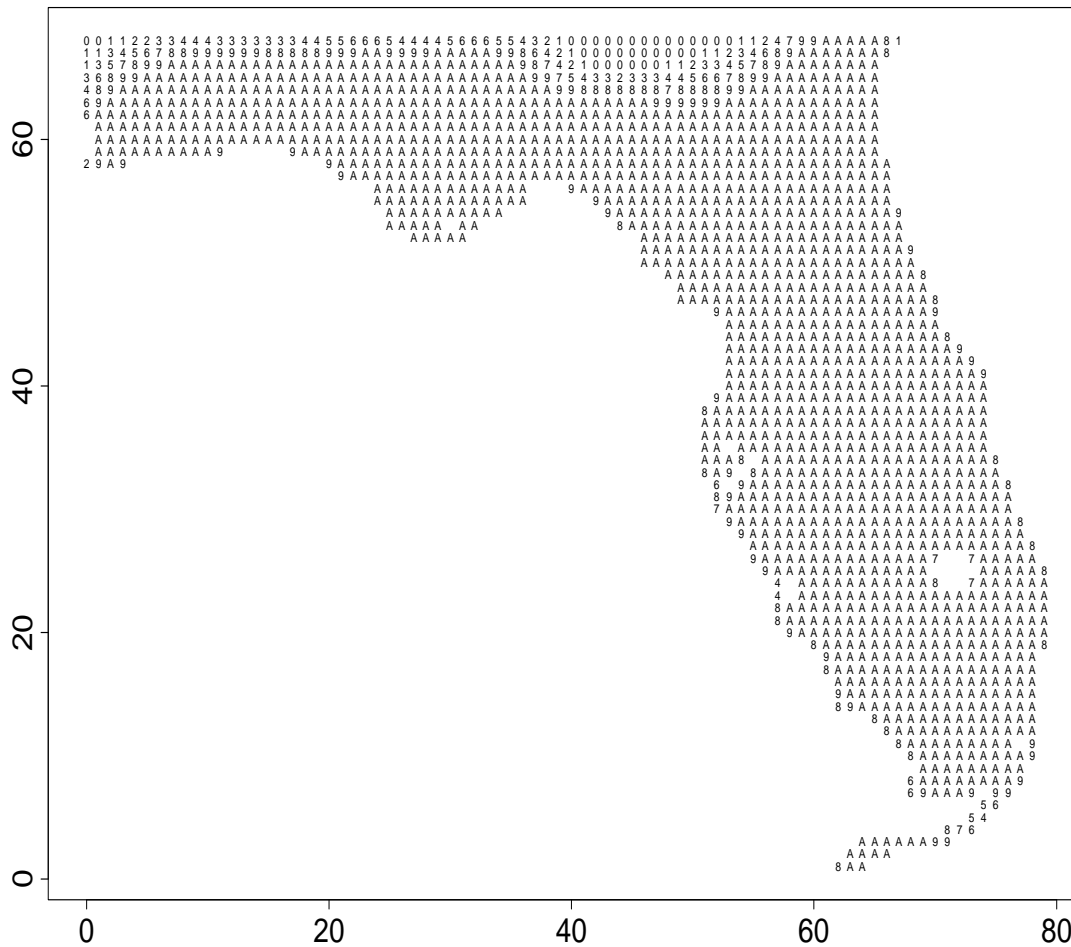


Figure 7: Digit plot of fitted unconditional probabilities for species No. 158: MPLE method ('0'= 0, '1'=0.1, '2'= 0.2, ..., '9'= 0.9, 'A'= 1.0).

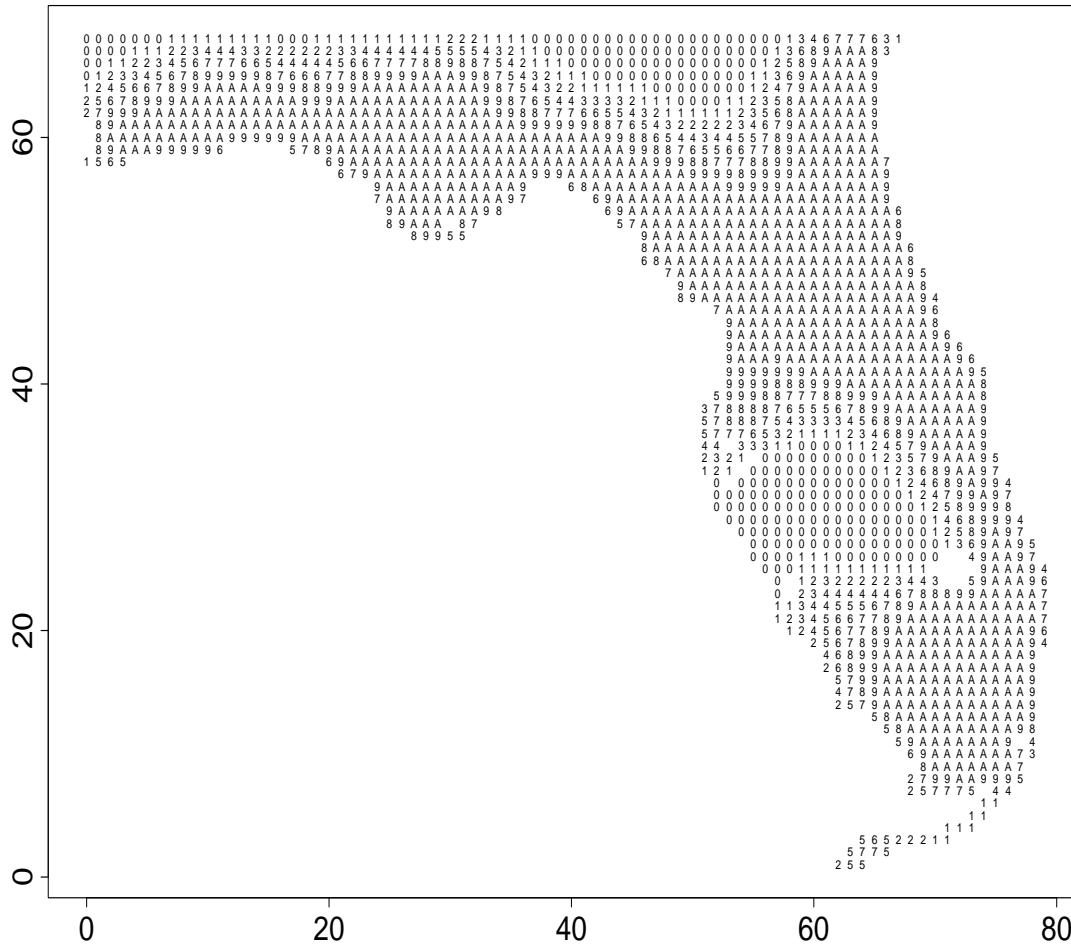


Figure 8: Digit plot of fitted unconditional probabilities for species No. 158: MCMC method ('0'= 0, '1'=0.1, '2'= 0.2, ..., '9'= 0.9, 'A'= 1.0).

Table 1: Mean estimates and standard errors of the three estimation methods.

True Parameters				
$(\beta_0, \beta_1, \gamma)$	Method	$\hat{\beta}_0$ (S.E.)	$\hat{\beta}_1$ (S.E.)	$\hat{\gamma}$ (S.E.)
(1, 2, 0)	CODE	1.0084 (0.257)	2.0117 (0.179)	-0.0005 (0.094)
	MPLE	1.0199 (0.255)	2.0142 (0.178)	-0.0060 (0.093)
	MCMC	1.0213 (0.250)	2.0151 (0.176)	-0.0066 (0.091)
(0.2, 2, 0.2)	CODE	0.1904 (0.259)	2.0066 (0.205)	0.2068 (0.102)
	MPLE	0.2038 (0.259)	2.0117 (0.204)	0.2002 (0.102)
	MCMC	0.2112 (0.252)	2.0169 (0.201)	0.1970 (0.099)
(-0.2, 2, 0.4)	CODE	-0.2144 (0.264)	2.0050 (0.220)	0.4092 (0.104)
	MPLE	-0.1981 (0.264)	2.0117 (0.220)	0.4011 (0.104)
	MCMC	-0.1868 (0.256)	2.0203 (0.215)	0.3963 (0.100)
(-0.6, 2, 0.6)	CODE	-0.6068 (0.286)	2.0155 (0.249)	0.6055 (0.113)
	MPLE	-0.5883 (0.286)	2.0231 (0.248)	0.5961 (0.113)
	MCMC	-0.5744 (0.273)	2.0341 (0.242)	0.5905 (0.106)
(-0.8, 2, 0.8)	CODE	-0.8037 (0.324)	2.0197 (0.281)	0.8062 (0.118)
	MPLE	-0.7825 (0.323)	2.0267 (0.280)	0.7954 (0.117)
	MCMC	-0.7608 (0.299)	2.0438 (0.267)	0.7874 (0.107)
(-1.0, 2, 1.0)	CODE	-0.9884 (0.360)	2.0400 (0.305)	1.0042 (0.121)
	MPLE	-0.9671 (0.355)	2.0434 (0.302)	0.9925 (0.119)
	MCMC	-0.9352 (0.323)	2.0661 (0.282)	0.9805 (0.106)

Table 2: MSE of estimates from the three estimation methods.

True Parameters				
$(\beta_0, \beta_1, \gamma)$	Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$
(1, 2, 0)	CODE	0.0662	0.0321	0.0088
	MPLE	0.0653	0.0317	0.0087
	MCMC	0.0629	0.0313	0.0083
(0.2, 2, 0.2)	CODE	0.0670	0.0421	0.0104
	MPLE	0.0668	0.0419	0.0103
	MCMC	0.0635	0.0408	0.0098
(-0.2, 2, 0.4)	CODE	0.0700	0.0484	0.0108
	MPLE	0.0698	0.0483	0.0108
	MCMC	0.0658	0.0466	0.0100
(-0.6, 2, 0.6)	CODE	0.0817	0.0619	0.0128
	MPLE	0.0819	0.0621	0.0128
	MCMC	0.0751	0.0595	0.0114
(-0.8, 2, 0.8)	CODE	0.1051	0.0790	0.0140
	MPLE	0.1046	0.0792	0.0138
	MCMC	0.0911	0.0730	0.0116
(-1.0, 2, 1.0)	CODE	0.1294	0.0945	0.0147
	MPLE	0.1269	0.0930	0.0143
	MCMC	0.1085	0.0838	0.0116

Table 3: Estimation results for species No. 38.

Method	Analysis	β_0	TAV	PRCP	MI	γ
Coding	Estimate	-11.552	13.229	-14.460	26.694	7.847
Method	Appr. S.E	3.921	5.722	5.366	9.980	2.510
MPL	Estimate	-7.013	8.200	-10.532	19.143	5.036
Method	Appr. S.E	1.735	3.442	3.696	6.549	1.038
MCMC	Estimate	-4.689	2.002	-2.994	6.574	3.035
Method	Appr. S.E	0.573	0.809	0.869	1.534	0.306

Table 4: Estimation results for species No. 158.

Method	Analysis	β_0	TM	TMM	ELV	γ
Coding	Estimate	-7.765	-4.282	3.994	-0.878	4.535
Method	Appr. S.E	0.881	1.010	1.030	0.293	0.478
MPL	Estimate	-7.047	-3.801	3.579	-0.741	4.041
Method	Appr. S.E	0.749	0.936	0.969	0.263	0.394
MCMC	Estimate	-4.777	-1.640	1.528	-0.318	2.553
Method	Appr. S.E	0.265	0.168	0.158	0.044	0.141

Table 5: Fitted errors for species No. 38 and No. 158.

Method	Species No. 38			Species No. 158		
	COD	MPL	MCMC	COD	MPL	MCMC
SAE	404.97	131.988	72.319	728.578	707.195	568.362
SSE	380.67	118.111	42.166	699.483	665.070	429.932