

Prediction of potential areas of species distributions based on presence-only data

JORGE A. ARGÁEZ,^{1,2} J. ANDRÉS CHRISTEN,¹
MIGUEL NAKAMURA¹ and JORGE SOBERÓN^{3,4}

¹*Centro de Investigación en Matemáticas, A. C., Apartado Postal 402, Guanajuato, Gto., 36000, México*

²*Centro de Investigación Científica de Yucatán A. C.*

E-mail: jargs@cicy.mx


³*Instituto de Ecología Universidad Nacional Autónoma de México, Liga Periférico Sur 4903, Parque del Pedregal, 14010, México*

⁴*Universidad Nacional Autónoma de México, Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, Liga Periférico Sur 4903, Parque del Pedregal, 14010, México*

Received July 2003; Revised September 2004

We introduce a methodology to infer zones of high potential for the habitat of a species, useful for management of biodiversity, conservation, biogeography, ecology, or sustainable use. Inference is based on a set of sites where the presence of the species has been reported. Each site is associated with covariate values, measured on discrete scales. We compute the predictive probability that the species is present at each node of a regular grid. Possible spatial bias for sites of presence is accounted for. Since the resulting posterior distribution does not have a closed form, a Markov chain Monte Carlo (MCMC) algorithm is implemented. However, we also describe an approximation to the posterior distribution, which avoids MCMC. Relevant features of the approach are that specific notions of data acquisition such as sampling intensity and detectability are accounted for, and that available *a priori* information regarding areas of distribution of the species is incorporated in a clear-cut way. These concepts, arising in the presence-only context, are not addressed in alternative methods. We also consider an uncertainty map, which measures the variability for the predictive probability at each node on the grid. A simulation study is carried out to test and compare our approach with other standard methods. Two case studies are also presented.

Keywords: biodiversity, ecology, mixture model, predictive probability map, prior elicitation

1352-8505 © 2005  Springer Science+Business Media, Inc.

1. Introduction

All species of animals and plants occupy a more or less well-defined geographical region, called their areas, or ranges, of distribution. It is one of the most fundamental

1352-8505 © 2005  Springer Science+Business Media, Inc.

expressions of its ecology and evolutionary history (Udvardy, 1969; Brown, Stevens and Kaufman, 1996; Gaston and Blackburn, 2000). A species distribution is the product of a complex combination of processes (Gaston and Blackburn, 2000), that begin with the apparition of the species by evolution, and its movements in ecological and geological times. The distribution is determined by factors affecting: (1) the regions in space where the right autoecological conditions for the survival of the species (climate, soil, energy availability, etc.) are met, (2) the regions of space that are available to the dispersal of the species, and (3) the presence or absence of crucial population interactions, that is, key competitors, predators, diseases, and mutualists (seed dispersers, pollinators, and so on). It is quite likely that at different spatial and temporal scales this complex of factors operate with different intensity.

Detailed knowledge of distribution areas is relevant to address basic questions in biogeography and ecology, but it is also useful in the management of biodiversity for conservation or sustainable use. Problems like the relative roles of ecological and historical factors in shaping them, how the shape of the area changes with the spatial scale of observation, and the relative importance of the local (or alpha) and the turnover (or beta) components of biodiversity, depend on being able to estimate in detail such areas. Moreover, such detailed knowledge can be used to determine the areas best suited for conservation, the regions where given activities can endanger protected species, and so on.

Unfortunately, for most species, the knowledge biologists have about distribution areas is very rough and often reduced to the few localities where a species has been observed. Ecologists and biogeographers determine the area of distribution by starting with “points” (in practice, localities) where the species has been registered or observed. A number of informal procedures are used to extrapolate from a cloud of points in the geographical space to a set of polygons that represent the area of distribution (Jennrich and Turner, 1969; Udvardy, 1969; Rapoport, 1975). Generally speaking, such extrapolation is entirely based on the field experience of the researchers and it is done at a very rough scale. The fundamental data that biogeographers use to base their extrapolations are the presence points. Detailed faunistic or floristic studies yield lists of observations of species in localities. Absences are mostly inferred from knowledge of the biology of the species, or from the experience of field biologists. Thus, an important feature of the data available in this setting is that one may only be certain of sites of presence, whereas sites of absence are not readily available. The problem in this paper is to infer zones of high probability for the habitat of species, using only reported sites of presence. When legitimate sites of absence are also available, this problem can be approached from many directions, such as generalized linear models (Austin, 2002), autologistic models (Pettitt, Weir and Hart, 2002), or Kriging (Heagerty and Lele, 1998).

We explicitly assume that (1) the scale is such that all the high similarity areas can be reached by the species, and that (2) the biotic interactions can be ignored. If the first assumption is false, our procedure will calculate the “fundamental niche” of the species (Peterson, Stockwell and Kluza, 2002). How important is the second assumption is a matter of empirical research and will depend on the specific cases.

We propose a Bayesian methodology for quantifying the probability that the species is present at each site, given that the sites in the region possess a known set of physical characteristics: the *covariates*. This probability will be estimated using

information on sites where individuals of the given species have been detected, with the capability of incorporating available prior knowledge.

An important feature in this setting is the fact that detected sites of presence typically occur clustered around roads, or near populated areas. From now on we refer to this as *spatial bias*. It describes heterogeneous distributions of sampled points, in a geographical sense. Since each site has associated values of additional covariates, any geographical distribution of points induces a distribution of points in the covariate space. These points may also be non-uniformly distributed, so in addition, a notion of *covariate bias* is relevant. Clearly, covariate bias depends on the nature of spatial bias and on the distribution of covariates over the whole region of interest. Although spatial correlation may also be present, we do not consider it explicitly in our approach because the nodes are sufficiently large so as to safely disregard local correlations. Spatial correlation is nevertheless indirectly considered through the covariates and spatial bias.

Assessment of potential zones of presence is based on values of the covariates. This means that even if samples were spatially biased, it is possible that they represent sampled covariates that are unbiased. However, in general we must allow for the fact that covariate bias may be present, induced by spatial bias. Covariate bias hence governs the probability that a site with a given set of covariate values appears as a physically examined site, over the period of observation considered.

In addition, there is the notion of *detectability* of a species. Even if the species is present at a physically examined site, the species may not be detected. Detectability is an intrinsic property of the species, for a given observation procedure implemented in the field. This is interpreted as the probability of detecting the presence of a species, given that it is present at an observed site. *Probability of observation* refers to the probability of actually registering the presence of a species at a site, once probability of presence, covariate bias, and detectability have been accounted for.

Some methods do exist and have been extensively used for constructing maps of distribution areas. However, few are formulated in statistical terms, and none appear to adequately take into account the available prior information. Methods mostly used are: *Bioclim* (Busby, 1991), *Domain* (Carpenter, Gillison and Winter, 1993), *FloraMap* (Jones and Gladkov, 1999), and *GARP* (Stockwell and Noble, 1991; Peterson and Cohoon, 1999; Stockwell and Peters, 1999; Peterson, Stockwell and Kluza, 2002). These algorithms are becoming increasingly popular, not only to address scientific questions (Peterson, Soberón, and Sánchez-Cordero, 1999), but also to estimate routes of entrance of invasive species (Soberón, Golubov and Sarukhán, 2001), risk of damage by plague species (Sánchez-Cordero and Martínez-Meyer, 2000), and other applied questions.

In all the above algorithms the opinion or knowledge of experts is used, *a posteriori* and informally, to correct blatant errors, mostly overprediction. The experts often reduce the surfaces predicted by the methods without resorting to explicit algorithm or criteria. This practice suggests that in applications, prior knowledge or expert opinion is indeed taken into consideration, although not transparently. One important aspect of the approach we consider in this paper is that prior knowledge is readily recognized and utilized in a clear-cut way for the production of relevant maps. In addition to establishing statistical inference for the map of probabilities of

presence, we propose a map of uncertainty, which allows for greater insight into the nature of potential areas of distribution.

2. The statistical model

2.1. Notation

Let \mathcal{R} be the set of nodes determined by a regular, square grid, which covers the region of interest. The probability of potential habitation at $s \in \mathcal{R}$ usually represents potential over a square centered on s , taken to be the same size as a square on the grid. For each $s \in \mathcal{R}$, an M -dimensional vector $\mathbf{e}(s) = (e_1(s), \dots, e_M(s))$ of covariates is assumed to be known. The M covariates are considered either categorical, or measured on discrete scales, so $e_k(s) \in \{1, \dots, R_k\}$, $1 \leq k \leq M$, being R_k the number of classes of the k th covariate. The set of all conceivable covariate configurations is $F = \{1, \dots, R_1\} \times \dots \times \{1, \dots, R_M\}$ (although many of them may not actually occur over \mathcal{R}), so that $\#(F) = \prod_{k=1}^M R_k$.

Observed data consists of n nodes, s_1, \dots, s_n , corresponding to n exact geographical locations of positive observation, that have been identified with the nearest $s \in \mathcal{R}$. Some of these nodes may be multiple, since two or more observations may have occurred at different locations sharing the same center. For $f \in F$, we denote by $C(f)$ the number of nodes in the sample such that $\mathbf{e}(s_i) = f$, $1 \leq i \leq n$. Let $\mathbf{C} = (C(f))_{f \in F}$ be the vector of all counts, arranged according to F 's lexicographical order. The vector \mathbf{C} summarizes observed data, and any modeling approach should be aimed at describing the probabilistic behavior of \mathbf{C} . Notice that $\sum_{f \in F} C(f) = n$, and that many of the $C(f)$ may actually be zero, since n is usually very small relative to $\#(F)$. A model parameterized to account for all $f \in F$ would be inconvenient in that it would pose an estimation problem with sparse data. Reduction in parameter dimensionality based on pairwise interactions between covariates will be considered and appropriate notation is required.

Let G be the set of index pairs (a, b) , $1 \leq a < b \leq M$, $J = (a, b)$ be a generic pair in G , $\mathbf{e}_J(s) = (e_a(s), e_b(s))$, and $F_J = \{1, \dots, R_a\} \times \{1, \dots, R_b\}$. For $g \in F_J$, let $C_J(g)$ be the number of nodes in the sample such that $\mathbf{e}_J(s_i) = g$, and $\mathbf{C}_J = (C_J(g))_{g \in F_J}$. Let p_s be a binary random variable which takes on the value 1 if the species is present at s , and the value 0 otherwise. The map of probabilities of presence for the species over \mathcal{R} is the probability $P(p_s = 1)$, as a function of s . A fundamental underlying notion is that presence is determined by covariates, rather than geographical location. Let $\mathbf{U} = (U_1, \dots, U_M)$ be the random vector of covariate values tacitly selected by the species when it makes itself present. The fundamental assumption that justifies inference of areas of high potential from reported sites of presence via the consideration of covariates is that $P(p_s = 1) = P(\mathbf{U} = \mathbf{e}(s))$. By simplifying this assumption, postulating that presence is determined only by the corresponding value $\mathbf{U}_J = (U_a, U_b)$ of the pair J , this translates to

$$P(p_s = 1|J) = P(\mathbf{U}_J = \mathbf{e}_J(s)|J). \quad (1)$$

For $g \in F_J$ define $\theta_J(g) = P(\mathbf{U}_J = g | J)$, and $\boldsymbol{\theta}_J = (\theta_J(g))_{g \in F_J}$ ($\boldsymbol{\theta}_J$ specifies the density of \mathbf{U}_J). To incorporate sampling bias, let $\delta(s)$ denote the probability that in a random excursion to the field, node s is examined for presence. This is spatial bias, and induces “covariate bias”, which we denote by $v_J(g)$. This last quantity is the probability that in a single outing, a node having value g for the covariate pair J is physically examined for presence. The relationship between spatial and covariate biases is

$$v_J(g) = \sum_{\{s: \mathbf{e}_J(s)=g\}} \delta(s). \quad (2)$$

As we have noted, detectability is an inherent property of the species, but it may depend on $\mathbf{e}(s)$. In what follows, for notational simplicity and because we believe it to be reasonable, we consider constant detectability d , that is, d is the probability of detecting a species given that it is present at a visited node. Considerations and notations may be defined to allow for non-constant detections as well, but we do not address them here. We make a few comments in Section 6 regarding these assumptions.

If o_s denotes a binary variable that takes on the value 1 if a species is observed at node s , and 0 otherwise, we have that

$$P(o_s = 1 | J) = P(\mathbf{U}_J = \mathbf{e}_J(s) | J) v_J(\mathbf{e}_J(s)) d. \quad (3)$$

The probability of presence, $P(\mathbf{U}_J = \mathbf{e}_J(s) | J)$, will not be identifiable without first discerning $v_J(\mathbf{e}_J(s))$ and d . Our method will assume that both of these quantities are given as inputs. Regarding $v_J(\mathbf{e}_J(s))$, we assume this is given exactly via (2) and specification of $\delta(s)$, which is either assumed to be spatially uniform, or to be given input generated by the user by previous means.

Notice that what is indeed random and observed is \mathbf{U}_J , the value of the covariate pair at a recorded site of presence, rather than o_s itself, which is fixed at the value 1 as a consequence of design. By incorporating the parameterization, and using (1) and (3), we obtain

$$P(o_s = 1 | \boldsymbol{\theta}_J, J) = P(\mathbf{U}_J = \mathbf{e}_J(s) | \boldsymbol{\theta}_J, J) v_J(\mathbf{e}_J(s)) d. \quad (4)$$

The notation $\mathbf{z} \boldsymbol{\theta}' = (\boldsymbol{\theta}_J)_{J \in G}$ and $\mathbf{C}' = (\mathbf{C}_J)_{J \in G}$ will also prove to be useful to denote pairwise parameters and observed pairwise counts. Dimensionality will be reduced if $\sum_{a < b} R_a R_b < \#(F)$, which is typical in our context because the R_k 's are not small.

2.2. Formulation

Consider a fixed pair J , and let N be the total number of nodes examined in the timeframe considered, that gave rise to the n nodes of presence. Temporarily assume N is known. If the N sampled nodes can be considered independent (if $\boldsymbol{\theta}_J$ is assumed as a random variable a weaker assumption of exchangeability may be used), each one can be viewed as having been randomly grouped into one of $\#(F_J) + 1$ bins. The first $\#(F_J)$ bins have the possible values of $g \in F_J$ as labels, and being classified into one of these bins signifies $o_s = 1$. The last bin corresponds to a node having resulted

in $o_s=0$. By (4), the probability of a node being classified into bin labeled g is $\theta_J(g) v_J(g) d$. This constitutes a standard multinomial setting, so that if $\mathbf{c}_J = (c_J(g))_{g \in F_J}$ is a vector such that $\sum_{g \in F_J} c_J(g) \leq N$, then

$$P(\mathbf{C}_J = \mathbf{c}_J | \theta_J, J) = \tau_J \left\{ 1 - \sum_{g \in F_J} \theta_J(g) v_J(g) d \right\}^{N - \sum_{g \in F_J} c_J(g)} \prod_{g \in F_J} \{ \theta_J(g) v_J(g) d \}^{c_J(g)}, \quad (5)$$

where τ_J is the normalizing constant $N! \{ \prod_{g \in F_J} c_J(g)! \}^{-1} [N - \sum_{g \in F_J} c_J(g)]!^{-1}$. For \mathbf{c} a vector of counts such that $\sum_{f \in F} c(f) \leq N$, we postulate the following model for \mathbf{C} :

$$P(\mathbf{C} = \mathbf{c} | \theta') = \sum_{J \in G} \pi(J) [k_J(\mathbf{c}_J, N)]^{-1} P(\mathbf{C}_J = \mathbf{c}_J | \theta_J, J), \quad (6)$$

where the constant $k_J(\mathbf{c}_J, N)$ is the number of different \mathbf{c} configurations that give rise to the same \mathbf{c}_J (this number does not depend on θ'), and $\pi(J)$ is a probability mass function over G . It is not hard to show that $\sum_{\{\mathbf{c}: \sum_{f \in F} c(f) \leq N\}} P(\mathbf{C} = \mathbf{c} | \theta') = 1$.

One interpretation of model (6) is probabilistic, based on the notion of mixing. A species is thought of as selecting a pair J at random from G , with probability $\pi(J)$. Conditioned on J , the probability of presence at any site s is specified by $\theta_J(\mathbf{e}_J(s))$. This induces a multinomial count structure for \mathbf{C}_J , which in turn induces a count structure for \mathbf{C} (namely, uniform probability is assigned to all values of \mathbf{C} that produce counts \mathbf{C}_J).

The distribution $\pi(J)$ may be thought of as summarizing the idiosyncrasy of the species with regard to its appraisal of a site according to covariates. The relatively simple structure of model (6) (pairs of covariates) is compatible with a principle stating that species focus on a small set of attributes and simple criteria when deciding a site for colonization. For example, it is known that for the GARP algorithm, more than about five variables do not add much predictive power (Peterson and Cohoon, 1999). Although sensible, this principle will require experimental testing, and our model could provide a contrasting hypothesis for such testing.

Regarding N , it is not the rule that a full record of visited sites is kept, especially considering historical data, and thus N must be considered to be unknown. However, for each J we expect that $C_J(g) \approx N \theta_J(g) v_J(g) d$ (for large N), and since $\sum_{g \in F_J} \theta_J(g) = 1$, we must have $N \approx N_J = \left\lceil \sum_{g \in F_J} C_J(g) v_J^{-1}(g) d^{-1} \right\rceil$ for all J . A simple way to proceed, as we do in the following Sections, is to postulate $N = \{\#(G)\}^{-1} \sum_{J \in G} N_J$ as a working approximation in (6), rather than considering N itself to be an unknown nuisance parameter.

3. Inference

3.1. Predictive probability

For each pair of covariates, a prior distribution, $f(\theta_J)$, is postulated for the parameter θ_J . A way to proceed is to consider J as a parameter (random variable),

and take the $\pi(J)$'s as its prior distribution. This is the usual procedure in the Bayesian analysis of mixture models (inclusion of a further hierarchy by taking the $\pi(J)$'s as random is irrelevant because we are assuming an arbitrary distribution for J). The elicitation of $f(\theta_J)$ and $\pi(J)$ is discussed in Section 2. As before, let $P(\mathbf{C}_J | \theta_J, J)$ denote the multinomial model (5), and $f(\theta_J | \mathbf{C}_J, J)$ denote the posterior distribution of θ_J given J . Notation $\pi(J | \mathbf{C}')$ is used for the posterior probability for pair J .

The law of total probability yields $P(p_s = 1 | \mathbf{C}') = \sum_{J \in G} P(p_s = 1 | \mathbf{C}', J) \pi(J | \mathbf{C}')$. The quantity $P(p_s = 1 | \mathbf{C}', J)$ is the predictive probability of presence given J , and is calculated by $\int P(p_s = 1 | \theta_J, J) f(\theta_J | \mathbf{C}', J) d\theta_J$. Since $P(p_s = 1 | \theta_J, J) = \theta_J(\mathbf{e}_J(s))$, we obtain by substitution that $P(p_s = 1 | \mathbf{C}', J) = E[\theta_J(\mathbf{e}_J(s)) | \mathbf{C}', J]$. Thus, the predictive probability at node s is given by

$$P(p_s = 1 | \mathbf{C}') = \sum_{J \in G} E[\theta_J(\mathbf{e}_J(s)) | \mathbf{C}', J] \pi(J | \mathbf{C}'). \quad (7)$$

For each pair J we postulate a Dirichlet distribution as prior for θ_J , whose expression is $f(\theta_J | J) = \Gamma(\alpha_J) \left[\prod_{g \in F_J} \Gamma(\alpha_J(g)) \right]^{-1} \prod_{g \in F_J} \theta_J(g)^{\alpha_J(g)-1}$, where $\alpha_J = \sum_{g \in F_J} \alpha_J(g)$, $\alpha_J(g) > 0$. The parameter for this distribution is $\alpha'_J = (\alpha_J(g))_{g \in F_J}$. However, there is no standard closed form for $f(\theta_J | \mathbf{C}_J, J)$ resulting from the model (5), and a Dirichlet prior, under the expressions for the bin probabilities (see Equation (A.1)). Therefore one needs to resort to numerical methods (MCMC, see Appendix A) to simulate values θ_J from $f(\theta_J | \mathbf{C}_J, J)$ to obtain the quantities $E[\theta_J(\mathbf{e}_J(s)) | \mathbf{C}', J]$, and $\pi(J | \mathbf{C}')$ involved in (7). The quantity $\pi(J | \mathbf{C}')$ can be interpreted as the posterior probability that the species assigns to pair J in its preference about colonizing \mathcal{R} .

However we discovered an alternative to avoid MCMC, by taking a Dirichlet with parameters $\mathbf{X}_J^* + \alpha_J$ as an approximation to the exact posterior distribution, where $\mathbf{X}_J^* = (X_J^*(g))_{g \in F_J}$, with $X_J^*(g) = C_J(g)(v_J(g)d)^{-1}$. Inspired by the observation that $X_J^*(g)$ represents an approximation of the actual multinomial count related to the cell probability $\theta_J(g)$, we would obtain the mentioned Dirichlet as a ‘‘posterior’’. Note that $X_J^*(g)$ may not be integer and the rigorous consideration of an alternative model of this type for the $C_J(g)$'s would entail the identification of an unknown normalization constant dependant on θ_J , thus prohibiting inference. Therefore, the $X_J^*(e_J)$'s should only be considered as a device for obtaining our approximation to the actual posterior. The required expected value in (7) is given by $[X_J^*(\mathbf{e}_J(s)) + \alpha_J(\mathbf{e}_J(s))] [N + \alpha_J]^{-1}$. The closed-form calculation of $\pi(J | \mathbf{C}')$ is shown in Appendix B. As far as the approximation is concerned, what is relevant is that, by examining the distributions $f(\theta_J | \mathbf{C}_J, J)$ and $f(\theta_J | \mathbf{X}_J^*, J)$, we observe (numerically) that $E[\theta_J(\mathbf{e}_J(s)) | \mathbf{C}_J, J]$ and $E[\theta_J(\mathbf{e}_J(s)) | \mathbf{X}_J^*, J]$ are virtually equal. Certainly, the mathematical tractability of $f(\theta_J | \mathbf{X}_J^*, J)$ (a Dirichlet) is more appealing. We compare both approaches in Section 4. It is relevant to note that a precise probabilistic definition for the concept of ‘‘potential’’ at node s has been established: the predictive probability given by (7).

In order to display the resulting map, we consider the arbitrary partition $I_j = ((j-1)/10, j/10]$, $1 \leq j \leq 10$, and a gray-scale to plot the predictive probability $P(p_s = 1 | \mathbf{C}')$ at each node, in accord with interval I_{j_s} where $P(p_s = 1 | \mathbf{C}') \in I_{j_s}$. We also evaluate $I'(s) = \int_{I_{j_s}} f(\theta_J(\mathbf{e}_J(s)) | \mathbf{C}', J) d\theta_J(\mathbf{e}_J(s))$. The quantity $I'(s)$ is the

posterior probability that $\theta_J(\mathbf{e}_J(s))$ lies in I_{J_s} , and motivated by (7), the quantity $I(s) = \sum_{J \in G} I^J(s) \pi(J|C')$ provides a level of certainty about the potential plotted in the first map. It depends both on the posterior distribution of each J , and on the partition used to display the first map, and may also be displayed using a gray-scale on the same partition. Maps of uncertainty obtained with MCMC and the approximation were qualitatively equivalent, even for the most extreme cases (small n , non-informative prior, and non-homogeneous bias).

The consideration of a measure of uncertainty in maps may be found in just a handful of papers, varying in flavor and presentation (see for example, Heikkinen and Högmänder, 1994; Högmänder and Möller, 1995; De Oliveira, 2000). In our experience, the usage of the map of uncertainty (or certainty) helps in the interpretation and understanding of the posterior distribution at hand, and enables more educated conclusions.

3.2. Prior elicitation

For the Dirichlet distribution it is a fact that $\alpha_J(g) = \alpha_J E[\theta_J(g)]$, where $E[\theta_J(g)]$ is the prior expected value for $\theta_J(g)$. That is, the values α_J and $E[\theta_J(g)]$ should be elicited in order to fix the $\alpha_J(g)$. It would be unusual that the expert provides values directly for these quantities, and a heuristic procedure to obtain them indirectly is proposed. We ask the user, based on prior experience and knowledge about the species (but not using data at hand), to divide \mathcal{R} into disjoint regions: region \mathcal{P} , where it is very likely that the species is present, and region \mathcal{A} , where it is very unlikely that the species is present. The complement, \mathcal{I} , is implicitly defined, and represents a region of ambiguity (see Fig. 3(a)). Either \mathcal{P} , \mathcal{A} , or both may be empty (see Fig. 4(a)).

Consider arbitrary nodes $s_1 \in \mathcal{P}$, $s_2 \in \mathcal{A}$, and $s_3 \in \mathcal{I}$. If $\mathbf{e}_J(s_1) = \mathbf{e}_J(s_2) = \mathbf{e}_J(s_3)$, then s_1, s_2, s_3 are called a 3-way contradiction, in the sense that the user's assessment is putting the same covariate values in areas with different *a priori* meaning. Region \mathcal{R} is examined until a 3-way contradiction (if any) is found, and the three nodes involved are excluded. The examination is repeated, each time with the remaining nodes, until 3-way contradictions are exhausted. Let $\mathcal{R}_2 \subseteq \mathcal{R}$ be the resulting set. Within \mathcal{R}_2 there can be other contradictions: If nodes $s_1 \in \mathcal{P} \cap \mathcal{R}_2$, $s_2 \in \mathcal{A} \cap \mathcal{R}_2$ (or $s_1 \in \mathcal{P} \cap \mathcal{R}_2$, $s_2 \in \mathcal{I} \cap \mathcal{R}_2$, or $s_1 \in \mathcal{A} \cap \mathcal{R}_2$, $s_2 \in \mathcal{I} \cap \mathcal{R}_2$) are such that $\mathbf{e}_J(s_1) = \mathbf{e}_J(s_2)$, then s_1, s_2 are called a 2-way contradiction. Following a similar procedure, 2-way contradictions are removed from \mathcal{R}_2 , and the remaining nodes conform the set \mathcal{R}_1 of non-contradictory nodes. Notice that \mathcal{R}_1 is not uniquely determined, because a node can be involved in several 3-way and/or 2-way contradictions, and the order in which contradictions are excluded is arbitrary. Nevertheless, the relevant information contained in \mathcal{R}_1 is $\#(\mathcal{R}_1)$, which is independent of the elimination sequence.

The set \mathcal{R}_1 contains the non-contradictory information in the covariates given by the user. One interpretation of parameter $\alpha_J > 0$ is the amount of information contained in the prior distribution (Gelman *et al.*, 1995, p. 76). Since the relevant information for establishment of the species depends on values of the covariates, we are thus motivated to define $\alpha_J = \#(\mathcal{R}_1) [\#(\mathcal{R} \setminus \mathcal{R}_1)]^{-1}$, which takes on values in the range $(0, \infty)$.

Regarding elicitation of $E[\theta_J(g)]$, the idea is to determine the probability of presence for each $g \in F_J$ that the user has (implicitly) specified by delimiting \mathcal{P}, \mathcal{A} ,

and \mathcal{I} . By postulating that “very likely” and “very unlikely” in the query above signify probabilities of 0.95 for \mathcal{P} , 0.05 for \mathcal{A} , and 0.5 for \mathcal{I} (denoting ambiguity), we define

$$w_J(g) = \frac{(0.95)\#\{s \in \mathcal{P} : \mathbf{e}_J(s) = g\} + (0.5)\#\{s \in \mathcal{I} : \mathbf{e}_J(s) = g\} + (0.05)\#\{s \in \mathcal{A} : \mathbf{e}_J(s) = g\}}{\#\{s \in \mathcal{R} : \mathbf{e}_J(s) = g\}}.$$

Using these values we establish $E[\theta_J(g)] = w_J(g) \left[\sum_{g' \in F_J} w_J(g') \right]^{-1}$. Finally, since α_J is the quantity of information contained in the prior for each J , a sensible value for $\pi(J)$ is found by normalizing the α_J 's: $\pi(J) = \alpha_J \left[\sum_{J' \in G} \alpha_{J'} \right]^{-1}$. Heuristic verification that elicitation is made sensibly is to calculate the *a priori* map of potential by means of $P(p_s = 1) = \sum_{J \in G} E[\theta_J(\mathbf{e}_J(s))] \pi(J)$. By inspection, we verify that contours of $P(p_s = 1)$ roughly coincide with the areas \mathcal{P} , \mathcal{A} , and \mathcal{I} established by the user (compare Figs 3(a)–(b), and 4(a)–(b)). In the absence of prior information (that is, $\mathcal{I} = \mathcal{R}$), one would set $\alpha_J(g) = (R_a R_b)^{-1}$, a well accepted non-informative prior. For further details of the elicitation process, see Argáez, Christen and Nakamura (in preparation).

4. Simulation study

The physical region and corresponding covariates are quite real – the Yucatan Peninsula in Mexico – but the actual sites of presence of a fictitious species were simulated. A regular grid of 761 nodes, separated approximately by 12 km, covers this region (scale 1:1,000,000). Three covariates are considered on this grid: mean temperature (5 levels), mean rainfall (10 levels), and vegetation type (11 levels).

Our fictitious species is postulated to prefer an “ideal” climate $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$. In order to prescribe how probability of presence depends on $\mathbf{e}(s)$, and to incorporate the notion that the species’ probability of presence decreases as the climate departs from $\boldsymbol{\mu}$, we set $P^*(p_s = 1) = \exp\{-0.5[\boldsymbol{\mu} - \mathbf{e}(s)]^T \mathbf{A} [\boldsymbol{\mu} - \mathbf{e}(s)]\}$ in the simulations. Note that model P^* is not a member of the family of models we have developed in our methodology. This is intentional. The proposed methodology, when fed with simulated data from (7), gives satisfactory results, and we do not document those examples here. Instead, we exemplify how procedures react to data generated by alternative realities that represent types of maps typically latent in biological applications. The symmetric matrix $\mathbf{A} = (a_{hl})$, $1 \leq h, l \leq 3$, allows for structure regarding interactions in the components of $\mathbf{e}(s)$. By varying $\boldsymbol{\mu}$ and \mathbf{A} we are able to simulate species with different degrees of sensitivity to an ideal climate. In the simulation study the function $P^*(p_s = 1)$ may be regarded as “reality”. Spatial bias is obtained by assigning a probability of visiting a node as inversely proportional to its distance to the nearest road, and covariate bias is defined by using the expression (2). We reproduce this fact by considering main highways on the Peninsula.

Data for simulations were generated as follows: a species is present at a node s according to probability $P^*(p_s = 1)$, the site s is visited by human observers with a probability inversely proportional to the distance from s to the nearest road (spatial bias), and an observation of the species is recorded with probability d (d is fixed at 1 from now on). Spatial bias is tuned in the simulations so that the (random) number n

has a desired order of magnitude. This simulation scheme produces spatial clustering that is strikingly akin to actual observed records of presence for species.

We compare our results with “reality”, and with results obtained with the alternative methods FloraMap and Domain. In addition, we produce the uncertainty map as explained in Section 1. Maps of potential using Bioclim and GARP were also obtained, but are not presented here, because these methods output practically all of the Yucatan Peninsula as high potential in all cases. We only display two representative examples. The first example represents a species with high sensitivity ($a_{11} = a_{22} = a_{33} = 1$, $a_{12} = a_{23} = 0.9$, $a_{13} = 0.85$), and the second example a species with low sensitivity ($a_{11} = a_{22} = a_{33} = 1$, $a_{12} = 0.6$, $a_{13} = 0.3$, $a_{23} = 0.1$). We only display a few representative figures. Additional figures are made available on the world wide web, <http://www.cimat.mx/~nakamura/potential.html>. In what follows these figures will be denoted by the prefix “W” (software currently under development will also be available there).

The idealized potential may be found in Figs 1(a) and 2(a). In both cases, the scenarios are difficult, in that there is spatial bias, non-informative prior information, and a small sample size. In Figs 1(b) and 2(b) the estimated potential map for each scenario is depicted. In both figures the presence of record sites located far away

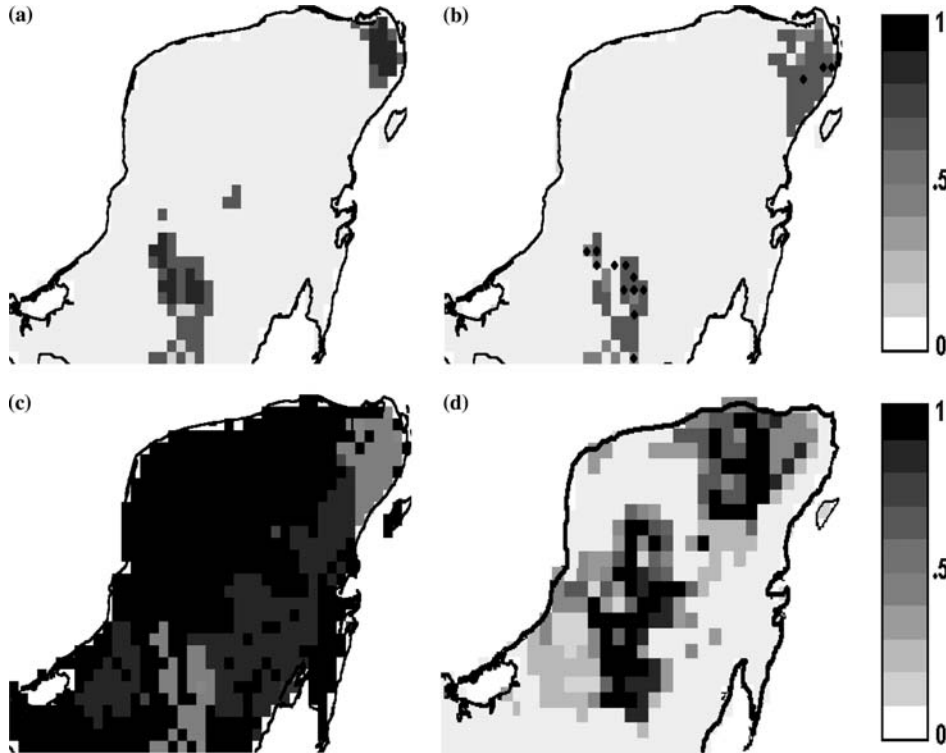


Figure 1. (a) Idealized potential produced by $P^*(p_s = 1)$. (b) Simulated points of presence ($n = 15$), and estimated potential using our method. (c) Map of uncertainty for the estimated potential using the Dirichlet approximation. (d) Estimated potential using FloraMap.

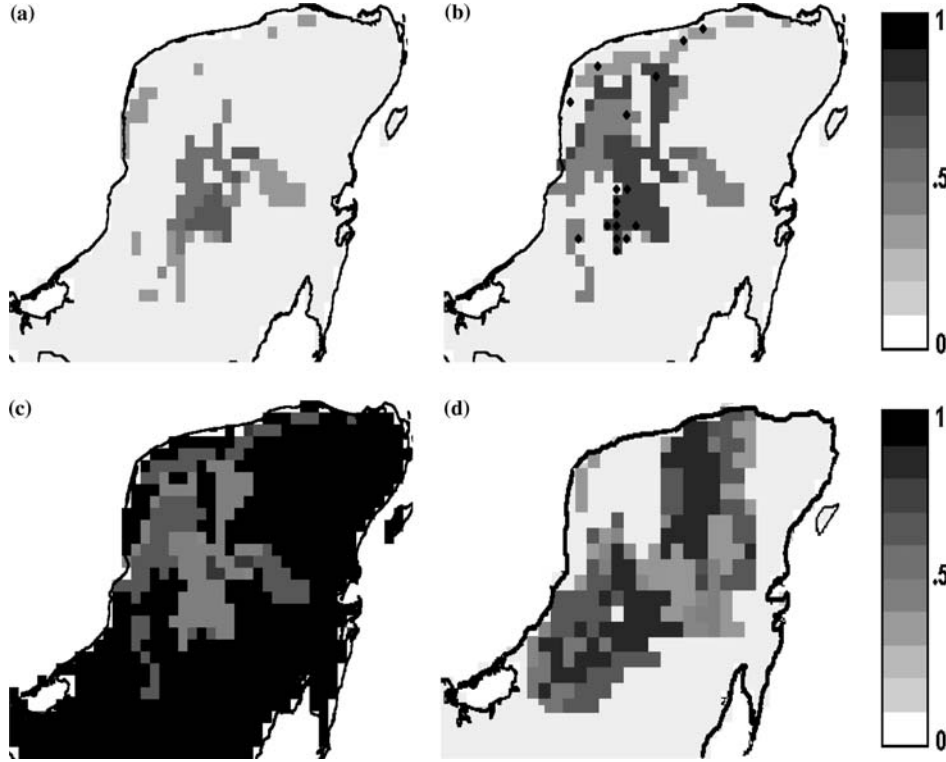


Figure 2. (a) Idealized potential produced by $P^*(p_s = 1)$. (b) Simulated points of presence ($n=20$), and estimated potential using our method. (c) Map of uncertainty for the estimated potential using the Dirichlet approximation. (d) Estimated potential using FloraMap.

from the actual high potential area are noted. Our method does not produce a high potential area around those sites, unlike FloraMap (Figs 1(d) and 2(d)) or Domain (Figs W1(f) and W2(f)). Moreover, maps depicted in Figs 1(c) and 2(c) show a low level of uncertainty for those sites. We also observe that low potential probability areas are associated with a low level of uncertainty. Our uncertainty maps depict that potential probabilities of about 0.5 are associated with the highest levels of uncertainty, resembling the standard setting of estimation of a binomial proportion.

An extensive simulation study may be found in Argáez, Christen and Nakamura (in preparation). Our methodology appears to behave correctly in all reasonable situations and also seems to be robust to isolated sites of presence located far away (geographically speaking) from the main area of high potential. These sites prompted FloraMap and Domain into determining high potential for a significant area around these points, inducing over-estimation. The region of high potential is recovered reasonably well despite the clustering of points of presence, so the method also appears to be robust to the spatial bias introduced by roads and towns. As expected, when n increases, the map of uncertainty tends to a region with low uncertainty.

Regarding the differences for the maps of uncertainty produced by the exact posterior (simulated using MCMC) and with the Dirichlet approximation, the maps

of uncertainty do not appear to have substantial differences that would lead to qualitatively different interpretations (see Figs W1(c)–(d) and W2(c)–(d)).

5. Case studies

5.1. *Coccothrinax readii*

The region of interest is the Yucatan Peninsula. The species under study, *Coccothrinax readii*, is an endemic plant belonging to the *palmeaceae* family, regarded as an endangered species. This species has been reported in 67 localities. The regular grid is as described in Section 4, and the matrix containing the values of covariates for each node of the grid was obtained from the Centro de Investigación Científica de Yucatán (CICY). The physical covariates used on the grid are: humidity (17 levels), mean temperature (5 levels), mean rainfall (10 levels), type of vegetation (11 levels), and type of soil (17 levels), which produce 10 pairs of covariates.

The *a priori* zones \mathcal{P} and \mathcal{A} , as produced by the expert from CICY, are shown in Fig. 3(a). The resulting map using our method and the uncertainty map are shown in Fig. 3(c) and (d), respectively. In this application, pair J defined by temperature–soil type produces $\pi(J|\mathbf{C}') = 0.9889$, and pair J' defined by humidity–temperature, produces $\pi(J'|\mathbf{C}') = 0.0111$. Other pairs produce a posterior probability less than 0.0002.

The potential map was observed by experts concerned with this species. Their appraisal on these zones of high potential given by our method is that they are quite sensible, unlike FloraMap (Fig. W3(e)), and Domain (Fig. W3(f)). Recent considerations suggest that this species is, at present, expanding its area of distribution. The zones highlighted by our method coincide with the expert's assessment about the areas where it is suspected that the species can colonize. Another comment regards the isolated reported site towards the center of the Peninsula. The validity of that site is actually under discussion. The combination of potential map in Fig. 3(c) (low predictive probability), with the uncertainty map in Fig. 3(d) (low level of uncertainty), leads to the suspicion that this record is anomalous.

5.2. *Baronia brevicornis*

The region of interest is the country of Mexico. *Baronia brevicornis* is a butterfly, which has been reported present in 40 localities. The matrix containing the values of covariates was obtained from the Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO). The regular grid consists of 136,875 nodes, with a separation of 4 km (scale 1:4,000,000). Covariates used on the grid are: climate (50 levels), humidity (9 levels), soil (79 levels), rain (19 levels), mean temperature (15 levels), maximum absolute temperature (18 levels), maximum average temperature (19 levels), minimum absolute temperature (20 levels), minimum average temperature (18 levels), and elevation (5 levels). These covariates lead us to consider 45 pairs. The map of *a priori* information is shown in Fig. 4(a).

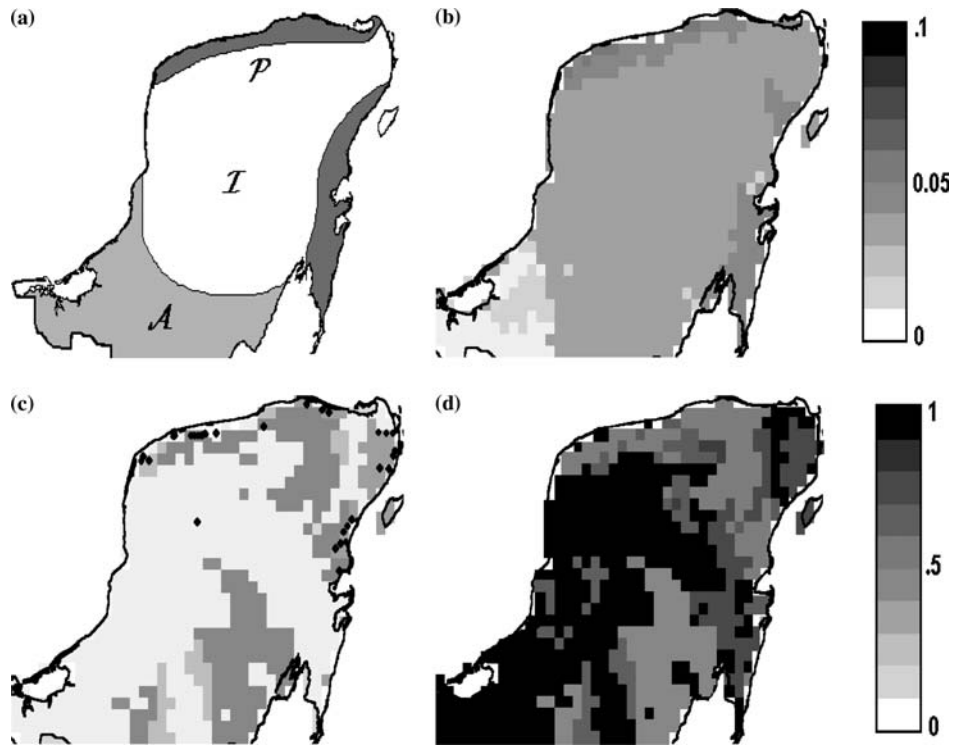


Figure 3. *Coccothrinax readii* (a) *a priori* regions provided by expert. (b) Resulting *a priori* potential. (c) Reported sites of presence ($n=67$), and estimated potential using our method. (d) Map of uncertainty for the estimated potential.

Fig. 4(c) is the potential map, with the sites of presence, and Fig. 4(d) is the map of uncertainty. In this case, the most influential pair is humidity-elevation, with posterior probability 0.999. Based on the field experience of one of us (JSM), Domain (Fig. W4(f)) overpredicts the actual or likely distribution area of *B. brevicornis*, which is a species strictly associated to the tropical deciduous forest, a very particular vegetation type. FloraMap (Fig. W4(e)) produces a slightly less overpredicted surface, but still including large tracts of unsuitable habitat, where the butterfly has never been seen. On the other hand, our method outlined areas where the likelihood of presence of *B. brevicornis* is good, without including obvious unsuitable habitat.

6. Discussion

The methodology postulated here has a series of technical advantages over the existing methodologies. It precisely defines “potential”, has a formal background in statistical inference to support it, has a version simple to implement, identifies and incorporates concepts specific to the genesis of curatorial data, and allows for inclusion of prior information in a convenient way. It might be argued that the

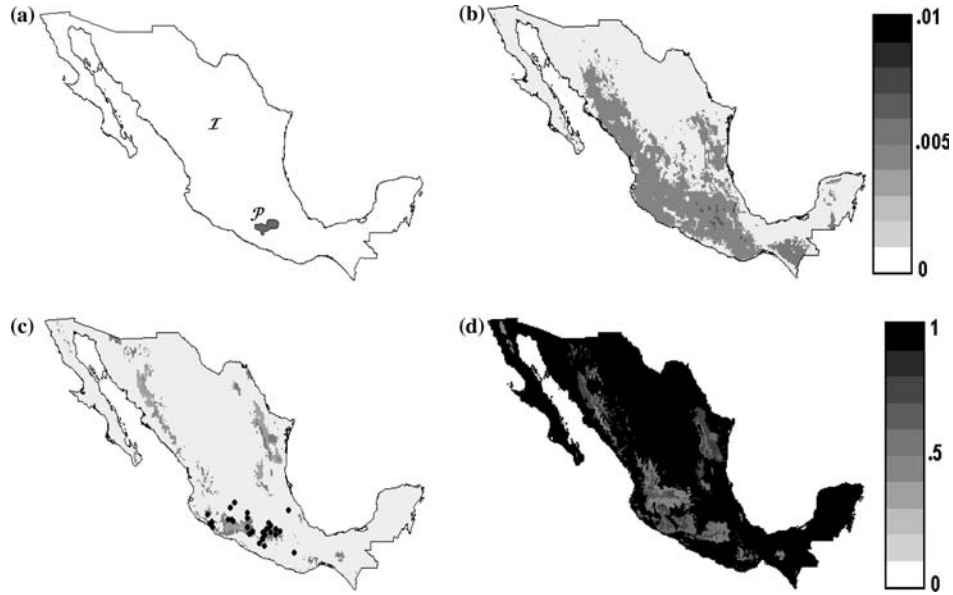


Figure 4. *Baronia brevicornis*. (a) *a priori* regions provided by expert. (b) Resulting *a priori* potential. (c) Reported sites of presence ($n=44$), and estimated potential using our method. (d) Map of uncertainty for the estimated potential.

consideration of only pairs of covariates could be too restrictive. Nevertheless, the mixture model proposed is rather flexible, reasonably parsimonious, and may well approximate higher interactions among covariates. Certainly, the techniques used in this paper may be easily generalized for higher interactions (3, 4-way, etc.), but we are not sure that the additional complexity would reflect in better results.

In a strict sense, Equation (2) may not hold, but the independence alluded does not seem too stringent. One does not intentionally plan to consider nodes having the same covariate pair as candidates for additional examination. Regarding the assumption of constant detectability, statement (1) is also saying that the species tends to make itself present at nodes such that $e_J(s)$ resembles probable values for U_J . Hence, since the species tends to be present at nodes of similar covariate pairs, it is sensible to assume that detectability does not depend on s at all sites where the species is present. An interesting possibility for research, is to establish spatial bias by using accumulated historical data. For example, in studying *B. brevicornis*, one has information on reported sites of presence for several species of butterflies, that may provide substantial information on sampling bias.

Bioclimatic predictive algorithms are becoming indispensable in many areas of ecological work. The need to predict the potential or actual distribution of species is acute in conservation work, invasive species management, bioprospecting, etc. From a user's perspective, the method we present here has several advantages over existing algorithms. In the first place, its Bayesian nature allows the inclusion of a large body of knowledge that experienced biologists have, but that could not be used by previous methodologies. In the second place, the preliminary examples we have analyzed suggest that our method suffers less from overprediction than some existing

alternatives. More work to assess the relative advantage of our method will be required, but our preliminary results are encouraging. Finally, the probabilistic logic of our algorithm is different from the approaches of Domain (clustering), or FloraMap (principal components). Perhaps our method will consistently provide better answers than the alternatives, but if this is not the case, having different tools to tackle the same class of problems will give flexibility to those requiring the prediction of biological species distributions.

Acknowledgments

The authors thank Celene Espadas of the LAB-GIS (CICY), and CONABIO for the data and input for prior elicitation provided for this paper. Argáez was supported by CONACYT Grant 115344. Nakamura and Christen's work was partially supported by CONACYT Grant 32156-E. The authors also thank two referees for interesting comments that resulted in a better manuscript.

Appendix A: MCMC details

A Metropolis–Hasting algorithm is implemented. Model $P(\mathbf{C}_J | \boldsymbol{\theta}_J, J)$ with a Dirichlet prior produces the joint posterior distribution

$$f(\boldsymbol{\theta}_J, J | \mathbf{C}') = \frac{\pi(J)N!\Gamma(\alpha_J)}{(N-n)! \prod_{g \in F_J} c_J(g)! \Gamma(\alpha_J(g))} \left\{ 1 - \sum_{g \in F_J} \theta_J(g) v_J(g) \right\}^{N-n} \quad (A.1)$$

$$\times \prod_{g \in F_J} \theta_J(g)^{c_J(g) + \alpha_J(g) - 1} v_J(g)^{c_J(g)}.$$

With probability p , given the set $(\boldsymbol{\theta}_J)_{J \in G}$ and pair J at iteration t , a candidate J' is selected uniformly from G . We take $J^{(t+1)} = J'$ with probability $\min \{1, \rho_1(J^{(t)}, J')\}$, where

$$\rho_1(J, J') = \frac{\left\{ \prod_{g \in F_{J'}} c_{J'}(g)! \right\} \pi(J') \Gamma(\alpha_{J'}) \prod_{g \in F_{J'}} \Gamma(\alpha_{J'}(g)) \left\{ 1 - \sum_{g \in F_{J'}} \theta_{J'}(g) v_{J'}(g) \right\}^{N-n}}{\left\{ \prod_{g \in F_{J'}} c_J(g)! \right\} \pi(J) \Gamma(\alpha_J) \prod_{g \in F_J} \Gamma(\alpha_J(g)) \left\{ 1 - \sum_{g \in F_J} \theta_J(g) v_J(g) \right\}^{N-n}}$$

$$\times \frac{\prod_{g \in F_{J'}} \theta_{J'}(g)^{c_{J'}(g) + \alpha_{J'}(g) - 1} v_{J'}(g)^{c_{J'}(g)}}{\prod_{g \in F_J} \theta_J(g)^{c_J(g) + \alpha_J(g) - 1} v_J(g)^{c_J(g)}}.$$

On the other hand, with probability $1-p$, given a fixed J , a candidate $\boldsymbol{\theta}'_J$ is selected from the Dirichlet distribution with parameters $\mathbf{X}_J^* + \alpha_J$, the approximation used for the posterior distribution. Since this approximation is commonly good, this

results in a high acceptance rate for this independent proposal and makes the MCMC quite efficient. We take $\boldsymbol{\theta}_J^{(t+1)} = \boldsymbol{\theta}'_J$ with probability $\min\left\{1, \rho_2(\boldsymbol{\theta}_J^{(t)}, \boldsymbol{\theta}'_J)\right\}$, where

$$\rho_2(\boldsymbol{\theta}_J, \boldsymbol{\theta}'_J) = \left(\frac{1 - \sum_{g \in F_J} \theta'_J(g) v_J(g)}{1 - \sum_{g \in F_J} \theta_J(g) v_J(g)} \right)^{N-n} \prod_{g \in F_J} \left(\frac{\theta_J(g)}{\theta'_J(g)} \right)^{X_J^*(g) - c_J(g)}.$$

We arbitrarily chose the value $p = 0.5$.

Appendix B: Approximating $\pi(J|\mathbf{C}')$

It is easy to see that $\pi(J|\mathbf{C}') \propto \pi(J) \int P(\mathbf{C}_J|\boldsymbol{\theta}_J, J) f(\boldsymbol{\theta}_J) d\boldsymbol{\theta}_J$. Having an approximation $f(\boldsymbol{\theta}_J|\mathbf{X}_J^*, J)$ for $f(\boldsymbol{\theta}_J|\mathbf{C}', J)$, by Bayes theorem we see that

$$\int P(\mathbf{C}_J|\boldsymbol{\theta}_J, J) f(\boldsymbol{\theta}_J) d\boldsymbol{\theta}_J \approx \frac{P(\mathbf{C}_J|\boldsymbol{\theta}_J^0, J) f(\boldsymbol{\theta}_J^0)}{f(\boldsymbol{\theta}_J^0|\mathbf{X}_J^*)}$$

for some fixed value $\boldsymbol{\theta}_J^0$ (where the approximation is good). From this we obtain

$$\begin{aligned} \pi(J|\mathbf{C}') &= \pi(J) \frac{N! \Gamma(\alpha_J)}{(N-n)! \Gamma(N + \alpha_J)} \prod_{g \in F_J} \frac{\Gamma(X_J^*(g) + \alpha_J(g)) v_J(g)^{c_J(g)}}{\Gamma(\alpha_J(g))} \\ &\quad \times \left\{ 1 - \sum_{g \in F_J} \theta_J^0(g) v_J(g) \right\}^{N-n} \prod_{g \in F_J} \{\theta_J^0(g)\}^{c_J(g) - X_J^*(g)}. \end{aligned}$$

In the examples we took $\theta_J^0(g) = [X_J^*(g) + \alpha_J(g)] [N + \alpha_J]^{-1}$.

References

- Argáez, J.A., Christen, J.A., and Nakamura, (In Prep) Quantifying information of a priori maps and simulation study. Centro de Investigación en Matemáticas A.C. Guanajuato, Mexico.
- Austin, M.P. (2002) Spatial prediction of species distribution: an inference between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.
- Brown, J., Stevens, G.C., and Kaufman, D.W. (1996) The geographic range: size, shape, boundaries and internal structure. *Annual Review of Ecology and Systematics*, **27**, 597–623.
- Busby, J.R. (1991) BIOCLIM – A bioclimate analysis and prediction system, in *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*, C.R. Margules and M.P. Austin (eds.), CSIRO, Australia, pp. 64–68.
- Carpenter, G., Gillison, A.N., and Winter, J. (1993) Domain: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, **2**, 667–680.
- De Oliveira, V. (2000) Bayesian prediction of clipped gaussian random field. *Computational Statistics and Data Analysis*, **34**, 299–314.

- Gaston, K. and Blackburn, T. (2000) *Pattern and Process in Macroecology*, Blackwell Science, Oxford.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995) *Bayesian Data Analysis*, Chapman & Hall, London.
- Heagerty, P.J. and Lele, S.R. (1998) A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, **93**, 1099–1111.
- Heikkinen, J. and Högmänder, H. (1994) Fully bayesian approach to image restoration with an application in biogeography. *Applied Statistics*, **43**, 569–582.
- Högmänder, H. and Möller, J. (1995) Estimating distribution maps from atlas data using methods of statistical image analysis. *Biometrics*, **51**, 393–404.
- Jennrich, R.I. and Turner, F.B. (1969) Measurement of a non-circular home range. *Journal of Theoretical Biology*, **22**, 227–237.
- Jones, P.G. and Gladkov, A. (1999) *FloraMap: a computer tool for predicting the distribution of plants and other organisms in the wild*; version 1, 1999, Annie L. Jones (ed.), CIAT CD-ROM Series, Centro Internacional de Agricultura Tropical, Cali, Colombia.
- Peterson, A.T. and Cohoon, K.P. (1999) Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling*, **117**, 159–164.
- Peterson, A.T., Soberón, J., and Sánchez-Cordero, V. (1999) Conservatism of ecological niches in evolutionary time. *Science*, **285**, 1265–1267.
- Peterson, A.T., Stockwell, D.R.B., and Kluza, D.A. (2002) Distributional prediction based on ecological niche modeling of primary occurrence data, in *Predicting Species Occurrences: Issues of Scale and Accuracy*, J.M. Scott, P.J. Heglund, and M.L. Morrison (eds.), Island Press, Washington, DC, pp. 617–623.
- Pettitt, A.N., Weir, I.S., and Hart, A.G. (2002) A conditional autoregressive gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing*, **12**, 353–367.
- Rapoport, E. (1975) *Aerografía. Estrategias Geográficas de las Especies*, Fondo de Cultura Económica, Mexico.
- Sánchez-Cordero, V. and Martínez-Meyer, E. (2000) Museum specimen data predict crop damage by tropical rodents. *Proceedings of the National Academy of Science of the United States of America*, **97** (13), 7074–7077.
- Soberón, J., Golubov, J., and Sarukhán, J. (2001) The importance of opuntia in Mexico and the routes of invasion and impact of *Cactoblastis cactorum*. *Florida Entomologist*, **84**, 486–492.
- Stockwell, D.R.B. and Noble, I.R. (1991) Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. *Mathematics and Computers in Simulation*, **32**, 249–254.
- Stockwell, D.R.B. and Peters, D. (1999) The GARP modeling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, **13**, 143–158.
- Udvardy, M. (1969) *Dynamic Zoogeography*, Van Nostrand Reinhold Company, New York.

Biographical Sketch

Jorge A. Argáez received the Ph.D. degree in statistics from Centro de Investigación en Matemáticas (CIMAT), and is currently a statistician at Centro de Investigación Científica de Yucatán. His primary interest is in applied statistical models for ecology. He has been involved in predicting the distribution of endemic species in the Yucatán Peninsula.

José Andrés Christen received the Ph.D. degree in statistics from the University of Nottingham in 1994, and is currently a researcher at CIMAT. His research interest is applied Bayesian statistics. His specific interests include applications in biodiversity, clinical trials, applied decision theory, and others.

Miguel Nakamura received the Ph.D. degree in statistics from the University of North Carolina at Chapel Hill in 1989. He is currently a researcher at CIMAT, and chairs the Statistics Lab there. His research interests include applications of statistics to environmental sciences and biology, namely, the modeling of species accumulation curves and statistical inference for areas of species distributions.

Jorge Soberón is a biologist at Universidad Nacional Autónoma de México, where he also obtained his M.Sc. He received the Ph.D. degree in theoretical ecology from Imperial College in 1982. Since 1996 he has been working on use of databases for the analysis of biodiversity patterns, and has published works on the applications of niche modeling and species distributions modeling to a variety of basic and applied subjects.