PREDICTION OF HIGH POTENTIAL AREAS OF HABITAT FOR MONITORED SPECIES

Jorge Argáez-Sosa, J. Andrés Christen, Miguel Nakamura and Jorge Soberón Mainero

Comunicación Técnica No I-02-06/16-04-2002
(PE/CIMAT)

Prediction of high potential areas of habitat for monitored species

Jorge Argáez–Sosa *jargs@cimat.mx*
J. Andrés Christen *jac@cimat.mx*
Miguel Nakamura *nakamura@cimat.mx*
Centro de Investigación en Matemáticas, A. C.
Apartado Postal 402
36000 Guanajuato, Gto., Mexico

Jorge Soberón Mainero *jsoberon@xolo.conabio.gob.mx*
Departamento de Ecología Evolutiva, Instituto de Ecología, UNAM
Comisión Nacional para el Conocimiento y Uso de la Biodiversidad

Abstract

The area of distribution of a species is one of the fundamental expressions of its ecology and evolutionary history. Detailed knowledge of distribution areas is relevant to address basic questions in biogeography and ecology and in the management of biodiversity for conservation or sustainable use. The problem we tackle in this paper is inferring the zones of high potential for the habitat of the species under study, based on reported sites of presence, where each site is associated with values of covariates, measured on a discrete scales. We compute the predictive probability that the species is present at each site, by means of a mixture involving all pairs of covariates. Possible spatial bias for sites of presence is accounted for. Since the posterior distribution does not have a closed form, MCMC is implemented. However, we also describe an approximation to the posterior distribution, which avoids MCMC. Available *a priori* information regarding the areas of distribution of the species can be incorporated in a clear-cut way. In addition we propose a map of uncertainty which allows for greater insight into the nature of potential areas of distribution. By simulations, we compare our approach with other standard methods. Two case studies are also presented.

Keywords: Area of distribution, Biodiversity, Mixture model, Predictive probability map, a priori elicitation.

1 Introduction

All species of animals and plants occupy a more or less well-defined geographical region, called their areas, or ranges, of distribution. The probability of presence of the species, as a

function of geographical location, would give rise to a map that highlights areas within the region where the species is most likely to be present. The area of distribution of a species is one of the most fundamental expressions of its ecology and evolutionary history (Udvardy 1969; Brown, Stevens and Kaufman 1996; Gaston and Blackburn 2000). 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 likely routes of economically damaging invasive species, 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 (very seldom formal, see Jennrich and Turner 1969; Rapoport 1975) 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 (Udvardy 1969). 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. Although very well-studied localities supply, by complement, also with "absence" information, generally speaking basic data is a set of coordinates providing localities where a given species has been observed. 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 we tackle in this paper is inferring the zones of high potential for the habitat of the species, based on reported sites of presence. We propose 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. In what follows we refer to this as *spatial bias*. Spatial bias deals with 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. Points in the covariate space may also be non-uniformly distributed, so in addition, a notion of *covariate bias* may also be present. Clearly, covariate bias depends on the nature of spatial bias and on the distribution of covariates over the whole region of interest.

Assessment of potential zones of presence are based on values of the covariates This means that even if samples were biased spatially, 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 site physically examined, over the period of observation considered.

In addition, there is the notion of *detectability* of a species. Even if a site with a high probability of presence is physically examined for the presence of the species at a given point in time, 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 in fact have been extensively used for constructing maps of distribution areas. However, few of these methods 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; Stockwell and Peters 1999; Peterson and Cohoon 1999; Peterson and Stockwell in press). FloraMap and GARP (run via internet at http://biodi.sdsc.edu) will be considered for comparisons in this paper. 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. Thus the experts often reduce by hand the surfaces predicted by the mathematical methods without resorting to explicit methods 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

The geographical region of interest is assumed to be covered by a regular, square, grid. Let s be a generic node on the grid. The probability of potential inhabitation at s usually represents potential over a square centered on s taken to be the same size as a square on the grid. Thus, in practice, the region is a set $\mathcal{R}$ of nodes specified by the grid. For each node $s \in \mathcal{R}$, an M-dimensional vector $\mathbf{e}(s) = (e_1(s), \ldots, e_M(s))$ of covariates is assumed to be known. Here M is the number of physical/climatic covariates, and it is assumed that all of them are either categorical or measured on discrete scales. Thus, we assume $e_k(s) \in \{1, ..., R_k\}$, $1 \leq k \leq M$,

where R_k is the number of possible classes for the k-th covariate. The set of all conceivable covariate configurations is $F = \{1, \ldots, R_1\} \times \ldots \times \{1, \ldots, R_M\}$ (although many of them may not actually occur over $\mathcal{R}$), thus $\#(F) = \prod_{k=1}^{M} R_k$.

Observed data consists of n nodes, s_1, ..., s_n, corresponding to n exact geographical locations of positive observation that have been identified with a nearest centerpoint s. Some of these nodes may be multiple, since two or more observations may have occurred at different locations sharing the same center s. For any $f \in F$, we denote by $C(f)$ the number of nodes in the sample such that $\mathbf{e}(s_i) = f$, $1 \leq i \leq n$. We use the notation $\mathbf{C} = (C(f))_{f \in F}$, to represent the vector of all counts, arranged according to F's lexicographic order. Notice that $\sum_{f \in F} C(f) = n$, and that many of the elements of $\mathbf{C}$ may actually be zero.

For a given node s, let u_s be a binary random variable which takes on the value 1 if the species is present at the site, and the value 0 otherwise. The probability $P(u_s = 1)$, as a function of s, constitutes the map of probabilities of presence for the species over $\mathcal{R}$. A fundamental notion is that presence is determined by covariates, rather than geographical location. Let $\mathbf{U} = (U_1, \ldots, U_M)$ be the vector of covariate values tacitly selected by the species when it makes itself present. We interpret $\mathbf{U}$ to be a random vector. The fundamental assumption that enables inference of areas of high potential from reported sites of presence via the consideration of covariates is that

$$P(u_s = 1) = P\{\mathbf{U} = \mathbf{e}(s)\}. \tag{1}$$

To incorporate sampling bias, let $\delta(s)$ denote the probability that node s is examined for presence within the timeframe of study. This is spatial bias, and induces "covariate bias", which we now denote by $v(f)$. This last quantity is the probability that a node having value f for the covariate vector is physically examined for presence. The relationship $v(f) = 1 - \prod_{s:\mathbf{e}(s)=f}(1 - \delta(s))$ is assumed, essentially meaning that nodes with constant covariates are independently visited. In a strict sense this may not hold, but independence does not seem too stringent. One does not intentionally plan to consider nodes having the same covariate vector as candidates for additional examination. In addition, the distance between nodes is usually large (*e.g.* 10–15 km), and therefore having visited a node does not necessarily increase the chances of visiting a neighbor.

As we have noted, detectability is an inherent property of the species. In general, however, detectability may also depend on $\mathbf{e}(s)$, but statement (1) is also saying that the species tend to be present at nodes such that $\mathbf{e}(s)$ resembles probable values for f. Hence, since the species tends to be present at nodes of similar covariates, it is sensible to assume approximately that detectability does not depend on s at all sites where the species is present. Accordingly, let d denote detectability for a node, that is, d is the (constant) probability of detecting a species given that it is present at a node. Considerations may be easily made to allow for non-constant detections, but we will not address them here.

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

$$P(o_s = 1) = P(u_s = 1)\, v(\mathbf{e}(s))\, d. \tag{2}$$

The development of the statistical model below reflects the fact that the only observable quantity in (2) is o_s, when $o_s = 1$. Therefore, the probability of presence is not identifiable

without first discerning $v(\mathbf{e}(s))$ and d. Our method will assume that $v(\mathbf{e}(s))$ is given exactly either by assuming uniform sampling over nodes, or by a given input generated by the user via specification of spatial bias, $\delta(s)$. Notice that what is indeed random is $\mathbf{U}$, the value of the covariates at a recorded site of presence, rather than o_s itself, which is fixed at the value 1 as a consequence of design.

2.2 Formulation

For $f \in F$, let $\theta(f)$ denote the probability that the species is present at a node s such that $\mathbf{e}(s) = f$, and let $\boldsymbol{\theta} = (\theta(f))_{f\in F}$. From (1), we note that $\theta(\mathbf{e}(s))$ is the probability of $\mathbf{U} = \mathbf{e}(s)$ given $\boldsymbol{\theta}$, so that $\boldsymbol{\theta}$ is actually the parameter of interest. Incorporating this parameterization, and using (1) and (2) we obtain

$$P(o_s = 1 \mid \boldsymbol{\theta}) = P(\mathbf{U} = \mathbf{e}(s) \mid \boldsymbol{\theta})\, v(\mathbf{e}(s))\, d. \tag{3}$$

Let N be the total number of nodes examined in the timeframe considered that gave rise to the n nodes of presence, and (temporarily) assume N is known. If the N sampled nodes can be considered independent (if $\boldsymbol{\theta}$ is assumed as a random variable a weaker assumption of exchangeability may be used; see Bernardo and Smith 1994, pp. 167–171) each sampled node can be viewed as having been randomly grouped into one of $\#(F) + 1$ bins. The first $\#(F)$ bins have the possible values of $f = \mathbf{e}(s)$ 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 (3), the probability of a node being classified into bin labeled f is $\theta(f)\, v(f)\, d$. This constitutes a standard multinomial setting, so that if $\mathbf{c} = (c(f))_{f\in F}$ is a vector such that $\sum_{f\in F} c(f) = n \le N$, then

$$P(\mathbf{C} = \mathbf{c} \mid \boldsymbol{\theta}) = \kappa \left\{1 - \sum_{f\in F} \theta(f)\, v(f)\, d\right\}^{N-n} \prod_{f\in F} \{\theta(f)\, v(f)\, d\}^{c(f)}, \tag{4}$$

where κ is the normalizing constant $N!\,\{\prod_{f\in F} c(f)!\}^{-1}[\{N - \sum_{f\in F} c(f)\}!]^{-1}$.

Because n is usually small, most of the observed counts result in zero. This causes the parameter $\boldsymbol{\theta}$ to be very inconvenient, in that it possess an estimation problem with sparse data. Meaningful reduction in parameter dimensionality is considered next.

Let G be the set of all index pairs (a, b), $1 \le a < b \le M$. To shorten notation, let $J = (a, b)$ denote a generic pair in G. Let $\mathbf{U}_J = (U_a, U_b)$, $\mathbf{e}_J(s) = (e_a(s),\, e_b(s))$ and $F_J = \{1, \ldots, R_a\} \times \{1, \ldots, R_b\}$. For $g \in F_J$, we denote by $C_J(g)$ the number of nodes in the sample such that $\mathbf{e}_J(s_i) = g$, and we let $\theta_J(g) = P(\mathbf{U}_J = g \mid \boldsymbol{\theta}_J)$, $\boldsymbol{\theta}_J = (\theta_J(g))_{g\in F_J}$, $\mathbf{C}_J = (C_J(g))_{g\in F_J}$, and $v_J(g) = 1 - \prod_{s:\mathbf{e}_J(s)=g}(1 - \delta(s))$. The object of introducing the J notation is to point out that if $M > 2$, there is a corresponding multinomial distribution (4) for each pair J: If $\mathbf{c}_J = (c_J(g))_{g\in F_J}$ is a vector such that $\sum_{g\in F_J} c_J(g) = n \le N$, then

$$P(\mathbf{C}_J = \mathbf{c}_J \mid \boldsymbol{\theta}_J) = \kappa_J \left\{1 - \sum_{g\in F_J} \theta_J(g)\, v_J(g)\, d\right\}^{N-n} \prod_{g\in F_J} \{\theta_J(g)\, v_J(g)\, d\}^{c_J(g)}, \tag{5}$$

with a corresponding expression for κ_J. Let $\boldsymbol{\theta}' = (\boldsymbol{\theta}_J)_{J\in G}$. Now let us assume that J is actually random, having a distribution $\pi(J)$, and that conditioned on the value of J, the probability of presence as a function of covariates in pair J is $\boldsymbol{\theta}_J$. The unconditional probability of presence at node s is then the mixture

$$\theta'(f) = P(u_s = 1 \mid \boldsymbol{\theta}') = \sum_{J\in G} \theta_J(\mathbf{e}_J(s))\,\pi(J). \tag{6}$$

There is a multinomial model for $\mathbf{C}_J$ for each value of J, so that observed data is regarded to be $\mathbf{C}' = (\mathbf{C}_J)_{J\in G}$ instead of $\mathbf{C}$.

Notice that by setting θ in (4) to be of the form (6), the model for $P(\mathbf{C} = \mathbf{c} \mid \boldsymbol{\theta}')$ is

$$\kappa \left\{1 - \sum_{f\in F} \theta'(f)\, v(f)\, d\right\}^{N-n} \prod_{f\in F} \{\theta'(f)\, v(f)\, d\}^{c(f)}. \tag{7}$$

Model (7) amounts to restricting the multinomial bin probabilities in (4) to be of a given form, (6), via a parameter $\boldsymbol{\theta}'$ (greatly reduced in dimension) and probabilities $\pi(J)$. The reduction is based on a mixture of all pairwise interactions. One interpretation of this restriction is probabilistic: A species is thought of as selecting a pair, J, at random from G, with probability $\pi(J)$, and then the probability of presence at any site s is determined by $\theta_J(\mathbf{e}_J(s))$. 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 only allows for resolution up to pairs of covariates, but is compatible with a principle stating that species focus on a small set of attributes and simple criteria in order to decide a site for colonization. Although sensible, this principle will require experimental testing and our model could provide a contrasting hypothesis for such testing. Currently it is known that for the GARP algorithm, more than about five variables do not add much predictive power (Peterson and Cohoon 1999). A map depicting the probabilities (6) for each node is the true map of potential for the species under study. In passing, note that a precise probabilistic definition for the concept of "potential" at each node s has been established. This contrasts with the rather lax use of the word "potential" in other approaches.

Regarding N, it is very unlikely that a full record of visited sites is kept, especially considering historical data, and thus N must be considered to be unknown. However, we expect $C(f) \approx N\,\theta(f)\,v(f)\,d$ (for large N) and since $\sum_{f\in F}\theta(f) = 1$, we must have $N \approx N^* = \left\lceil \sum_{f\in F} C(f)/(v(f)\,d) \right\rceil$. A simple way to proceed, as we do in the following sections, is to postulate $N = N^*$ as a working approximation in (7), rather than considering N itself to be an unknown, nuisance, parameter.

3 Inference

3.1 Predictive probability

We calculate the predictive probability of presence of the species at each node s, $P(u_s = 1 \mid \mathbf{C}')$. For each pair of covariates, a prior distribution is postulated for the parameter $\boldsymbol{\theta}_J$,

denoted by $f(\boldsymbol{\theta}_J)$. 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 themselves as random is irrelevant because we are assuming an arbitrary distribution for J). The elicitation of $f(\boldsymbol{\theta}_J)$ and $\pi(J)$ is discussed in Section 3.2. We also introduce the notation $f(\mathbf{C}_J \mid \boldsymbol{\theta}_J,\, J)$ for the (multinomial) model (5) and $f(\boldsymbol{\theta}_J \mid \mathbf{C}_J,\, J)$ for the posterior distribution given J. Notation $\pi(J \mid \mathbf{C}')$ is used for the posterior probability for pair J.

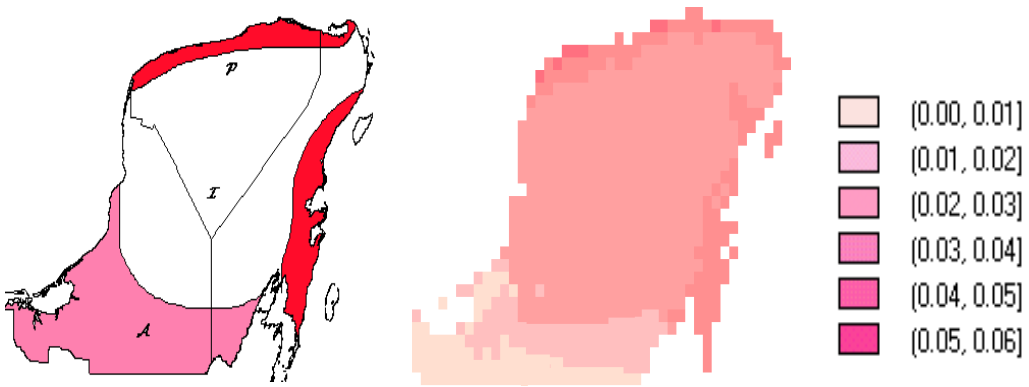
The law of total probability yields $P(u_s = 1 \mid \mathbf{C}') = \sum_{J \in G} P(u_s = 1 \mid J,\, \mathbf{C}')\, \pi(J \mid \mathbf{C}')$. The quantity $P(u_s = 1 \mid J, \mathbf{C}')$ is the predictive probability of presence given J, and is calculated by $\int P(u_s = 1 \mid J, \boldsymbol{\theta}_J)\,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

$$P(u_s = 1 \mid \mathbf{C}') = \sum_{J \in G} E[\theta_J(\mathbf{e}_J(s)) \mid J, \mathbf{C}')]\, \pi(J \mid \mathbf{C}'). \tag{8}$$

For each pair J we postulate a Dirichlet distribution as prior for $\boldsymbol{\theta}_J$, whose expression is $f(\boldsymbol{\theta}_J) = \Gamma(\alpha_J) \left[\prod_{g \in F_J} \Gamma(\alpha_J(g))\right]^{-1} \prod_{g \in F_J} \theta_J(g)^{\alpha_J(g)-1}$, where $\alpha_J = \sum_{g \in F_J} \alpha_J(g)$, $\alpha_J(g) > 0$. The parameter for this distribution is $\boldsymbol{\alpha}_J = (\alpha_J(g))_{g \in F_J}$. The Dirichlet distribution is commonly used to model vectors of probabilities. In case of a multinomial model, under certain general conditions, every prior distribution for parameter $\boldsymbol{\theta}_J$ can be approximated by a mixture of Dirichlet distributions (Walley 1996). However, there is no standard closed form for the posterior distribution resulting from the multinomial model (5) and *a priori* Dirichlet given the expressions for the bin probabilities (see expression 10 in Appendix). Therefore, one needs to resort to numerical methods (MCMC, as indicated in the Appendix) to simulate values from the posterior of $\boldsymbol{\theta}_J$ to obtain the quantities $E[\theta_J(\mathbf{e}_J(s)) \mid J, \mathbf{C}')]$ and $\pi(J \mid \mathbf{C}')$ involved in (8). The quantity $\pi(J \mid \mathbf{C}')$ can be interpreted as the posterior probability that 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 $\mathbf{X}^*_J + \boldsymbol{\alpha}_J$, where $\mathbf{X}^*_J = (X^*_J(g))_{g \in F}$ with $X^*_J(g) = C_J(g)\, (v_J(g)\, d)^{-1}$, as an approximation to the exact posterior distribution. Inspired by the observation that $X^*_J(g)$ represents an approximation to the actual multinomial count related to the cell probability $\theta_J(g)$, we would obtain the mentioned Dirichlet as a "posterior" (note that the formal consideration of an alternative model of this type for the $C_J(g)$'s would entail the identification of an unknown normalization dependant on $\boldsymbol{\theta}_J$). The required expected value in (8) is given by $E[\theta_J(\mathbf{e}_J(s)) \mid J, \mathbf{X}^*_J] = [X^*_J(\mathbf{e}_J(s)) + \alpha_J(\mathbf{e}_J(s))][N + \alpha_J]^{-1}$. The closed-form calculation of $\pi(J \mid \mathbf{C}')$ is shown in the Appendix. As far as the approximation is concerned, what is relevant is that, by examining the distributions $f(\boldsymbol{\theta}_J \mid \mathbf{C}_J)$ and $f(\boldsymbol{\theta}_J \mid \mathbf{X}^*_J)$, we observe (numerically) that the corresponding expected values, $E[\theta_J(\mathbf{e}_J) \mid J, \mathbf{C}_J]$ and $E[\theta_J(\mathbf{e}_J) \mid J, \mathbf{X}^*_J]$ are virtually equal. The approximation device produces slightly smaller marginal posterior variances. Certainly, the mathematical tractability of $f(\boldsymbol{\theta}_J \mid \mathbf{X}^*_J)$ (a Dirichlet) is more appealing. We compare both approaches in Section 4.

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 color scale to plot the predictive probability $P(u_s = 1 \mid \mathbf{C}')$ at each node. That is, we plot node s with the color associated with interval I_{j_s}, where $P(u_s = 1 \mid$



(a) (b)

Figure 1: (a) *a priori* regions provided by the user for species *Coccothrinax readii*. $\mathcal{P}$ represents a region where the user is quite certain that the species is present, $\mathcal{A}$ where the user is quite certain that the species is not present, and $\mathcal{I}$ the complement of these two. (b) Resulting *a priori* potential.

$\mathbf{C}') \in I_{j_s}$. We also evaluate $I^J(s) = \int_{I_{j_s}} f(\theta_J(\mathbf{e}_J(s)) \mid \mathbf{C}')\, d\theta_J(\mathbf{e}_J(s))$. The quantity $I^J(s)$ is the posterior probability that $\theta_J(\mathbf{e}_J(s))$ lies in I_{j_s}, and motivated by (8), the quantity $I(s) = \sum_{J \in G} I^J(s)\, \pi(J \mid \mathbf{C}')$ provides a level of certainty about the potential (predictive probabilities) 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 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, see Figures 2 and 3).

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 Högmander 1994, Högmander and Möller 1995, Diggle, Tawn and Moyeed 1998 or De Oliveira 2000). In our experience, the usage of the map of certainty (or uncertainty) helps in the interpretation and understanding of the posterior distribution at hand and leads to more educated conclusions.

3.2 Prior elicitation

In this Section, for fixed $J \in G$, the parameters of the *a priori* distribution, $(\alpha_J(g))_{g \in F_J}$ and $\pi(J)$, are elicited. For the Dirichlet distribution it is a fact that $\alpha_J(g) = \alpha_J E\left[\theta_J(g)\right]$, where $E\left[\theta_J(g)\right]$ is the prior expected value for $\theta_J(g)$. That is, the values α_J and $E\left[\theta_J(g)\right]$ should be elicited. It will be unusual that the expert provides directly values 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 Figure 1(a)). Either $\mathcal{P}$, $\mathcal{A}$ or both may be empty.

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 involved three nodes are excluded. The examination is repeated, each time with the remaining nodes, until 3-way contradictions are exhausted. Let $\mathcal{R}_2 \subset \mathcal{R}$ 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. For further details of this elicitation process, see Argáez-Sosa, Christen and Nakamura (In Prep.).

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, Carlin, Stern and Rubin 1995, p. 76). Since the relevant information for establishment of the species depends on values of the covariates, we are thus motivated to define $\alpha_J = \#(\mathcal{R}_1)\left[\#(\mathcal{R} \setminus \mathcal{R}_1)\right]^{-1}$, which takes on values in the range $(0, \infty)$. In the absence of prior information (that is, $\mathcal{I} = \mathcal{R}$), one would set $\alpha_J(g) = 1 \diagup R_a R_b$, a well accepted non-informative prior.

Regarding elicitation of $E\left[\theta_J(g)\right]$, 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 .95 for $\mathcal{P}$, .05 for $\mathcal{A}$ and .5 for I (denoting ambiguity), we define

$$w_J(g) = \frac{(.95)\#\{s \in \mathcal{P} : \mathbf{e}_J(s) = g\} + (.5)\#\{s \in \mathcal{I} : \mathbf{e}_J(s) = g\} + (.05)\#\{s \in \mathcal{A} : \mathbf{e}_J(s) = g\}}{\#\left\{s \in \mathcal{R} : e_J(s) = g\right\}}.$$

Using these values we normalize and establish $E\left[\theta_J(g)\right] = w_J(g)\left[\sum_{g' \in F_J} w_J(g')\right]^{-1}$. Finally, we elicit $\pi(J)$. 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, namely $\pi(J) = \alpha_J \diagup \sum_{J' \in G} \alpha_{J'}$.

Heuristic verification that elicitation is made sensibly is to calculate the *a priori* maps of potential, by means of $P(u_s = 1) = \sum_{J \in G} E\left[\theta_J(\mathbf{e}_J(s))\right] \pi(J)$. By inspection, we verify that contours of $P(u_s = 1)$ roughly coincide with the areas $\mathcal{P}$, $\mathcal{A}$ and $\mathcal{I}$ established by the user (See Figures 1(a) and (b)).

4 Simulation Study

A simulation study is considered to examine peculiarities of our methodology and alternative methods such as Domain and FloraMap. The physical region and corresponding covariates will be quite real—the Yucatan Peninsula in Mexico—but the actual sites of presence of a fictitious species will be simulated. A regular grid of 761 nodes covers this region, separated approximately by 12 km. 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 its ideal value, we set

$$P(u_s = 1) = e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{e}(s))^t \mathbf{A}(\boldsymbol{\mu}-\mathbf{e}(s))} \tag{9}$$

in the simulations. Note that model (9) 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 (8) gives satisfactory results and we do not document those examples here. Instead, we here 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 varying degrees of sensitivity to an ideal climate. In the simulation study, the function (9) may thus 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 for $\nu(f)$ given in Section 2. We reproduce this fact by considering main highways on the Peninsula.

Data for simulations were generated by superimposing (9) and the spatial clustering induced by highways. A species is present at a node s according to probabilities (9), the site s is visited by human observers with a probability inversely proportional to the distance from s to the nearest road, and an observation of the species is recorded with probability d (d is fixed at 1 in what follows). 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 alternative methods FloraMap and Domain. In addition, we produce the uncertainty map as explained in Section 3.1. Maps of potential using Bioclim and GARP were also obtained, but are not presented here because these methods over-estimate and output practically all of the Yucatan Peninsula as high potential in all cases.

We only display two representative examples (Figures 2 and 3). The first example represents a species with high sensitivity ($a_{11} = 1$, $a_{12} = .9$, $a_{13} = .85$, $a_{22} = 1$, $a_{23} = .9$, $a_{33} = 1$) and the second example a species with low sensitivity ($a_{11} = 1$, $a_{12} = .6$, $a_{13} = .3$, $a_{22} = 1$, $a_{23} = .1$, $a_{33} = 1$). The idealized potential may be found in Figures 2(a) and 3(a). In both cases, the scenarios are difficult, in that there is spatial bias, non–informative prior information, and a small sample size.

In Figures 2(b) and 3(b) the estimated potential map for each scenario is depicted. In both figures the presence of record sites located far away from the real high potential area are noted. Our method does not produce a high potential area around those sites, unlike FloraMap (Figures 2(e) and 3(e)) and Domain (Figures 2(f) and 3(f)). Moreover, maps depicted in Figures 2(c)–(d) and 3(c)–(d), 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 .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–Sosa *et al.* (In Prep.). Our methodology appears to behave correctly in all reasonable situations (*e.g.* 1 holds, non extreme spatial bias and priors not grossly in error), 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 Domain and FloraMap into determining high potential for a significant area around these points. The method also appears to be robust to the spatial bias introduced by roads and towns, because the region of high potential is recovered reasonably well despite the clustering of points of presence. Both Domain and FloraMap tend to over-estimate due to spatial bias. As expected, when the sample size 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 found in Figure 2(c)–(d) and 3(c)–(d) do not appear to have substantial differences that would lead to qualitatively different interpretations. This suggests that the Dirichlet approximation is useful.

5 Case Studies

5.1 *Coccothrinax readii*

The region of interest, $\mathcal{R}$, is the Yucatan Peninsula in Mexico. The species under study, *Coccothrinax readii,* is an endemic plant belonging to the *palmacea* family, regarded as an endangered species. This species has been reported in 67 localities. The regular grid is as described in the simulations, and the matrix containing the values of covariates for each node of the grid was obtained from researchers in botany at 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 produces ten pairs of covariates.

The *a priori* zones $\mathcal{P}$ and $\mathcal{A}$, as produced by the researchers are shown in Figure 1(a). The resulting maps using our method, the uncertainty map and Domain and FloraMap outputs, are shown in Figure 4(b), (c) and (d), respectively.

We also computed the quantities $\pi(J \mid \mathbf{C}')$ for each J. In this application pair J defined by temperature-soil type produces $\pi(J \mid \mathbf{C}') = .9889$, and for pair J' defined by humidity-temperature $\pi(J' \mid \mathbf{C}') = .0111$. Other pairs produce a posterior probability less than .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. Recent considerations suggest that this species is, at present, expanding its area of distribution. The zones highlight by our method coincides 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 Figure 5(a) (producing a low predictive probability), with the uncertainty map in Figure 5(b) (producing a low level of uncertainty

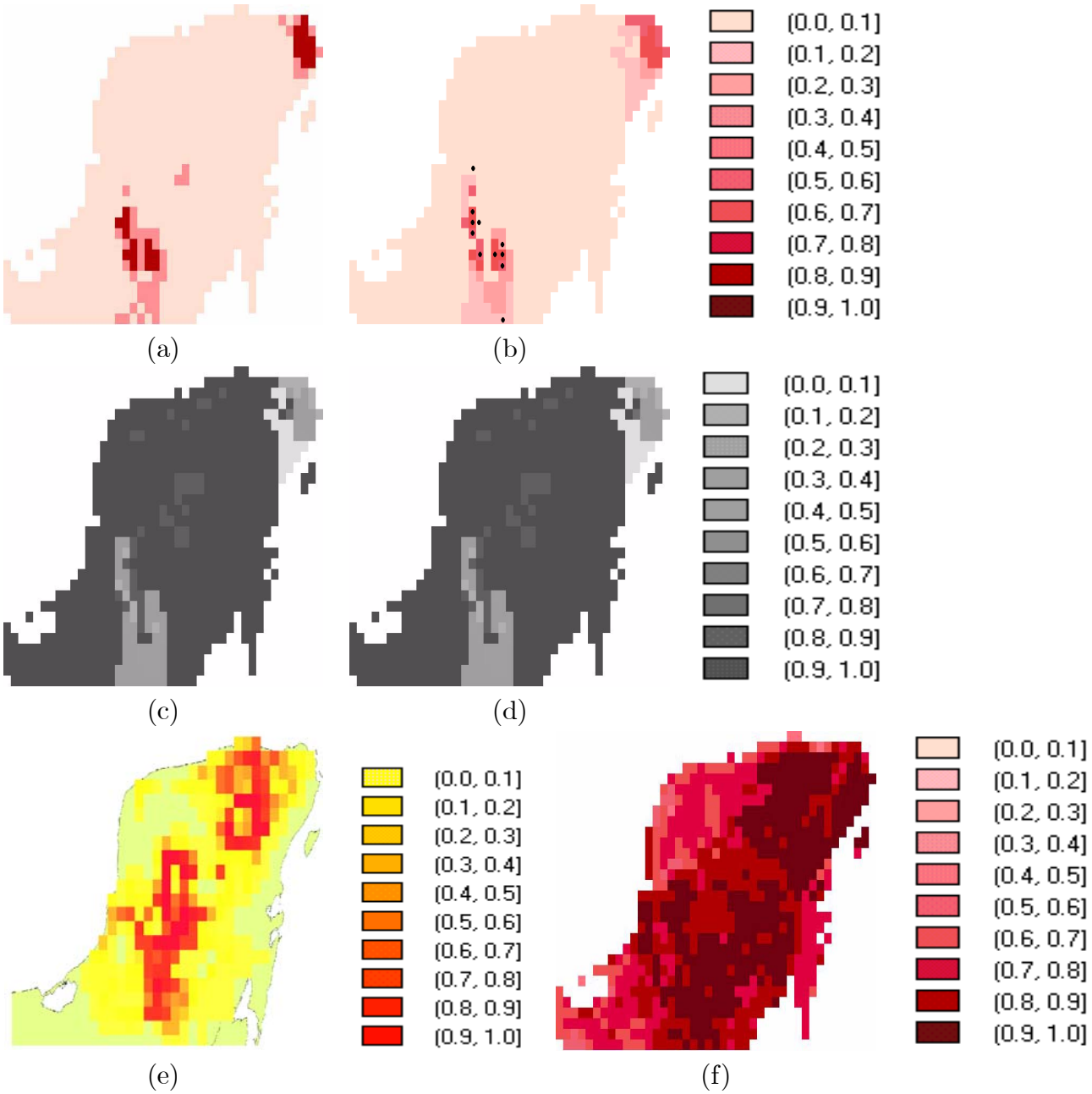


Figure 2: (a) Idealized potential produced by the model in (9). (b) Simulated points of presence ($n = 15$) and estimated potential using our method. (c) Map of uncertainty for the estimated potential using the Dirichlet approximation. (d) Map of uncertainty for the estimated potential for the exact posterior using MCMC. (e) Estimated potential using FloraMap. (f) Estimated potential using Domain.

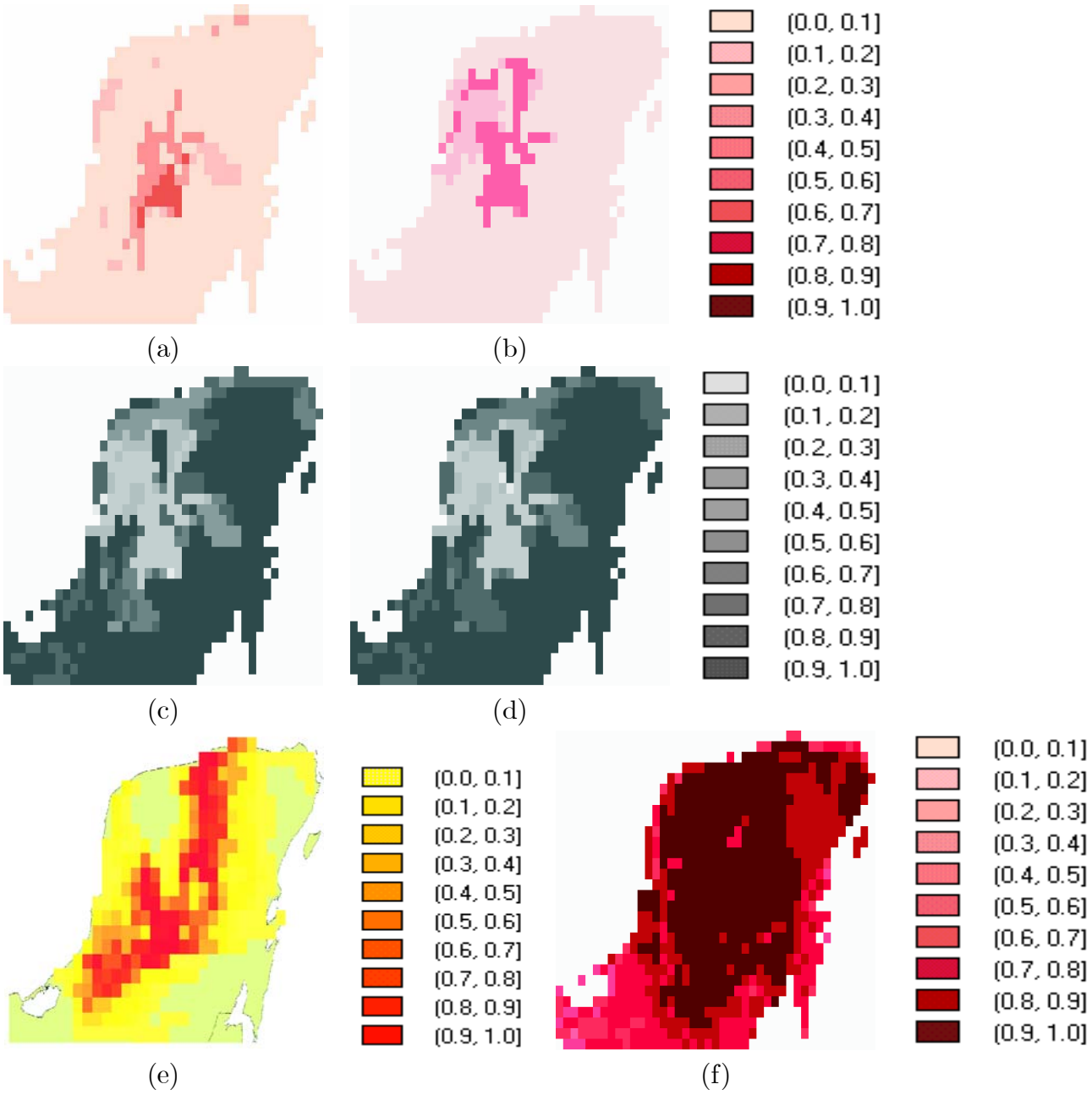


Figure 3: (a) Idealized potential produced by the model in (9). (b) Simulated points of presence ($n = 15$) and estimated potential using our method. (c) Map of uncertainty for the estimated potential using the Dirichlet approximation. (d) Map of uncertainty for the estimated potential for the exact posterior using MCMC. (e) Estimated potential using FloraMap. (f) Estimated potential using Domain.

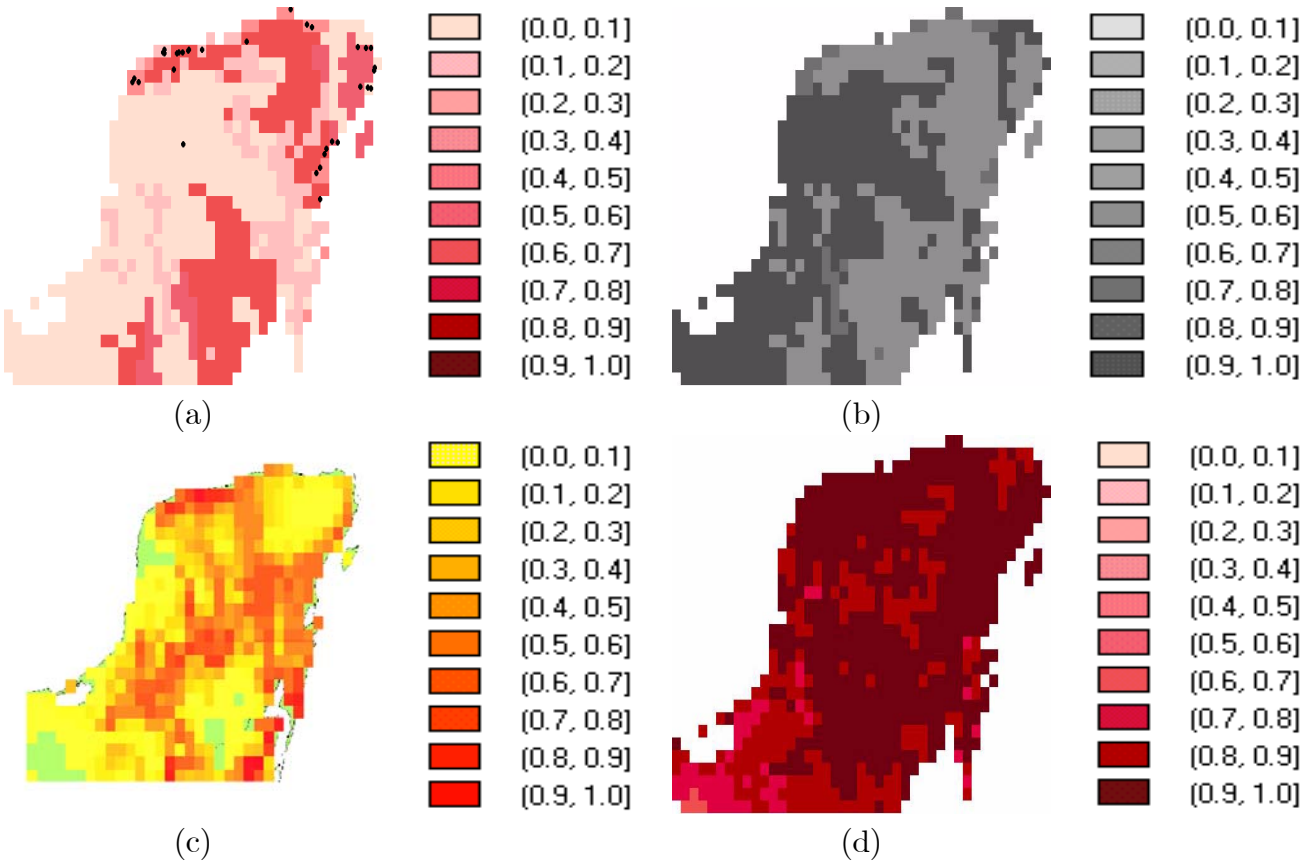


Figure 4: (a) Reported sites of presence for the species *Coccothrinax readii* ($n = 67$) and estimated potential using our method. (b) Map of uncertainty for the estimated potential. (c) Estimated potential using FloraMap. (d) Estimated potential using Domain.

around this site), 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 consist 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 to consider 45 pairs. The map of *a priori* information is shown in Figure 5.

Figure 6(a) is the potential map, with the sites of presence, and Figure 6(b) is the map of uncertainty. The corresponding maps obtained with FloraMap and Domain are Figures 6(c) and 6(d). In this case, the most influential pair is humidity-elevation, with posterior probability .999.

Figure 5: *a priori* region $\mathcal{P}$ provided by the user for the species *Baronia brevicornis.*

Based on the field experience of one of us (JSM), Domain 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 produced 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 formally defines "potential", has a formal background in statistical inference to support it, has a version simple to implement and allows for inclusion of prior information in a convenient way. It might be argued that the 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, as suggested in Section 4. Certainly, the techniques used in this paper may be easily generalized for higher (3, 4–way, *etc.*) interactions, but we are not sure that the additional complexity would reflect in better results.

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 could not be used by previous methodologies. In second place, the preliminary examples we have analyzed suggest that our method suffers less from overprediction that some existing alternatives, like FloraMap or Domain. Field checking the predictions of distribution algorithms is expensive and time-consuming. 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

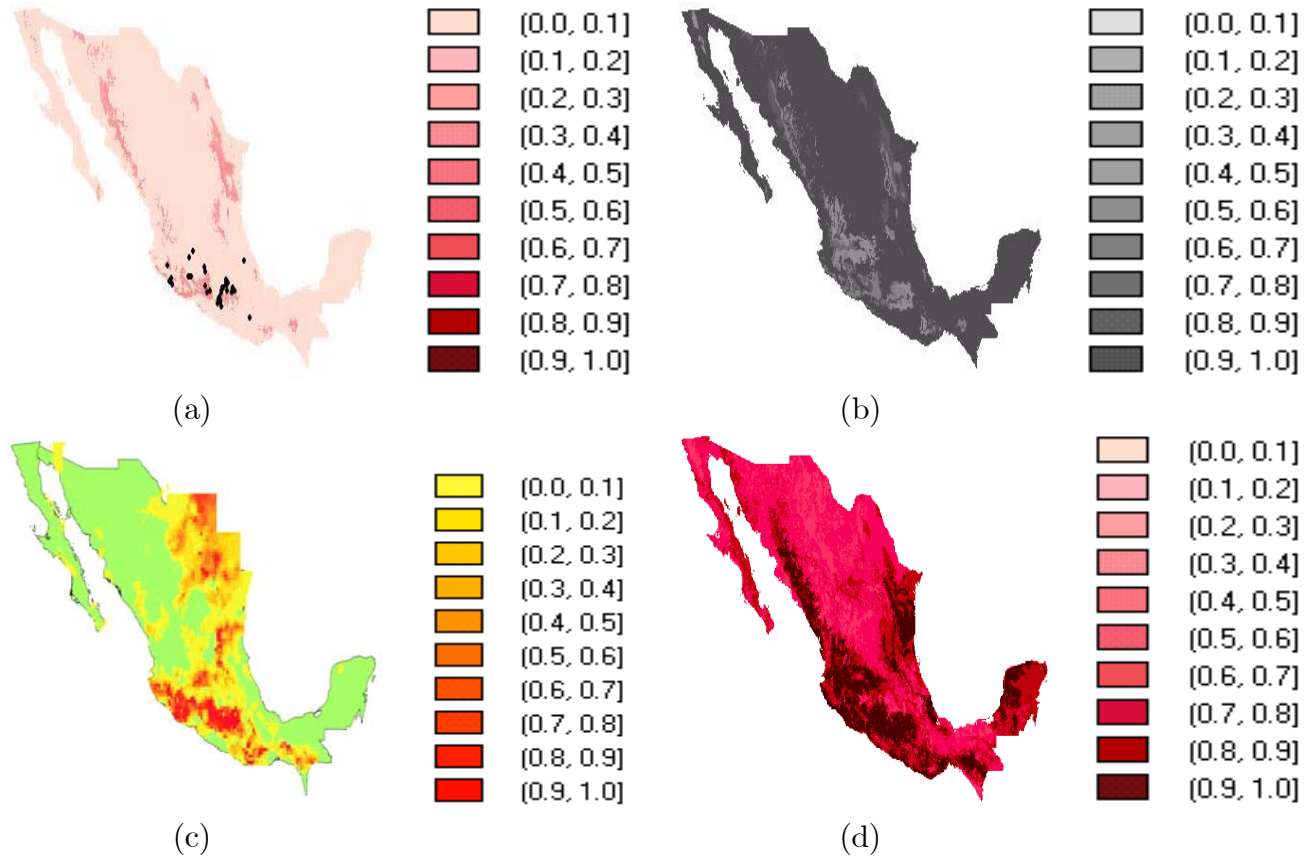


Figure 6: (a) Reported sites of presence for the species *Baronia brevicornis* ($n = 40$) and estimated potential using our method. (b) Map of uncertainty for the estimated potential. (c) Estimated potential using FloraMap. (d) Estimated potencial using Domain.

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 to predict biological species distributions.

7 Acknowledgments

J. Argáez-Sosa was supported by CONACYT Grant 115344. Dr. Nakamura's work was partially supported by CONACYT Grant 32156–E. This work was completed while J. A. Christen was visiting CIMAT. We would like to thank the Centro de Investigación Cientifica de Yucatán, for the data and input for prior elicitation provided for this paper.

A Appendix

A.1 Approximating $\pi(J \mid \mathbf{C}')$

It is easy to see that

$$\pi(J \mid \mathbf{C}') \propto \pi(J) \int f(\mathbf{C}_J \mid J, \boldsymbol{\theta}_J)\, f(\boldsymbol{\theta}_J)\, d\boldsymbol{\theta}_J.$$

Having an approximation $f(\boldsymbol{\theta}_J \mid \mathbf{X}^*_J)$ for $f(\boldsymbol{\theta}_J \mid \mathbf{C}')$, by Bayes theorem we see that

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

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

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

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

A.2 MCMC

A Metropolis-Hasting (See Robert and Casella 1999) algorithm is implemented. Model $f(\mathbf{C}_J \mid \boldsymbol{\theta}_J)$ with a Dirichlet prior produces the joint posterior distribution

$$\begin{aligned}f(\boldsymbol{\theta}_J, J \mid \mathbf{C}') = \frac{\pi(J) N! \Gamma(\alpha_J)}{(N-n)! \prod_{g \in F_J} c_J(g)! \Gamma(\alpha_J(g))} \left\{1 - \sum_{g \in F_J} \theta_J(g) \nu_J(g)\right\}^{N-n} \times \\ \prod_{g \in F_J} \theta_J(g)^{c_J(g)+\alpha_J(g)-1} \nu_J(g)^{c_J(g)}.\end{aligned} \tag{10}$$

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

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

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 $\mathbf{X}^*_J + \boldsymbol{\alpha}_J$ (that is, the approximation used for the posterior; 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

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

That is, the transition kernel considered is $K(\eta, \eta') = pK_1(\eta, \eta') + (1-p)K_2(\eta, \eta')$, $\eta = (J, \boldsymbol{\theta}_J)$, $p \in (0, 1)$. We arbitrarily chose the value $p = .5$.

References

[1] Argáez-Sosa, J., Christen, J. A., and Nakamura, M. (In Prep.), "Quantifying Information of *a priori* Maps and Simulation Study", Centro de Investigación en Matemáticas A. C., Guanajuato, Mexico, http://www.cimat.mx/~jac/papers/elicit.pdf

[2] Bernardo, J. O., and Smith, A. M. F. (1994), *Bayesian Theory,* New York: Wiley.

[3] 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.

[4] Busby, J. R. (1991), "BIOCLIM - A Bioclimate Analysis and Prediction System", in *Nature Conservation: Cost Effective Biological Surveys and Data Analysis,* eds. Margules, C. R., and Austin, M. P. pp. 64-68 (CSIRO Australia).

[5] 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.

[6] De Oliveira, V. (2000), "Bayesian Prediction of Clipped Gaussian Random Field", *Computational Statistics and Data Analysis,* 34, 299–314.

[7] Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based Geostatistics" (with discussion), *Applied Statistics,* 47, 299–326.

[8] Gaston, K., and Blackburn T. (2000), *Pattern and Process in Macroecology*, Blackwell Science, Oxford.

[9] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London: Chapman & Hall.

[10] Heikkinen, J., and Högmander H. (1994), "Fully Bayesian Approach to Image Restoration With an Application in Biogeography", *Applied Statistics,* 43, 569–582.

[11] Högmander, H., and Möller, J. (1995), "Estimating Distribution Maps From Atlas Data Using Methods of Statistical Image Analysis", *Biometrics,* 51, 393–404.

[12] Jennrich, R. I., and Turner F. B. (1969), "Measurement of a Non-circular Home Ranges", *Journal of Theoretical Biology,* 22, 227–237.

[13] 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, Edited by Annie L. Jones. CIAT CD-ROM Series. Cali, Colombia: Centro Internacional de Agricultura Tropical.

[14] Peterson, A. T., and Cohoon, K. P. (1999), "Sensitivity of Distributional Prediction Algorithms to Geographic Data Completeness", *Ecological modelling,* 117, 159–164.

[15] Peterson, A. T., Soberon, J., and Sanchez-Cordero, V. (1999), "Conservatism of Ecological Niches in Evolutionary Time", *Science*, 285, 1265-1267.

[16] Peterson, A. T., and Stockwell, D. R. B. (in press), Distributional Prediction Based on Ecological Niche Modeling of Primary Occurrence Data, Predicting Species Distributions (J. M. Scott, ed.), Island Press, Washington, D. C.

[17] Rapoport, E. (1975), *Aerografía. Estrategias Geográficas de las Especies,* Fondo de Cultura Económica, México.

[18] Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods,* New York, Springer-Verlag.

[19] Sanchez-Cordero, V., and Martinez-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.

[20] Soberón, J., Golubov, J., and Sarukhan, J. (2001), "The Importance of Opuntia in Mexico and the Routes of Invasion and Impact of *Cactoblastis cactorum"*, *Florida Entomologist*, 84, 486-492.

[21] 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.

[22] 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.

[23] Udvardy, M. (1969), *Dynamic Zoogeography,* Van Nostrand Reinhold Company, New York.

[24] Walley, P. (1996), "Inferences for Multinomial Data: Learning About a Bag of Marbles" (with discussion), *Journal of the Royal Statistical Society, Ser. B,* 58, 3-34.