Modelling Species Diversity Through Species Level Hierarchical Modeling

Alan E. Gelfand, Alexandra M. Schmidt, Shanshan Wu, John A. Silander Jr., Andrew Latimer and Anthony G. Rebelo[*]

April 2003

Abstract

Understanding spatial patterns of species diversity and the distributions of individual species is a consuming problem in biogeography and conservation. The Cape Floristic Region (CFR) of South Africa is a global hotspot of diversity and endemism, and the Protea Atlas Project, with some 60,000 site records across the region, provides an extraordinarily rich data set to model biodiversity patterns. Model development is focussed spatially at the one minute grid-cell scale ($\sim 37,000$ cells total for the region). We report on results for 23 species of a flower plant family known as Proteceae (of about 330 in the CFR) for a defined subregion. Using a Bayesian framework, we developed a two stage, spatially explicit, hierarchical logistic regression. Stage one models the *potential* probability of presence/absence for each species at each cell, given species attributes, grid cell (site-level) environmental data with species-level coefficients, and a spatial random effect. The second level of the hierarchy models the probability of observing each species in each cell given it is present. Because the atlas data are not evenly distributed across the landscape, grid cells contain variable number of sampling localities. Thus this model takes the sampling intensity at each site into account by assuming that the total number of times that a particular species was observed within a site follows a binomial distribution. After assigning prior distributions to all quantities in the model, samples from the posterior distribution were obtained

[*]Alan E. Gelfand is a Professor in the Institute of Statistics and Decision Sciences at Duke University. Alexandra M. Schmidt is an Associate Professor in the Department of Statistical Methods at the Federal University of Rio de Janeiro, Brazil. Shanshan Wu is a graduate student in the Department of Statistics at the University of Connecticut. John A. Silander Jr. is a Professor and Andrew Latimer is a graduate student in the Department of Ecology and Evolutionary Biology at the University of Connecticut, and Anthony G. Rebelo is Scientist at the Conservation Biology Unit, National Botanic Institute, Kirstenbosch, South Africa. This research was conducted while the second author was a Post-Doctoral Researcher in the Department of Statistics of the University of Connecticut. The work of all the authors was supported in part by NSF grant DEB0089801 and by the National Center for Ecological Analysis and Synthesis. The authors wish to thank Richard Cowling and Henri Laurie for valuable discussions.

via MCMC methods. Results are mapped as the model estimated probability of presence for each species across the domain. This provides an alternative to customary empirical ”range of occupancy” displays. Summing yields the predicted species richness over the region. Summaries of the posterior for each environmental coefficient show which variables are most important in explaining species presence. Our initial results describe biogeographical patterns over the modeled region remarkably well. In particular, species local population size and dispersal mode contribute significantly to predicting patterns, along with annual preciptation, C.V. in rainfall, and elevation.

KEY WORDS: adaptive rejection method, Markov random field, spatial logistic regression, species range, species richness.

1 Introduction

Why are there so many species in some areas and so few in others? A universal explanation for this has been the Grail of biogeographers since Darwin and other explorer-naturalists of the Nineteenth Century began cataloging global patterns in plant and animal distributions:

”if we compare this moderate number [of plant species in New Zealand or England] with the species that swarm over equal areas ... at the Cape of Good Hope, we must admit that some cause, independent of different conditions, has given rise to so great a difference in number” (Darwin, 1872).

To read the purported answers to this ancient challenge, one is left with the impression that there are many different universal explanations, each explicitly or implicitly claiming supremacy. Palmer (1996) lists 120 named hypotheses to explain patterns in biodiversity, and Rohde (1992) lists 28 that claim to explain just latitudinal patterns. This points up the difficulties encountered in developing explanatory models, and the utility of a richer, flexible modeling. The past couple of years have seen at least 3 universal (ecological) explanations of species richness patterns championed: 1) geometric constraints (Colwell and Lees, 2000), 2) scaling of constrained resource acquisition (Ritchie and Olff, 1999), and 3) species neutrality in saturated systems (Hubbell, 2001). Prior to this we have seen area proposed as the universal explanation for biodiversity patterns (e.g. Rosenzweig (1995)), as well as productivity (e.g. Currie (1991)), environmental heterogeneity (Huston, 1994), historical factors (e.g. Latham and Ricklefs (1993)), and indeed many others. The arguments marshaled are often compelling, but it is also disconcerting to see the same data used to illustrate different claims.

The advent of inexpensive high speed computation including widely available Geographic Information System (GIS) software has revised the way many ecologists think about data on species distributions. In particular, a variety of statistical and algorithmic methods have been proposed, in conjunction with GIS to enable spatial prediction of species distribution. The survey paper of Guisan and Zimmerman (2000) provides an extensive review of these developments and an enormous list of references. Here, we just note a few of the key themes (with selected references) in this work.

What we can envisage is a region which has been surveyed at a number of sites. At each site, presence (hence, implicitly absence) of a collection of species has been recorded resulting a site (rows) by species (columns) presence/absence matrix. A classification - then - modeling strategy gathers either the sites into groups containing similar species (”communities”) or the species into groups occurring at similar sites (”assemblages”). Regression modeling follows, using environmental factors for the communities or species attributes for the assemblages. See, e.g., Ferrier *et al.* (2002) and references therein. Marginalizing across rows yields richness at a site, possibly standardized by the area of the site. Marginalizing down columns produces species prevalence. Again these can be explained using regression model as in, for instance, Owen (1989) or Heikkinen (1996).

Rather than aggregating we might model directly at the species/site level. Regressions in this case and, in fact, in the above cases through the use of generalized linear and generalized additive models are receiving considerable attention in the ecology literature. See Guisan *et al.* (2002) for a review. In particular, the recently proposed Generalized Regression Analysis and Spatial Prediction (GRASP) methodology as in Lehmann *et al.* (2002) appends a spatial prediction technique onto a generalized additive model.

Our intent is also to work at the species/site level. However, in our application, as described in the ensuing paragraphs, we face very irregular sampling intensity, ecological factors measured at much lower resolution than our sampling sites and human intervention to transform land use. To accommodate these aspects, we adopt an explicitly spatial hierarchical modeling approach and fit the model to the data within a Bayesian framework. More elementary Bayesian approaches develop prior probabilities of observing species (e.g., Aspinall (1992); Aspinall and Veitch (1993)) or communities (e.g., Brzeziecki *et al.* (1993)). Linkage between occurrence and discretized environmental predictions is made, enabling a posterior predicted probability for the modeled entity at a site with specified environmental features. Also, for us, spatial structure is introduced through random effects rather than at the data stage. This contrasts with, e.g., Hoeting *et al.* (2000) as well as the GRASP approach.

Hence the objective of this paper is to develop a fully model-based multilevel approach to illuminate biodiversity concepts such as species range, richness and turnover. The novelty in our contribution is to work at the species level modeling presence/absence across the region for each species under study. As we clarify below, possibly confounded insight arises when implementing standard regression modeling for the observed richness; it is preferable to build regression models at the species level. In addition, we introduce spatial association in presence/absence across the domain of investigation. Causal ecological explanations such as dispersal as well as omitted (unobserved) variables with spatial pattern such as local smoothness of geological features, suggest that at sufficiently high resolution we anticipate that presence/absence of species at one location will be associated with their presence/absence at neighboring locations.

The domain we study here is a portion (Kogelberg-Hawequas subregion) of the

Cape Floristic Region in South Africa. Arguably, the data set we use is the largest and highest quality of its kind in the world for studying biodiversity. Still, while in some parts of this domain sampling is fairly intensive, in others it is much less or nonexistent. Also, in many places the region has been transformed due to human involvement. The "natural" state has been replaced by an alternative land use, e.g. an agricultural, residential or commercial use. This implies that there is a notion of *potential* presence/absence as well as *transformed* (or adjusted) presence/absence. These notions will be defined at areal unit (1 minute by 1 minute pixel) level. However, relative to this scale of resolution, observed presence/absence for a sampling location is at the point level.

Hence, we envision a multilevel model. That is, we model potential presence/ absence, transformed presence/absence given potential absence and observed presence/absence given transformed presence/absence. With regard to the biodiversity questions above, potential presence/absence is of primary interest. We set this multilevel model within a Bayesian framework. The output of the Bayesian model fitting enables model measures to convey species range, to capture species richness, to explain species richness, to study species turnover across the domain.

The format of the paper is as follows. Sections 2 and 3 provide the ecological motivation for the problem and description of the dataset used to address it. Section 4 develops the species level modeling. Section 5 presents the novel biodiversity measures which arise under this modeling. Section 6 presents the analysis of the data under this modeling. Finally, section 7 offer some discussion and extension.

2 Motivation: The Cape Floristic Kingdom

The focal area for this study of patterns of species distributions and biodiversity is the Cape Floristic Kingdom or Region (CFR), the smallest of the world's six floral kingdoms (Takhtajan, 1986). This encompasses a very small region of southwestern South Africa, about 90,000 km^2, centered on the Cape of Good Hope. It has long been recognized for high levels of plant species diversity and endemism across all spatial scales. The region includes about 9000 plant species, 69% of which are found nowhere else. This is globally one of the highest concentrations of endemic plant species in the world (Meyers *et al.*, 2000) – as diverse as many of the world's tropical rain forests. The CFR also apparently has the highest density of globally endangered plant species (Rebelo, 2002b).

The plant diversity in the CFR is concentrated in relatively few groups, like the icon flowering plant family of South Africa, the Proteaceae. We have chosen to focus on modeling the biogeography and biodiversity patterns of this family because the data on species distribution patterns are sufficiently rich and detailed to allow complex modelling. The Proteaceae have also shown a remarkable level of speciation with about 400 species across Africa, of which 330 species are 99% restricted to the CFR. Of those 330 species at least 152 are listed as "threatened" with extinction by the International Union for the Conservation of Nature.

3 Data Description

To model species distribution patterns and biodiversity, we have relied on the Protea Atlas data set (Rebelo, 2002a). These data were collected beginning in 1991 as part of a 10-year project to document the distribution of this Flagship Family in Southern Africa, the Proteaceae (Rebelo, 2001). The original purpose of the project was to provide adequate data to determine the biogeographical and vegetation patterns within the CFR; to determine the optimal areas, reserve location and strategies to conserve the flora; and to obtain data at a scale suitable for modeling biogeographic patterns. Data were collected at ”record localities”: relatively uniform, geo-referenced areas typically of 50m in diameter. In addition to the presence (or absence) at the locality of protea species, abundance of each species along with selected environmental and species-level information were also tallied (Rebelo, 1991). To date some 60,000 localities have been recorded (including null sites), with a total of about 250,000 species counts from among some 375 proteas. The CFR and the Proteaceae together provide an extraordinarily detailed and rich dataset to model patterns of biogeography and biodiversity. This is one of the hottest hotspots of plant diversity and the protea data may be the closest there is to a complete presence/absence inventory of species for any biogeographic region.

The explanatory data we employ here were obtained from the South African Atlas of Hydrology and Climatology (Schultze, 1997) and downloaded from the Computing Centre for Water Research (CCWR), University of Natal. A large number of climatological traits are available as GIS raster layers with a minimum pixel resolution of 1 minute latitude by 1 minute longitude. We used the following geographical data as explanatory variables: elevation, mean annual precipitation, inter-annual coefficient of variation in precipitation, July (winter) minimum temperature, and January (summer) maximum temperature. In this analysis we restricted the areal extent of our analysis to a small sub-region of the full CFR: a roughly rectangular region with its upper left corner at 33 23.5’ S, 18 50.5’ E, and its lower right at 34 20.5’ S, 19 16.5’ E, with total area of 4,456 km^2. It comprises a rectangular area including the Kogelberg Biosphere Reserve and beyond, extending 41km east and 107km north from Cape Hangklip. We further restricted the analysis to 23 species of Proteaceae out of roughly 150 found within this rectangular area. For each species we scored the following traits: height (continuous), local population size (ordinal), dispersal mode (categorical), and ability to resprout after fire (categorical).

Transformed areas (by agriculture, afforestation, alien plants and urbanization) were obtained as a GIS data layer from R. Cowling (private communication). 25% of the Cape has been transformed, mainly in the lowlands on more fertile soils where rainfall is adequate. Most of the transformation outside of these areas, on the infertile mountains, is due to dense alien invader species, which are the single biggest threat to Fynbos vegetation and, the Proteaceae. There is no sampling in transformed areas since no protea are currently found there.

4 Modeling and Implementation

We begin by proposing a model to infer about the distribution of individual species over a region of interest. It is assumed that this distribution depends upon the locally varying nature of the region. But also it depends upon attributes of the species. Since many of the variables which define the local features are observed at pixel level (at some scale of resolution) we suppose a regular lattice of cells over the region. The model must address several important issues, such as the fact that a pixel is never explored extensively for presence or absence, that only a subset of the pixels are actually ever observed resulting in 'holes' in the region, that for many pixels at least a portion has been transformed by human activity (as described above). After introducing the model and obtaining the likelihood we discuss the computational implementation and describe how to obtain inference of interest under the model specification.

4.1 The Proposed Model

In order to model potential presence/absence for a species we have to clarify the meaning of this binary outcome. Ecologists customarily view species range as an areal construct, e.g., the range of occupancy interpreted as the convex hull of the occurrence locations. This suggests that we adopt an areal unit conceptualization for presence/absence. In fact we view presence/absence with regard to a regular grid of cells. Moreover, the data layers providing local features have been prepared in minute by minute grid cells. So, we assume this scale for presence/absence as well resulting in roughly $37,000$ units for the entire CFR and 1554 areal units (pixels) in our study region. In this subregion the pixels are rectangular, approximately 1.85 km × 1.55 km. If we were to formalize potential presence/absence as a binary spatial process over this region, the value of the process on a grid cell becomes a block average (see, e.g. Cressie (1993)). With probability 1 the value will belong to $(0, 1)$; a binary response for an areal unit can not be modeled using a binary process. However, it can be modelled using a latent binary process.

Suppose we let $X_i^{(k)}$ denote the *potential* presence/absence state for the k^{th} species in the i^{th} site with presence=1 and absence=0. Then we set $P(X_i^{(k)} = 1) = p_i^{(k)}$, and we model $p_i^{(k)}$, the probability that species k is potentially found in areal unit i, using a binary process. That is, let $\lambda^{(k)}(\mathbf{s})$ be a binary process over the region and let $p_i^{(k)}$ be the block average of this process over unit i. That is,

$$p_i^{(k)} = \frac{1}{|A_i|} \int_{\text{pixel i}} \lambda^{(k)}(\mathbf{s}) d\mathbf{s} = \frac{1}{|A_i|} \int_{\text{pixel i}} \mathbf{1}(\lambda^{(k)}(\mathbf{s}) = 1)\ d\mathbf{s} \tag{1}$$

where $|A_i|$ denotes the area of unit i. The interpretation associated with (1) is that $\lambda^{(k)}(\mathbf{s})$ indicates the *suitability* of species k at location $\mathbf{s}$. The more $\lambda^{(k)}(\mathbf{s})$ in A_i which equal 1, the more suitable species k for A_i, hence the greater the chance for potential presence.

Next, let $V_i^{(k)}$ denote the transformed presence/absence state for the k^{th} species in the i^{th} unit. Let $T(\mathbf{s})$ be an indicator process indicating whether location $\mathbf{s}$ is transformed ($T(\mathbf{s}) = 1$) or not ($T(\mathbf{s}) = 0$). Then at $\mathbf{s}$ we need both $T(\mathbf{s}) = 0$ and $\lambda^{(k)}(\mathbf{s}) = 1$ in order that location $\mathbf{s}$ be suitable under transformation, i.e., we need both suitability and availability. Therefore,

$$P(V_i^{(k)} = 1) = \frac{1}{|A_i|} \int_{\text{pixel i}} \mathbf{1}(T(\mathbf{s}) = 0)\mathbf{1}(\lambda^{(k)}(\mathbf{s}) = 1)\ d\mathbf{s}. \tag{2}$$

If we make the simplifying (and hopefully plausible) assumption that availability is uncorrelated with suitability, then (2) reduces to

$$P(V_i^{(k)} = 1) = (1 - U_i)p_i^{(k)} \tag{3}$$

where U_i denotes the proportion of area in the i^{th} unit which is transformed, $0 \leq U_i \leq 1$. We adopt (3) in the sequel.

Next, assume that unit i has been visited n_i times in untransformed areas within the unit. Further, let $Y_{ij}^{(k)}$ be the presence/absence status of the k^{th} species in the i^{th} unit at the j^{th} sampling location within that unit. We need to model $P(Y_{ij}^{(k)}|V_i^{(k)} = 1)$. Given $V_i^{(k)} = 1$, we view the $Y_{ij}^{(k)}$ as i.i.d. Bernoulli trials with success probability $q_i^{(k)}$. Of course, given $V_i^{(k)} = 0$, $Y_{ij}^{(k)} = 0$ with probability 1. Based upon its interpretation as a conditional probability, $q_i^{(k)}$ is thought of as a ratio of integrals, i.e.,

$$q_i^{(k)} = \frac{\int_{\text{pixel i}} 1(T(\mathbf{s}) = 0)1(\tilde{\lambda}^{(k)}(\mathbf{s}) = 1)d\mathbf{s}}{\int_{\text{pixel i}} 1(T(\mathbf{s}) = 0)1(\lambda^{(k)}(\mathbf{s}))d\mathbf{s}} \tag{4}$$

In (4), $\tilde{\lambda}^{(k)}(\mathbf{s})$ is another binary process which indicates actual presence/absence of species k at location $\mathbf{s}$. Note that $\tilde{\lambda}^{(k)}(\mathbf{s}) = 1$ implies that $\lambda^{(k)}(\mathbf{s}) = 1$, i.e., presence implies suitability so $0 \leq q_i^{(k)} \leq 1$. But also, $\tilde{\lambda}^{(k)}(\mathbf{s}) = 1$ implies $T(\mathbf{s}) = 0$, i.e., presence implies availability. So the numerator simplifies to $\int_{\text{pixel i}} 1(\tilde{\lambda}^{(k)}(\mathbf{s}) = 1)d\mathbf{s}$ which, divided by $|A_i|$ is the expected probability of presence/absence at a randomly selected location in A_i. As a result, using (3), $P(Y_{ij}^{(k)} = 1) = q_i^{(k)}(1 - U_i)p_i^{(k)}$.

Note that the probabilities associated with $X_i^{(k)} = 1$, $V_i^{(k)} = 1$ and $Y_{ij}^{(k)} = 1$ all have interpretations through extent of "switches turned on". That is, in modeling for the $p_i^{(k)}$ and $q_i^{(k)}$, we look for ecological variables or species attributes which are expected to affect the "number" of $\lambda^{(k)}(\mathbf{s})$ or $\lambda^{\widetilde{(k)}}(\mathbf{s})$. Also, note that given $V_i^{(k)} = 1$, by sufficiency, we can work with $Y_{i+}^{(k)} = \sum_{j=1}^{n_i} Y_{ij}^{(k)} \sim Bi(n_i, q_i^{(k)})$. For an unsampled pixel ($n_i = 0$) there will be no contribution to the likelihood. For a sampled pixel ($n_i \geq 1$) there will be a contribution to the likelihood and, in fact, we can marginalize over $V_i^{(k)}$ to give, for $y > 0$, $P(Y_{i+}^{(k)} = y) = \begin{pmatrix} n_i \\ y \end{pmatrix} \left(q_i^{(k)}\right)^y \left(1 - q_i^{(k)}\right)^{n_i - y} (1 - U_i)p_i^{(k)}$, and for $y = 0$, $\left(1 - q_i^{(k)}\right)^{n_i} (1 - U_i)p_i^{(k)} + \left(1 - (1 - U_i)p_i^{(k)}\right)$. The two components of this latter expression have immediate interpretation. The first provides the probability that the species exists in pixel i but has not been observed while the second provides the probability that it is not present in the pixel.

We next turn to modeling $p_i^{(k)}$ and $q_i^{(k)}$. For $p_i^{(k)}$ we use a logistic regression conditional on unit level characteristics, unit level spatial random effects, species level attributes and species level random effects. Logistic regression for presence/absence modeling has been widely used in the ecological literature. The survey paper of Guisan and Zimmerman (2000) provides discussion and extensive referencing.

Let

$$\log\left(\frac{p_i^{(k)}}{1-p_i^{(k)}}\right) = \mathbf{w}_i'\boldsymbol{\beta}_k + \Psi_k + \rho_i, \tag{5}$$

where $\mathbf{w}_i$ is a vector of pixel-level characteristics, and the $\boldsymbol{\beta}_k$'s are species level coefficients associated with the pixel-level covariates. Therefore, the model allows the flexibility of each species having a different coefficient for each pixel-level covariate, i.e., that each species can react differently to the local environment. The assumption that $\boldsymbol{\beta}_k$ is constant across species converts (5) to an additive form in i and k which need not be appropriate. The Ψ_k's are defined below using species level attributes and an overall intercept. The Ψ_k's are viewed as an intercept specification for each of the species. Hence, there is no intercept in $\boldsymbol{\beta}_k$. We will discuss further the parameterization of Ψ_k in Section 4.2. The ρ_i's denote spatially associated random effects. In other words we believe that the potential probability of presence/absence of species k at pixel i, is also affected by its direct neighbors. We expect pixels which are close together to behave in a similar fashion in terms of their species distribution. We employ a CAR model (Besag, 1974) to capture the spatial association in the ρ_i. In this regard, Hoeting *et al.* (2000) employ a single-stage autologistic model to directly describe spatial association between the $X_i^{(k)}$ across i. To accommodate the untractable calculation of the normalizing constant arising under this model, they employ a pseudo-likelihood approximation.

From above, $q_i^{(k)}$ is interpreted as the chance of falling in the realized patch for species k in pixel i relative to the potential patch for this species in pixel i. We model $q_i^{(k)}$ on the logit scale setting

$$\log\left(\frac{q_i^{(k)}}{1-q_i^{(k)}}\right) = \tilde{\mathbf{w}}_i{}'\tilde{\boldsymbol{\beta}}_k + \tilde{\mathbf{z}}_k\tilde{\boldsymbol{\gamma}}. \tag{6}$$

In (6), $\tilde{\mathbf{w}}_i$ are location characteristics and $\tilde{\mathbf{z}}_k$ are species attributes which may affect $q_i^{(k)}$. Factors which are anticipated to affect realized patch size should accord with those used to model potential patch size. In fact if we are modeling the joint distribution of the $X_{ij}^{(k)}$ and $V_i^{(k)}$ given these factors, then the marginal specification for $V_i^{(k)}$ and the conditional specification for $Y_{ij}^{(k)}$ given $V_i^{(k)}$ should both reflect these factors. There is no concern with regard to confounding.

From the equations above and defining $\boldsymbol{\theta}$ as the vector containing all the parameters involved in the model, we can thus immediately write the logarithm of the likelihood for $\mathbf{Y} = \{Y_{i+}^{(k)}\}$ as

$$l(\boldsymbol{\theta};\mathbf{Y}) \;=\; \sum_{i=1}^{N}\sum_{k=1}^{K} \min\,(1, Y_{i+}^{(k)})\left[Y_{i+}^{(k)}(\tilde{\mathbf{w}}_i{}'\tilde{\boldsymbol{\beta}}_k + \tilde{\mathbf{z}}_k\tilde{\boldsymbol{\gamma}}) - \right.$$

$$
\begin{aligned}
&- n_i \log\left(1+\exp(\tilde{\mathbf{w}}_i{}'\tilde{\boldsymbol{\beta}}_k + \tilde{\mathbf{z}}_k\tilde{\boldsymbol{\gamma}})\right) + \log\left((1-U_i)p_i^{(k)}\right)\Big] + \qquad (7)\\
&+(1-\min(1,Y_{i+}^{(k)}))\Big[\log\left((1-q_i^{(k)})^{n_i}(1-U_i)p_i^{(k)} + 1 - (1-U_i)p_i^{(k)}\right)\Big].
\end{aligned}
$$

With priors on $\boldsymbol{\beta}_k$, Ψ_k, $\tilde{\boldsymbol{\beta}}_k$, $\tilde{\boldsymbol{\gamma}}$, and ρ_i, we have a fully specified Bayesian model.

As noted above, we can still use (7) in a formal way for the likelihood even if $n_i = 0$. There will just be no contribution from the i^{th} pixel. However, from (5), we can learn about $p_i^{(k)}$. That is, $\mathbf{w}_i$ is known, we learn about $\boldsymbol{\beta}_k$ and Ψ_k from other pixels and, due to the spatial modeling for ρ_i, we can still learn about it from its neighbors through $\rho_i \mid \rho_j, j \neq i$. The special case where $U_i = 1$ implies $n_i = 0$. Hence our modeling can accommodate ”holes” in the region resulting from totally transformed regions or unsampled regions.

4.2 Details of the Prior Specification and Sampling the Posterior Distribution

From the Bayesian point of view, a model is fully specified after assigning the prior distribution to all its unknown quantities. Here, we have to assign prior distributions to the coefficients of the area level characteristics $\boldsymbol{\beta}_k$, the species effects Ψ_k, the spatial random effects ρ_i, and also the coefficients of the second level of hierarchy $\tilde{\beta}_k$ and $\tilde{\gamma}_k$. For each of the parameters β_k, $\tilde{\beta}_k$ and $\tilde{\gamma}_k$ we assign independent normal prior distributions centered at 0 and with large variance.

As previously noted, the Ψ_k's are species random effects. *A priori*, we assume that, conditioned on μ, $\boldsymbol{\gamma}$ and σ^2_ψ, the Ψ_ks are independent and identically distributed following a normal distribution with mean $\mu + \mathbf{z}_k'\boldsymbol{\gamma}$ and common variance σ^2_ψ. In other words, analogous to (6), each species effect Ψ_k can be described by an overall intercept plus, say, L species level covariates. We then assign a normal prior distribution to μ centered at zero with a large variance, and also a normal prior distribution to $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_L)'$ centered at 0 with a large variance. For the variance of Ψ_k, σ^2_ψ, we assign an Inverse Gamma prior with infinite variance. One could introduce the mean structure of Ψ_k into the first level of hierarchy, together with the area level covariates and the spatial random effects plus a species random effect. However, computation would become unstable because the species random effects would be collinear with the species level attributes in the design matrix.

We are now left to assign the prior distribution of the spatial random effects. We presume there exists local spatially structured variation in the presence/absence of species at each pixel i. This prior knowledge can be described through a nearest neighbor Markov random field model (Besag, 1974). In other words we also correct the overall trend of the logistic regression in equation (5) for spatial association. For this class of prior model, the conditional distribution of the spatial random effect in pixel i, given values for the spatial random effect in all other areas $j \neq i$, depends only on the spatial random effect of the neighbouring pixels, δi of i. Here we say that pixel i is a neighbor of j if they share the same boundary. In particular, with a Gaussian Markov random field, the distribution of the spatial random effect at pixel i, conditioned on all the other pixels has the distribution

$$\rho_i|\rho_j \sim N\left(\frac{\sum_{j\in\delta i} w_{ij}\rho_j}{w_{i+}}, \frac{\sigma_\rho^2}{w_{i+}}\right), j \neq i \tag{8}$$

where w_{i+} denotes the total number of sites which are neighbors of i, $w_{ij} = 1$ if sites i and j share the same boundary and 0 otherwise. For the variance of the Gaussian Markov random field, σ_ρ^2, we also assign an Inverse Gamma prior with infinite variance.

Now, following the Bayes paradigm, the posterior distribution is proportional to prior times likelihood. Inference for the resulting posterior is done through simulation based model fitting using Gibbs sampling (Gelfand and Smith, 1990) in order to obtain samples from the posterior distribution.

In implementing the Gibbs sampling one needs to specify all the posterior full conditional distributions of all unknown quantities in the model. Following the model specifed above we obtain log concave full conditional distributions for the following parameters: β_k, γ_l, and ρ_i. Therefore we can use the adaptive rejection method introduced by Gilks and Wild (1992). The parameter μ, the intercept of the species random effect, have a Normal full conditional, which can be sampled from directly. The variance of the species random effects and of the spatial random effects both have Inverse Gamma full conditionals which are also immediate to sample from. The full conditional distributions for the parameters $\tilde{\beta_k}$ and $\tilde{\gamma_k}$ are not log concave. But as they are close to log concavity, this is not a problem; we can use adaptive rejection Metropolis sampling within Gibbs introduced by Gilks *et al.* (1995).

5 Inference with regard to biodiversity

The model developed in section 4 evidently enables information about the importance of particular environmental factors as well as species attributes in explaining species presence or absence. However, it also enables us to introduce several model summaries which shed light on key issues in the study of biodiversity.

We begin with species range. Common presentation of species range is based upon extent of occupancy and range of occupancy. For the observed $\{Y_{ij}^{(k)}\}$, the convex hull of the set $\{Y_{ij}^{(k)} = 1\}$ provides the "observed" range. This estimate is purely descriptive allowing no inference. It fails to recognize holes in the hull where the species almost surely can not be present. It also fails to recognize edge effects in that presence/absence need not have a *hard* edge but perhaps a *soft* edge characterized by diminishing chance of presence. This is precisely what $p_i^{(k)}$ can capture. Moreover, since $p_i^{(k)}$ is a parametric function of $\boldsymbol{\theta}$, given samples from $p(\boldsymbol{\theta}|\mathbf{Y})$ we obtain a posterior distribution for $p_i^{(k)}$ at each k and i.

Using, for example, $E(p_i^{(k)}|\mathbf{Y})$ we can create a posterior surface for presence of species k. In fact, the display could take the form of a chloropleth or grey scale map or a smoothed contour plot. We can also obtain lower and upper surfaces to capture individual $1 - \alpha$ intervals estimates for the $p_i^{(k)}$. We suggest using the posterior mean surface as a species range (see Heikkinen and Högmander (1994), and Högmander and Møller (1995) in this regard). It is obviously more informative than the above observed range and it allows quantification of uncertainty. The range can be hardened by replacing expected probabilities below a specified threshold by 0. The surface plot of the $E(p_i^{(k)}|\mathbf{Y})$ provides a picture of the potential range for species k. That is, in the absence of human intervention, where in the region it is likely that the species would be found. A surface plot of $(1-U_i)\,E(p_i^{(k)}|\mathbf{Y})$ provides an adjusted or *transformed* range reflecting where the species is likely to be found, adjusting for human intervention. We note that the ranges we have proposed can only be interpreted with respect to the domain of study.

Another important feature is the species richness. The *observed* species richness in pixel i is $\sum_{k=1}^{K} 1(Y_{i+}^{(k)} > 0)$ for pixels where $n_i > 0$ and $U_i > 0$. Again, this is a purely descriptive summary. Regression models can be used to explain these observed richness values using environmental features and enable interpolation to unobserved sites. Work of this sort has been mentioned in the introduction and enables refined prediction of species richness. See Guisan and Zimmerman (2000) in this regard. Under our model, the analogue for pixel i is the posterior distribution of $\sum_{k=1}^{K} X_i^{(k)}|\mathbf{Y}$. This posterior speaks to potential richness. That is, in the absence of human intervention, it is the *number* of species we would expect to find in pixel i. Converting to the distribution of $(1-U_i)\sum_{k=1}^{K} X_i^{(k)}|\mathbf{Y}$ modifies to transformed richness, i.e., the number of species we expect to find in the pixel, adjusting for human intervention. Each is of ecological interest but the latter will better align with observed richness.

Using the posterior mean across i we can create a posterior potential richness surface by plotting $E(\sum X_i^{(k)}|\mathbf{Y}) = \sum E(p_i^{(k)}|\mathbf{Y})$ versus i; similarly a posterior transformed richness surface can be obtained. These can be displayed in a fashion similar to that proposed above for species range. It is important to note that, under our modeling, species richness can only be inferred within the domain of study and is only relative to the set of species which have been modeled.

Since traditional modeling of species richness attempts an explanation in terms of local environmental characteristics, what does our model, implemented at the species level, offer in this regard? We note that a regression model to explain richness can be misleading. For a particular ecological feature such as altitude or rainfall, one species may prefer high levels for both, another species high for one, low for the other. How can a single regression coefficient make sense of this? Indeed, this is why we work with species level coefficients. Expressed in different terms, when similar species richness is observed at two different locations, the set of species present at one location need not be the same as those at the second. Are those at the second "replacements" for those at the first, i.e., ones which respond to the ecology in a similar way to those at the first? Or do we have a much different ecology with a quite different set of species?

In our setting we can offer the same clarification. Since $\log \frac{p_i^{(k)}}{1-p_i^{(k)}}$ strictly increases in $p_i^{(k)}$, suppose we look at $\sum_{k=1}^{K} E \log\left(\frac{p_i^{(k)}}{1-p_i^{(k)}}\right) \mid \mathbf{Y}$ rather than $\sum E(p_i^{(k)}|\mathbf{Y})$. With regard to environmental characteristics, the former involves $\mathbf{w}_i' E\left(\sum_{k=1}^K \boldsymbol{\beta}^{(k)}|\mathbf{Y}\right)$. We see that $\sum \boldsymbol{\beta}^{(k)}$ plays the role of the coefficient vector when modeling species richness directly. Thus we can see that for say the l^{th} component of $\sum \boldsymbol{\beta}^{(k)}$, it can be the case that for some k, $\sum \boldsymbol{\beta}_l^{(k)}$ is significantly positive while for other k it may be significantly negative. In aggregate, we need not find significance. (To work on the same scale of the $\sum \boldsymbol{\beta}_l^{(k)}$'s we might use the posterior of $K^{-1}\sum_{k=1}^K \boldsymbol{\beta}^{(k)}|\mathbf{Y}$).

Useful diagnostic displays are the scatterplots of $\sum_{k=1}^K E(p_i^{(k)}|\mathbf{Y})$ vs the components of $\mathbf{w}_i$. In addition, recalling that there is a spatial effect ρ_i associated with each i, a scatterplot of $\sum_{k=1}^K E(p_i^{(k)}|\mathbf{Y})$ or preferably $\sum_{k=1}^K E\log\left(\frac{p_i^{(k)}}{1-p_i^{(k)}}\right)|\mathbf{Y}$ vs $E(\rho_i|\mathbf{Y})$ can inform about the strength of spatial explanation. The display will show an "increasing" point cloud but the variability in trend of this cloud will reveal whether high (low) richness can arise without a large (small) spatial effect. We explore the issues of explaining richness in some detail in the data analysis of section 6.

A related comment is to note that an inappropriate alternative is to treat $\sum E(p_i^{(k)}|\mathbf{Y})$ as the "data" and fit a regression with spatial effects to this data. Apart from the possible confounding problems above, viewing $\sum E(p_i^{(k)}|\mathbf{Y})$ as the data, i.e., conditioning on them as fixed will result in underestimation of variability in the regression.

Finally, related to the foregoing discussion, we consider the issue of species turnover. That is, not only do we expect similar richness in neighboring pixels but also, that it arises from essentially common species. With increasing distance between pixels, not only do we expect less similarity in richness but also less overlap in species. We propose to, as well, use the $E(p_i^{(k)}|\mathbf{Y})$ to investigate this. Defining $E(\mathbf{p}_i|\mathbf{Y})$ to be the $K\times 1$ vector whose entries are $E(p_i^{(k)}|\mathbf{Y})$, overlap (turnover) is reflected by the similarity (difference) between these vectors. Using a neighborhood structure (first or second order), for each pixel i we propose to compute

$$d_i = \exp\left(-\sum_{j\in\delta i}\frac{||E(\mathbf{p}_i|\mathbf{Y}) - E(\mathbf{p}_j|\mathbf{Y})||}{\text{number of neighbors of } i}\right). \tag{9}$$

For cell i, d_i yields an average similarity (first or second order) of cell i with its neighbors. When d_i is large, high overlap is indicated; when d_i is small high turnover is indicated. A chloropleth map of the d_i will reveal where in the region overlap is high, where it is low. Scatterplots of d_i vs the components of $\mathbf{w}_i$ can illuminate which environmental attributes encourage turnover.

6 Analysing a subsample of the Cape Floristic Region

The study region (referred to as the Kogelberg-Hawequas subregion) lies in the western portion of the Cape Floristic Kingdom occupying 1554 grid cells. Figure 1 shows the region with the transformed areas indicated and the sampling locations overlaid. There are a total of 6957 sampling locations within the region including null sites (sites where nothing was observed). Pixel-level characteristics (the $\mathbf{w}_i$) include July minimum temperature, January maximum temperature, intra-cell coefficient of variation in annual preciptation (PPTCV), altitude, and mean annual precipitation. Grey scale maps of these data layers are supplied in Figures 2-6. As can be observed, July minimum temperature presents some high values in the South-West portion of the region, whereas January maximum temperature presents higher values in the North-West. The range of the January maximum temperature is higher than the July minimum temperature. PPTCV tends to be lower in the center portion of the sub-region. This region presents some variability with respect to elevation, presenting higher elevation in its mid-east. Finally, mean annual preciptation presents higher values towards the central portion of the sub-region.

Twenty three species were selected somewhat arbitrarily. They are listed alphabetically with abbreviated versions of their full latin names in Table 1. The most frequently occurring, *Leucadendron salignum*, was found at 622 of the sampling locations (42.09%). The least frequently occurring, *Sorocephalus imbricatus* was found at 3 locations (0.06%). Species level attributes (the $\mathbf{Z}_k$) include dispersal mechanism which has classifications "ant and mammal" (0) or "wind" (1), response to fire which has classifications "killed by fire" (0) or "resprouting after fire from the bole" (1), local population size which has classifications "less than 1000 plants" and "greater than 1000 plants", and average height of the species. The species attribute classifications are given in Table 1.

Figures 7-11 show posterior box plots for the $\beta_l^{(k)}$. In particular, Figure 7 shows these estimates for January maximum temperature. From this figure it is clear that an increase in January maximum temperature decreases the probability of presence for most of the species. Figure 8 shows these estimates for July minimum temperature. The importance of considering different coefficients for each of the species is clear, as different species present considerably different behavior in terms of July minimum temperature. For some species, an increase in July minimum results in an increase in the chance of presence whereas for others, the probability of presence decreases with higher July minimum temperature. Figure 9 shows the 23 point and interval estimates for elevation. In this case, most of the coefficients are not significant, as 0 lies within the 95% interval estimates. Figure 10 shows these estimates for mean annual preciptation. Notice that for most of the species, the 95% posterior intervals of the different coefficients are quite wide, indicating that the data present little information to learn about this set of coefficients. Finally, Figure 11 shows these estimates for PPTCV with interpretation similar to that of Figure 10.

Table 2 provides a summary of the inference for the coefficients of the species level attributes (gammas) implicitly within (5). Potential patch size is encouraged more through wind than ant or mammal dispersal. It is also somewhat encouraged by a bole response to fire and by a tendency toward larger local population size. It appears to be discouraged by increasing average plant height. Figure 12 shows the spatial adjustment to the $p_i^{(k)}$ in (5) using the posterior means of the ρ_i's. Spatial pattern, smoothed through the CAR model, is evident. For instance, spatial effects are small in the north/west portion, larger in central east and south east portion.

The former diminish potential patch size, the latter enhance it. Note, by comparison with Figure 1, that areas where there has been substantial transformation by humans do not appear to be associated with high or low spatial effects.

With regard to the modeling for $q_i^{(k)}$, as suggested below (6), we included all of the covariates that are in the model for $p_i^{(k)}$, i.e., we set $\tilde{\mathbf{w}}_i = \mathbf{w}_i$, $\tilde{\mathbf{z}}_k = \mathbf{z}_k$. Since only local population size emerged as significant, we have omitted a table of posterior summaries for this model.

Next, we turn to the patterns of species distributions and ranges described in the previous section. For species range we illustrate with three species, which are quite different from each other with regard to abundance and range, *Leucandendron salicifolium*, *Sorocephalus imbricatus*, and *Mimetes arboreus.* In each case we present the observed range, i.e. the locations where the species was observed as well as the transformed ranges. These maps are shown in Figure 13. Figure 14 shows the potential ranges for these three species. They are larger than the corresponding ranges in Figure 13 but, comparison across panels shows that this occurs in a species-specific fashion. Informally, we see that the model predicts quite well in terms of the probability of presence for each of the species in panels (a) and (b) of Figure 13. Notice that for the species in panel (c) of this figure the model is assigning some high probabilities to sites where it was not observed. Actually, what is happening is that the model is predicting the presence of another species, *Mimetes argenteus*, which is similar to, or a "sister species" of, *Mimetes arboreus*, but was not included in the present analysis (See Figure 15).

Turning to species richness, in Figure 16 we present observed richness (in the form of a grey scale map attaching an observed richness to each cell) as well as potential and transformed richness. When one compares the transformed richness with the observed one, it is clear that the model is able to predict the richness quite well. Following the discussion in section 5, with regard to explaining richness, Table 3 summarizes the posterior distribution for the $\sum_{k=1}^{23} \beta_l^{(k)}$. Altitude is suggestively significant while January maximum temperature is not in explaining richness. An increase in July minimum temperature represents an increase in richness, whereas mean annual preciptation, PPTCV and elevation have a behavior in the opposite direction. In this regard the scatterplot of potential richness vs each of the w_i's are informative. These are given in Figure 17. Also included in this figure is a scatterplot of potential richness vs (mean) spatial effect.

Finally to examine turnover, Figure 18 summarizes similarity of the $E(\mathbf{p}_i|\mathbf{Y})$ using first and second order neighbors. In both maps, the north-west portion of the region tends to present high values for the d_i, meaning that the probabilities for each of the species tend to be very similar among the cells there. Again this is all in accordance with the data, as we know that this is a transformed area and the probabilities of finding any of the species there tend to be quite low. On the other hand, the mid-south portion has some of the lowest values of d_i; the probabilities for each of the species in this area tend to be quite different across cells. Therefore we tend to find a greater number of different species in this area.

7 Discussion and Extensions

In several respects the proposed model marks a significant advance over previous efforts to model biogeographic patterns in species distributions. The model form enables quantification of uncertainty for all parameters. Moreover, the proposed model incorporates species-specific parameters for environmental or pixel-level characteristics, enabling it to capture and predict the differential responses of species to a full suite of ecological factors. By treating survey locations as Bernoulli trials, the proposed model can handle variation in areal sampling intensity. Finally, predictions of presence or absence take into account spatial autocorrelation, so that predictions for a particular pixel are influenced by the neighbouring pixels, whether or not the state of the pixel was actually observed.

From an ecological perspective, this approach provides a new, rigorous means of specifying the range or distribution of a species, the spatial pattern of species richness, and the spatial patterns in species turnover. In place of the conventional species range concepts - typically either the set of observed point locations, the convex hull for these points, or some more arbitrary encompassing polygon - our implementation specifies species range as a probability surface for an areal grid, with point estimates and confidence intervals available for each grid cell. This range specification is useful and intuitive, as it incorporates gaps in the species distribution as well as any declining probabilities of presence near distributional limits. Clearly this is a more realistic and practical means of specifying species ranges than conventional solid polygons. This approach also provides testable predictions about the potential range of a species. For example, Figure 13c shows that the modeled species was not observed in a discrete portion of its predicted range, but that this predicted range is a reasonably good description of a closely related or ”sister” species (Figure 15). Of course one may also test predictions empirically. The proposed model appears to perform well even with relatively sparse data: for example, the species predicted in Figure 13b was observed in only 18 grid cells out of 1554 in the study region.

As with individual species range, the proposed model provides a rigorous way to predict species richness in a particular grid cell. Since our richness measure is the sum of the probabilities of presence of individual species, it includes both an estimate and a quantification of uncertainty. Predicted species turnover is similarly a summation of the differences of probability of presence for each species between a grid cell and its neighborhood. This measure of turnover is novel and avoids the problems in conventional ecological concepts of turnover, which cannot provide any estimate of uncertainty, cannot differentiate between turnover caused by range edges and patchiness within distributions, and cannot directly suggest how individual species and site-level characteristics may influence turnover rate.

Future work with this modeling approach would incorporate other explanatory data layers. For example, geological information capturing fertility, texture and pH of the soil would be very important in explaining species presence/absence. Another possibility would be to introduce phylogenetic information into the modeling to replace species attributes. Indeed, the entire hierarchical modeling and inference strategy provided here can serve as a prototype for other biodiversity data analysis.

For instance, in principle, species abundance could be studied replacing the binary $Y_{ij}^{(k)}$ with counts. However, the modeling will require some modification since we would want to associate abundance with an area rather than a point. Of course, a sufficiently rich dataset will be required to take advantage of the opportunities which are available. The approach enables assessment of scale of resolution effects by fitting a given model at different resolutions. It also enables comparison of regions with regard to biodiversity issues. Finally, the Bayesian approach for fitting such hierarchical models is clearly advantageous in enabling such rich inference possibilities. Moreover, there may not be another feasible approach for actually fitting the models proposed here.

Bibliography

Aspinall, R. (1992) An Inductive Modeling Procedure Based on Bayes' Theorem for Analysis of Pattern in Spatial Data. *International Journal of Geographic Information Systems*, **6,** 105–121.

Aspinall, R. and Veitch, N. (1993) Habitat Mapping from Satellite Imagery and Wildlife Survey using a Bayesian Modeling Procedure in a GIS. *Photogrammetric Engineering and Remote Sensing*, **59,** 537–543.

Besag, J. E. (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series* B, **36,** 192–225.

Brzeziecki, B., Kienast, F. and Wildi, O. (1993) A Simulated Map of the Potential Natural Forest Vegetation of Switzerland. *Journal of Vegetation Science*, **4,** 499–508.

Colwell, R.K. and Lees, D.C. (2000) The Mid-Domain Effect: Geometric Constraints on the Geography of Species Richness. *Trends in Ecology and Evolution*, **15,** 70–76.

Cressie, N.A.C. (1993) *Statistics for Spatial Data. Revised Edition.* John Wiley & Sons, Inc.

Currie, D.M. (1991) Energy and Large-Scale Patterns of Animal - and Plant - Species Richness. *American Naturalist*, **137,** 27–40.

Darwin, C. (1872) *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life. Sixth edition.* John Murray, London, UK.

Ferrier, S., Drielsma, M., Manion, G. and Watson, G. (2002) Extended Statistical Approaches to Modelling Spatial Pattern in Biodiversity in Northeast New South Wales. II. community-level modeling. *Biodiversity and Conservation*, **11,** 2309–2338.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85,** 398–409.

Gilks, W.R., Best, N. and Tan, K. K. C. (1995) Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Applied Statistics*, **44,** 455–472.

Gilks, W.R. and Wild, P. (1992) Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, **41,** no. 2, 337–348.

Guisan, A., Edwards, Jr. T.C. and Hastie, T. (2002) Generalized Linear and Generalized Additive Models in Studies of Species Distributions: Setting the Scene. *Ecological Modelling*, **157,** 89–100.

Guisan, A. and Zimmerman, N.E. (2000) Predictive Habitat Distribution Models in Ecology. *Ecological Modelling*, **135,** 147–186.

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

Heikkinen, R.K. (1996) Predicting Patterns of Vascular Plant Species Richness with Composite Variables: A Mesoscale Study in Finnish Lapland. *Vegetation*, **126,** 151–165.

Hoeting, J.A., Leecaster, M., and Bowden, D. (2000) An Improved Model for Spatially Correlated Binary Responses. *Journal of Agricultural, Biological and Environmental Statistics*, **5,** no. 1, 102–114.

Högmander, H. and Møller, J. (1995) Estimating Distribution Maps from Atlas Data Using Methods of Statistical Image Analysis. *Biometrics*, **51,** 393–404.

Hubbell, S.P. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography.* Princeton University Press, Princeton, NJ, USA.

Huston, M.A. (1994) *Biological Diversity. The Coexistence of Species on Changing Landscapes.* Cambridge University Press, Cambridge, UK.

Latham, R.E. and Ricklefs, R.E. (1993) Global Patterns of Tree Species Richness in Moist Forests: Energy-Diversity Theory Does Not Account for Variation in Species Richness. *Oikos*, **67,** 325–333.

Lehmann, A., Overton, J.M. and Leathwick, J.R. (2002) GRASP: Generalized Regression Analysis and Spatial Prediction. *Ecological Modelling*, **159,** 189–207.

Meyers, N., Mittermeier, R.A., C.G., Mittermeier, de Fonesca G.A.B. and Kent, J. (2000) Biodiversity Hotspots for Conservation Priorities. *Nature*, **403,** 853–858.

Owen, J.G. (1989) Patterns of Herpetofaunal Species Richness:Relation to Temperature, Precipitation and Variance in Elevation. *Journal of Biogeography*, **16,** 141–150.

Palmer, M.W. (1996) Variation in Species Richness: Towards a Unification of Hypotheses. *Folia Geobotanica et Phytotaxonomica (Praha)*, **29,** 511–530.

Rebelo, A.G. (1991) *Protea Atlas Manual: Instruction Booklet to the Protea Atlas Project.* Protea Atlas Project, Cape Town.

Rebelo, A.G. (2001) *Proteas: A Field Guide to the Proteas of Southern Africa.* Fernwood Press, Vlaeberg, South Africa (2nd Edition).

Rebelo, A.G. (2002a) The Protea Atlas Project. Technical Report. Retrieved on-line 12 May, 2002 from: `http://protea.worldonline.co.za/default.htm`.

Rebelo, A.G. (2002b) The State of Plants in the Cape Flora. In *Proceedings of a conference held at the Rosebank Hotel in Johannesburg*, pp. 18–43. G.H. Verdoorn and J. Le Roux (editors) The State of South Africa's Species. Endangered Wildlife Trust.

Ritchie, M.E. and Olff, H. (1999) Spatial Scaling Laws Yield a Synthetic Theory of Biodiversity. *Nature*, **400,** 557–560.

Rohde, K. (1992) Latitudinal Gradients in Species Diversity: the Search for the Primary Cause. *Oikos*, **65,** 514–527.

Rosenzweig, M.L. (1995) *Species Diversity in Space and Time.* Cambridge University Press, Cambridge, UK.

Schultze, R.E. (1997) South African Atlas of Agrohyrdology and Climatology. Technical Report. Report TT82/96. Water Research Commission, Pretoria, South Africa.

Takhtajan, A. (1986) *Floristic Regions of the World.* University of California Press, Berkeley, CA, USA.

Table 1: The List of Species (alphabetical order) with their Attributes

Species	Dispersal Mechanism	Response to fire	Height	Location Population Size
Aulax pallasia Stapf	Wind (1)	bole (1)	2.0	< 1000
A. umbellata (Thunb.)R.Br.	Wind (1)	killed (0)	2.0	> 1000
Leucadendron corymbosum P.J. Bergius	Wind (1)	killed (0)	1.5	< 1000
L. daphnoides (Thunb.)Meisn.	Mammal (0)	killed (0)	1.00	> 1000
L. microcephalum (Grand.)Gand. & Schinz	Wind (1)	killed (0)	1.25	> 1000
L. salicifolium (Salisb.)I. Williams	Wind (1)	killed (0)	2.00	< 1000
L. salignum P.J. Bergius	Wind (1)	bole (1)	0.50	> 1000
L. sessile R.Br.	Mammal (0)	killed (0)	1.0	> 1000
L. spissifolium (Salisb. ex Knight)I. Williams	Wind (1)	bole (1)	1.00	< 1000
L. tinctum I. Williams	Mammal (0)	killed (0)	0.75	< 1000
Leucospermum bolusii Grand.	Ant (0)	killed (0)	1.00	> 1000
L. oleifolium (P.J. Bergius)R. Br.	Ant (0)	killed (0)	0.75	< 1000
L. grandiflorum (Salisb.)R.Br.	Ant (0)	killed (0)	1.5	< 1000
Mimetes arboreus Rourke	Ant (0)	killed (0)	3.00	< 1000
M. cucullatus (L.)R.Br.	Ant (0)	bole (1)	1.0	< 1000
Orothamnus zeyheri Pappe ex Hook.f.	Ant (0)	killed (0)	2.90	< 1000
Protea acuminata Sims	Wind (1)	killed (0)	1.50	< 1000
P. nana (P.J.Bergius) Thunb.	Wind (1)	killed (0)	1.00	< 1000
P. neriifolia R.Br.	Wind (1)	killed (0)	2.50	> 1000
Serruria fasciflora Salisb. ex Knight	Ant (0)	killed (0)	0.5	> 1000
S. elongata (P.J. Bergius) R.Br.	Ant (0)	killed (0)	1.00	> 1000
Sorocephalus imbricatus (Thunb.)R.Br.	Ant (0)	killed (0)	1.2	< 1000
Spatalla curvifolia Salisb. ex Knight	Ant (0)	killed (0)	0.65	< 1000

Table 2: Posterior Summary of the Coefficients of the Species Level Attributes (γ's).

Covariate	Mean	2.5%	50.0%	97.5%
Dispersal Mechanism	1.40	-0.21	1.39	2.95
Response to fire	2.72	0.41	2.73	4.78
Height	-1.31	-2.47	-1.30	-0.11
Local Pop. Size	1.14	-0.69	1.16	2.87

Table 3: Posterior Summary of the Area Level Attributes in terms of Potential Richness ($\sum_k \beta_l^{(k)}$).

Covariate	2.5%	50%	97.5%
MAP	-26.73455	-19.43816	-12.70147
Julmint	10.1223	16.58706	23.73233
Janmaxt	-7.890144	-0.101283	6.829859
PPTCV	-13.45996	-7.281156	-1.342811
Elevation	-16.43027	-8.035193	0.4785387

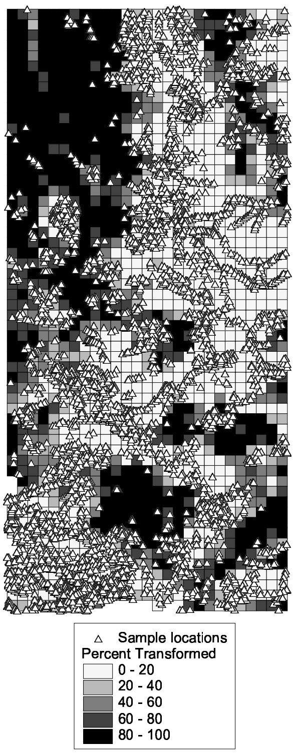

Figure 1: The Kogelberg-Hawequas sub-region overlaid with the sampling locations.

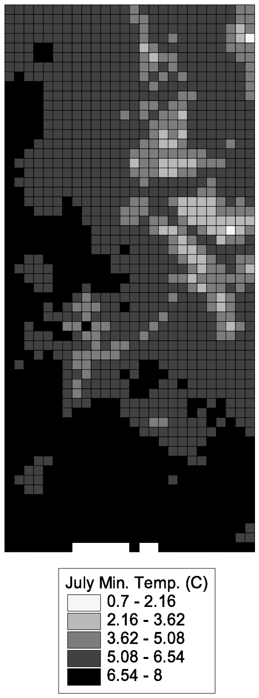

Figure 2: Data layer of July Minimum Temperature.

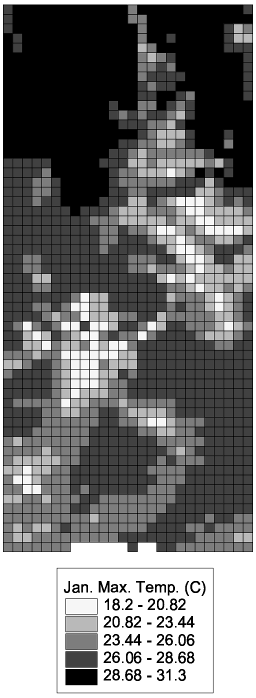

Figure 3: Data layer of January Maximum Temperature.

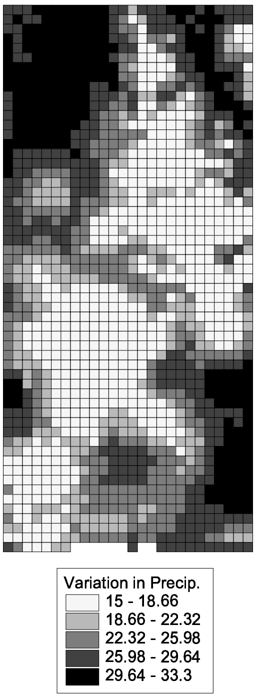

Figure 4: Data layer of PPTCV.

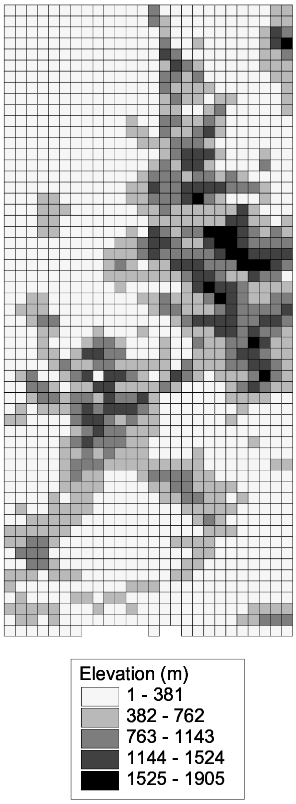

Figure 5: Data layer of elevation.

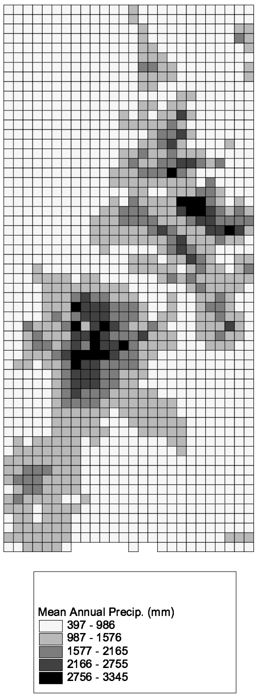

Figure 6: Data layer of mean annual preciptation.

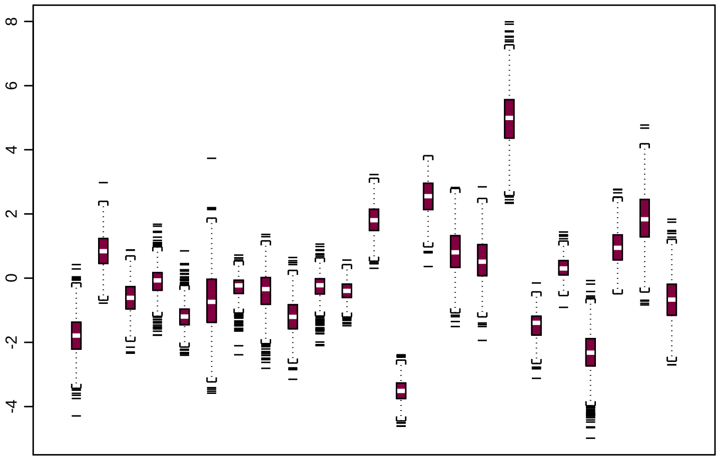

Figure 7: Posterior summary of the coefficients of January maximum temperature for each of the 23 species (Species in alphabetical order as per Table 1).

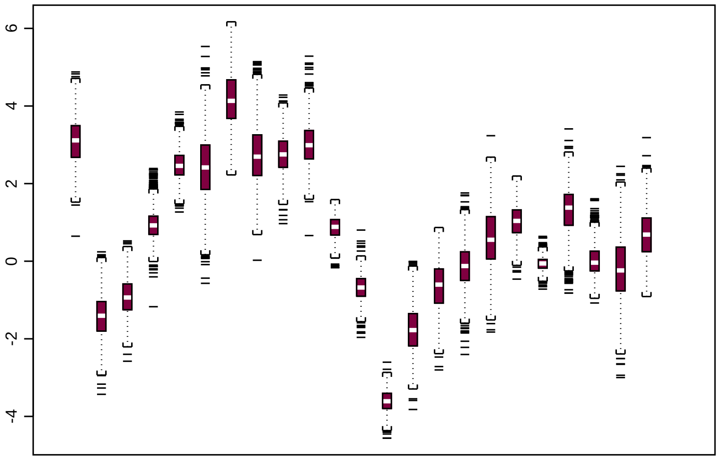

Figure 8: Posterior summary of the coefficients of July minimum temperature for each of the 23 species (Species in alphabetical order as per Table 1).

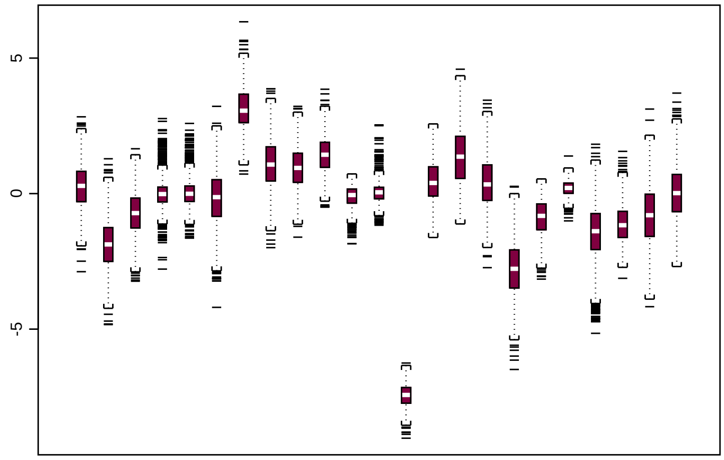

Figure 9: Posterior summary of each of the coefficients of elevation for each of the 23 species (Species in alphabetical order as per Table 1).

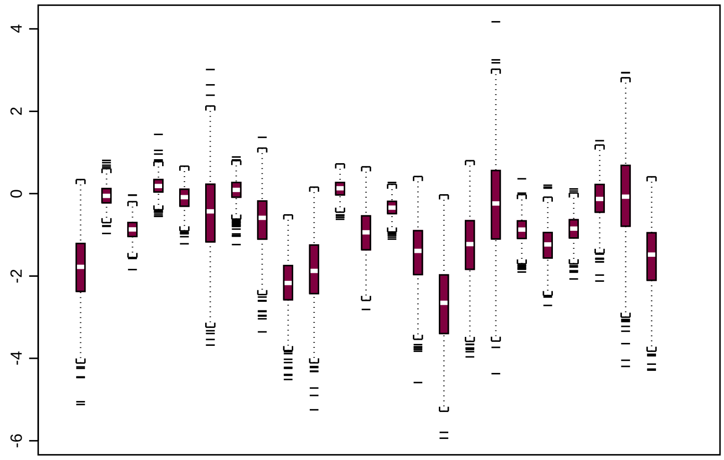

Figure 10: Posterior summary of each of the coefficients of mean annual preciptation for each of the 23 species (Species in alphabetical order as per Table 1).

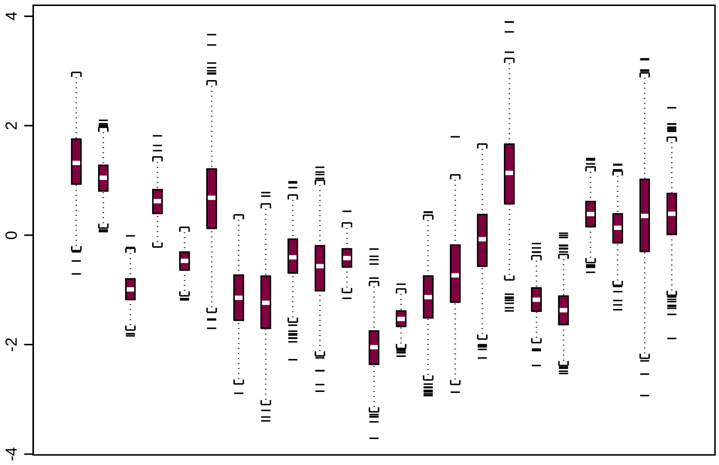

Figure 11: Posterior summary of each of the coefficients of PPTCV for each of the 23 species (Species in alphabetical order as per Table 1).

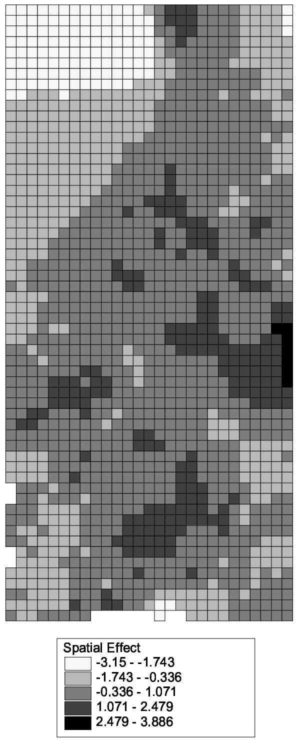

Figure 12: Posterior mean of the spatial effects $(\rho_i' s)$.

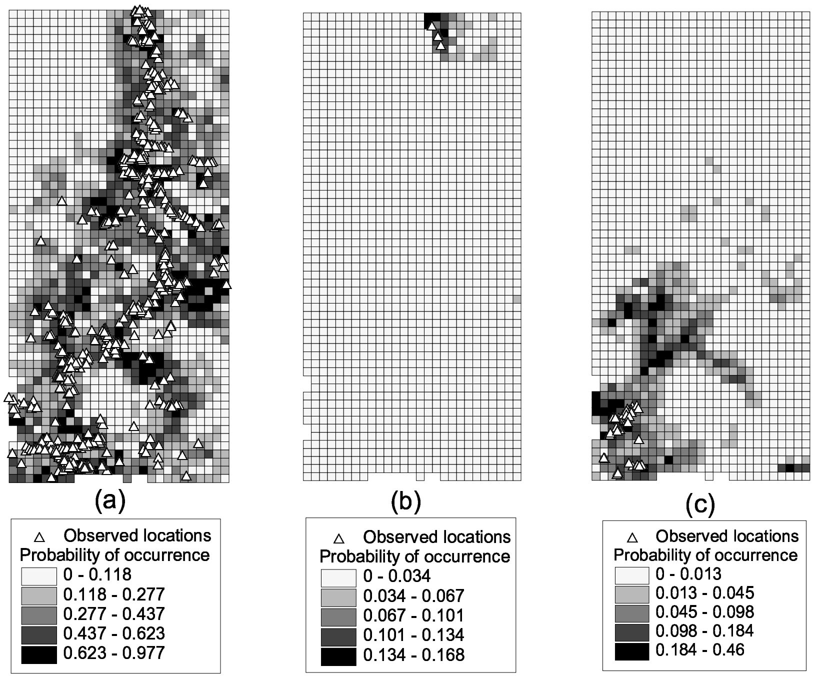

Figure 13: Observed and adjusted ranges for (a) *Leucondendron salicifolium*, (b) *Sorocephalus imbricatus*, and (c) *Mimetes arboreus*.

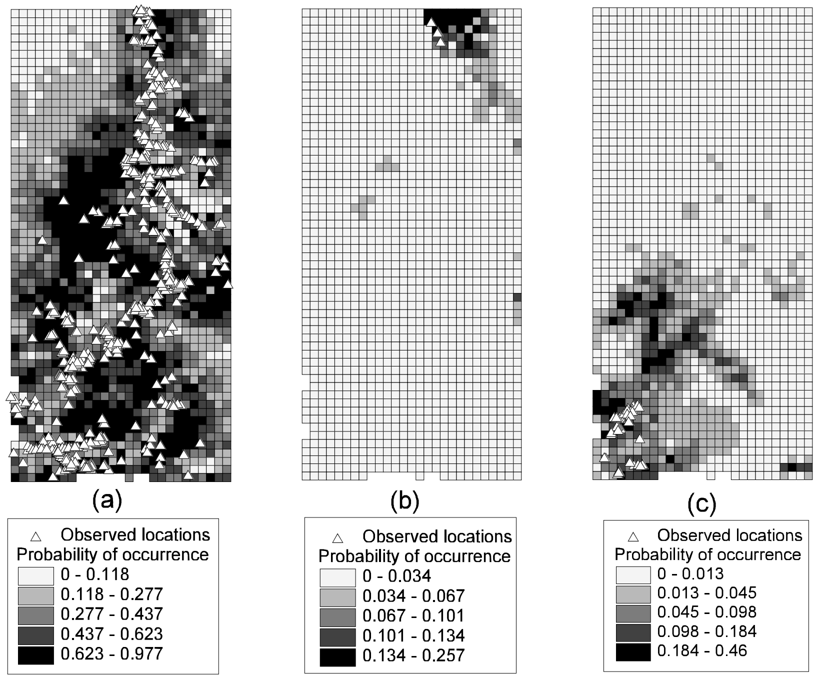

Figure 14: Potential range for (a) *Leucondendron salicifolium*, (b) *Sorocephalus imbricatus*, and (c) *Mimetes arboreus*.

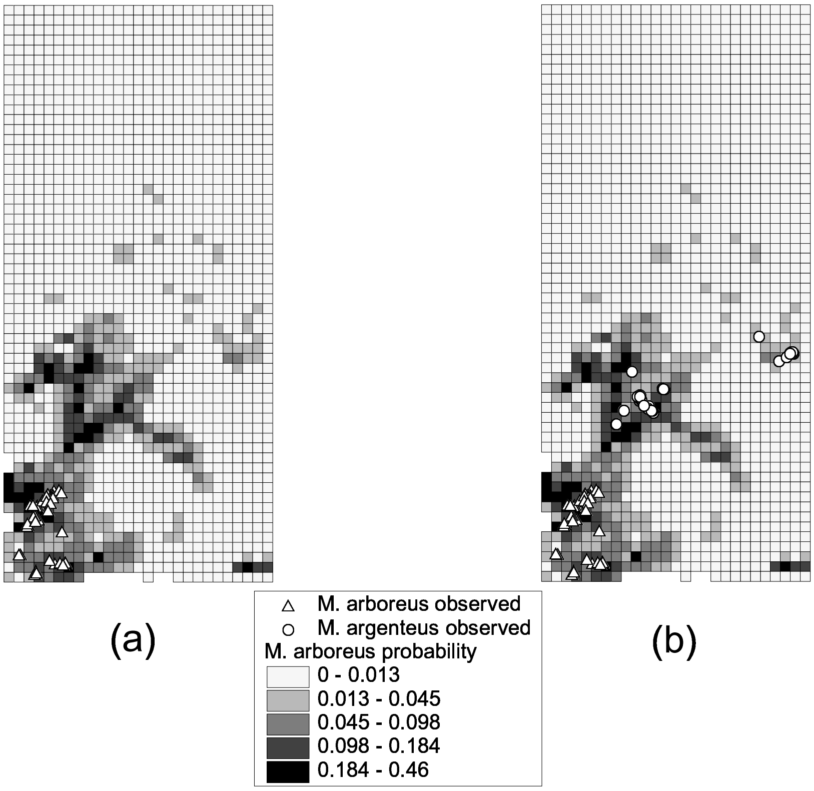

Figure 15: (a) Observed and adjusted range of the species *Mimetes arboreus* and (b) *Mimetes arboreus* overlayed with the observed range of the species *Mimetes argenteus* ("sister species")

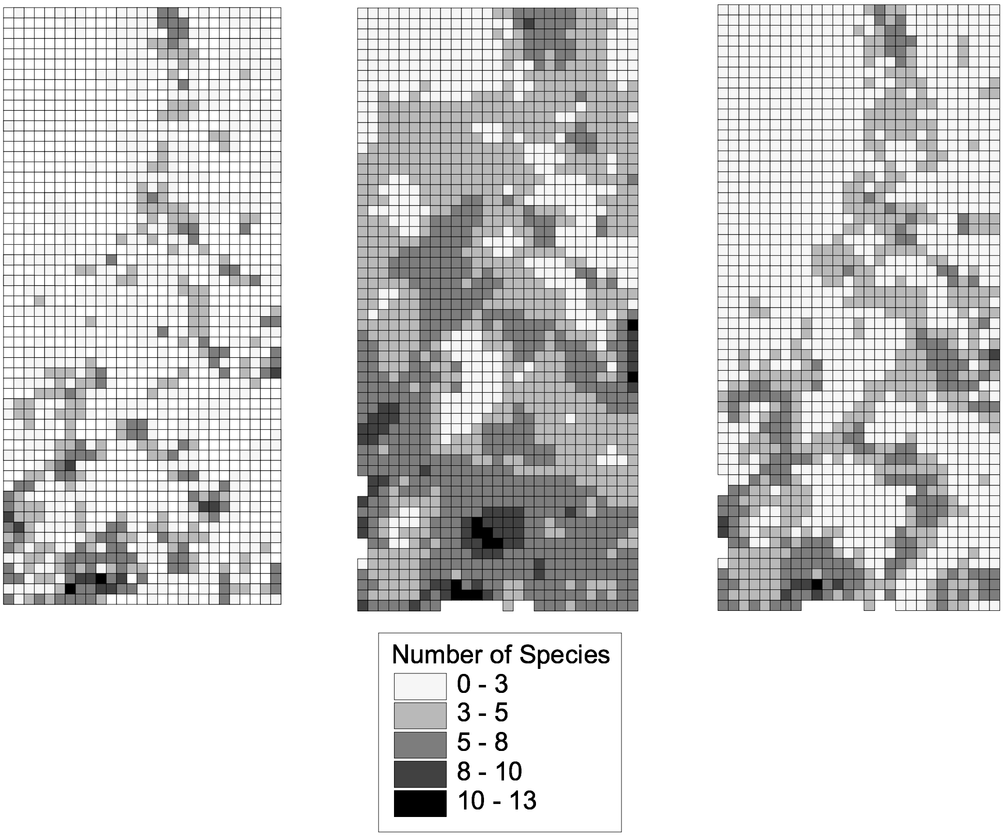

Figure 16: Observed, potential and adjusted richness.

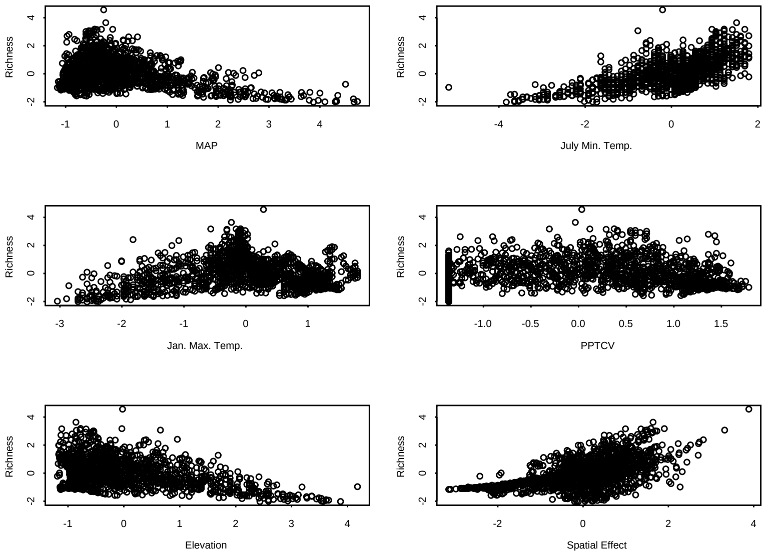

Figure 17: Area Level Covariates versus Potential Richness

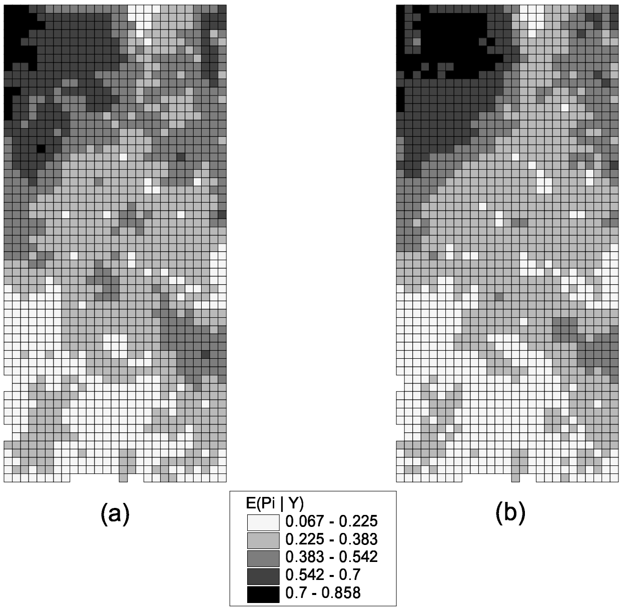

Figure 18: Maps of the similarity of the $E(\mathbf{p}_i|\mathbf{Y})$ using (a) first and (b) second order neighbors.