Appl. Statist. (2005)
54, *Part* 1, *pp.* 1–20

Modelling species diversity through species level hierarchical modelling

Alan E. Gelfand,

Duke University, Durham, USA

Alexandra M. Schmidt,

Federal University of Rio de Janeiro, Brazil

Shanshan Wu, John A. Silander, Jr, and Andrew Latimer

University of Connecticut, Storrs, USA

and Anthony G. Rebelo

National Botanic Institute, Kirstenbosch, South Africa

[Received August 2002. Final revision November 2003]

Summary. Understanding spatial patterns of species diversity and the distributions of individual species is a consuming problem in biogeography and conservation. The Cape floristic region of South Africa is a global hot spot of diversity and endemism, and the Protea atlas project, with about 60 000 site records across the region, provides an extraordinarily rich data set to model patterns of biodiversity. Model development is focused spatially at the scale of 1′ grid cells (about 37 000 cells total for the region). We report on results for 23 species of a flowering plant family known as *Proteaceae* (of about 330 in the Cape floristic region) for a defined subregion. Using a Bayesian framework, we developed a two-stage, spatially explicit, hierarchical logistic regression. Stage 1 models the *potential* probability of presence or 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 that it is present. Because the atlas data are not evenly distributed across the landscape, grid cells contain variable numbers 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 via Markov chain Monte Carlo 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 richness of species over the region. Summaries of the posterior for each environmental coefficient show which variables are most important in explaining the presence of species. Our initial results describe biogeographical patterns over the modelled region remarkably well. In particular, species local population size and mode of dispersal contribute significantly to predicting patterns, along with annual precipitation, the coefficient of variation in rainfall and elevation.

Keywords: Adaptive rejection method; Markov random field; Spatial logistic regression; Species range; Species richness

Address for correspondence: Alexandra M. Schmidt, Departamento de Métodos Estastísticos, Instituto de Matemática, Universidade Federal do Rio de Janerio, CP 68530, Rio de Janerio, CEP 21.945-970, Brazil.
E-mail: alex@dme.ufrj.br

 0035–9254/05/54001

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 19th century began cataloguing 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, we are left with the impression that there are many different universal explanations, each explicitly or implicitly claiming supremacy. Palmer (1996) listed 120 named hypotheses to explain patterns in biodiversity, and Rohde (1992) listed 28 that claim to explain just latitudinal patterns. This points up the difficulties that are encountered in developing explanatory models, and the utility of a richer flexible modelling. The past couple of years have seen at least three universal (ecological) explanations of species richness patterns championed:

(a) geometric constraints (Colwell and Lees, 2000),
(b) scaling of constrained resource acquisition (Ritchie and Olff, 1999) and
(c) neutrality of species in saturated systems (Hubbell, 2001).

Before this we have seen area proposed as the universal explanation for patterns of biodiversity (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 that are marshalled 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 that many ecologists think about data on distributions of species. In particular, a variety of statistical and algorithmic methods have been proposed, in conjunction with GISs, to enable spatial prediction of the distribution of species. The survey paper of Guisan and Zimmerman (2000) has provided 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, the presence (hence, implicitly, the absence) of a collection of species has been recorded resulting in a site- (rows) by-species (columns) presence–absence matrix. A classification-then-modelling strategy gathers either the sites into groups containing similar species ('communities') or the species into groups at similar sites ('assemblages'). Regression modelling follows, using environmental factors for the communities or species attributes for the assemblages. See, for example, 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 prevalences of species. Again these can be explained by using regression models 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 implemented 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 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 modelling 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) and 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 modelled entity at a site with specified environmental features. Also, for us, spatial structure is introduced through random effects in the modelling of the probability of presence rather than at the data stage. This contrasts with, for example, Hoeting *et al.* (2000) as well as the generalized regression analysis and spatial prediction approach.

Hence the objective of this paper is to develop a fully model-based multilevel approach to illuminate concepts of biodiversity such as the range of species, richness and turnover. The novelty in our contribution is to work at the species level, modelling presence or absence across the region for each species under study. As we clarify below, possibly confounded insight arises when implementing standard regression modelling for the observed richness; it is preferable to build regression models at the species level. In addition, we introduce spatial association in presence or 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 expect that the presence or absence of species at one location will be associated with their presence or absence at neighbouring locations.

The domain that we study here is a portion (Kogelberg–Hawequas subregion) of the Cape floristic region (CFR) in South Africa. Arguably, the data set that we use is the largest and highest quality of its kind in the world for studying biodiversity. Still, whereas in some parts of this domain sampling is fairly intensive, in others it is sparse or non-existent. Also, in many places the region has been transformed because of 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 or absence as well as *transformed* (or adjusted) presence or absence. These notions will be defined at an areal unit (1′-by-1′ pixel) level. However, relative to this scale of resolution, the observed presence or absence for a sampling location is at the point level.

Therefore, we envision a multilevel model, i.e. we model potential presence or absence, transformed presence or absence given potential absence and observed presence or absence given transformed presence or absence. With regard to the biodiversity questions above, potential presence or 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 the range of species, to capture the richness of species, to explain the richness of species and to study the turnover of species 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 data set that is used to address it. Section 4 develops the species level modelling. Section 5 presents the novel biodiversity measures which arise under this modelling. Section 6 presents the analysis of the data under this modelling. Finally, Section 7 offer some discussion and extension.

The data that are analysed in the paper can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Motivation: the Cape floristic kingdom

The focal area for this study of patterns of distributions of species and biodiversity is the CFR, the smallest of the world's six floral kingdoms (Takhtajan, 1986). This encompasses a very small region of south-western South Africa, about 90000 km^2, centred 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, 2002a).

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 modelling 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. Description of the data

To model species distribution patterns and biodiversity, we have relied on the Protea atlas data set (Rebelo, 2002b). These data were collected beginning in 1991 as part of a 10-year project to document the distribution of *Proteaceae*, the flagship family in Southern Africa (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 that is suitable for modelling biogeographic patterns. Data were collected at 'record localities': relatively uniform georeferenced areas typically of 50 m in diameter. In addition to the presence (or absence) at the locality of protea species, the abundance of each species along with selected environmental and species level information were also tallied (Rebelo, 1991). To date about 60000 localities have been recorded (including null sites), with a total of about 250000 species counts from among about 375 proteas. The CFR and the *Proteaceae* together provide an extraordinarily detailed and rich data set to model patterns of biogeography and biodiversity. This is one of the hottest hot spots of plant diversity and the protea data may be the closest that there is to a complete presence–absence inventory of species for any biogeographic region.

The explanatory data that 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, University of Natal. A large number of climatological traits are available as GIS raster layers with a minimum pixel resolution of 1′ latitude by 1′ longitude. We used the following geographical data as explanatory variables: the elevation, mean annual precipitation, interannual 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 subregion 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 a total area of 4456 km^2. It comprises a rectangular area including the Kogelberg Biosphere Reserve and beyond, extending 41 km east and 107 km north from Cape Hangklip. We further restricted the

analysis to 23 species of *Proteaceae* out of roughly 150 that are 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 these areas, on the infertile mountains, is due to dense alien invader species, which are a major threat to Fynbos vegetation and, in particular, to the *Proteaceae*. There is no sampling in transformed areas since no protea are currently found there.

4. Modelling 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 on the locally varying nature of the region. But also it depends on 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 is actually ever observed, resulting in 'holes' in the region, and 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 model proposed

To model potential presence or absence for a species we must clarify the meaning of this binary outcome. Ecologists customarily view the range of species as an areal construct, e.g. the range of occupancy interpreted as the convex hull of the locations of occurrence. 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, for example, Cressie (1993)). With probability 1 the value will belong to (0,1); a binary response for an areal unit cannot be modelled by using a binary process. However, it can be modelled by using a latent binary process.

Suppose that we let $X_i^{(k)}$ denote the *potential* presence or absence state for the kth species in the ith site with presence 1 and absence 0. Then we set $P(X_i^{(k)} = 1) = p_i^{(k)}$, and we conceptualize $p_i^{(k)}$, the probability that species k is potentially found in areal unit i, by using a binary process, i.e. 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,

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

where $|A_i|$ denotes the area of unit i. The interpretation that is associated with equation (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 cell A_i is for species k; hence the greater the chance for potential presence.

Next, let $V_i^{(k)}$ denote the transformed presence or absence state for the kth species in the ith 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 $\lambda^{(k)}(\mathbf{s})=1$ so that location $\mathbf{s}$ is suitable under transformation, i.e. we need both suitability and availability. Therefore,

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

If we make the simplifying (and hopefully plausible) assumption that, for each species, availability is uncorrelated with suitability, then equation (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 ith cell which is transformed, $0\leqslant U_i\leqslant 1$. We adopt equation (3) in what follows.

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 kth species in the ith unit at the jth 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 independent and identically distributed Bernoulli trials with success probability $q_i^{(k)}$. Of course, given $V_i^{(k)}=0$, $Y_{ij}^{(k)}=0$ with probability 1. On the basis of its interpretation as a conditional probability, $q_i^{(k)}$ is thought of as a ratio of integrals, i.e.

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

In equation (4), $\tilde{\lambda}^{(k)}(\mathbf{s})$ is another binary process which indicates the actual presence or 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 $0\leqslant q_i^{(k)}\leqslant 1$. But, also, $\tilde{\lambda}^{(k)}(\mathbf{s})=1$ implies that $T(\mathbf{s})=0$, i.e. presence implies availability. So the numerator simplifies to

$$\int_{\text{pixel } i}\mathbf{1}\{\tilde{\lambda}^{(k)}(\mathbf{s})=1\}\,\mathrm{d}\mathbf{s}$$

which, divided by $|A_i|$, is the expected probability of presence or absence at a randomly selected location in A_i. As a result, using equation (3), $P(Y_{ij}^{(k)}=1)=q_i^{(k)}(1-U_i)\,p_i^{(k)}$.

Note that the probabilities that are associated with $X_i^{(k)}=1$, $V_i^{(k)}=1$ and $Y_{ij}^{(k)}=1$ all have interpretations through the extent of 'switches turned on', i.e., in modelling 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 $\tilde{\lambda}^{(k)}(\mathbf{s})$ that is turned on in cell i. Also, note that, given $V_i^{(k)}=1$, by sufficiency, we can work with $Y_{i+}^{(k)}=\Sigma_{j=1}^{n_i}Y_{ij}^{(k)}\sim\text{Bi}(n_i,q_i^{(k)})$. For an unsampled pixel ($n_i=0$) there will be no contribution to the likelihood. For a sampled pixel ($n_i\geqslant 1$) there will be a contribution to the likelihood and, in fact, we can marginalize over $V_i^{(k)}$ to give, for $y>0$,

$$P(Y_{i+}^{(k)}=y)=\binom{n_i}{y}(q_i^{(k)})^y(1-q_i^{(k)})^{n_i-y}(1-U_i)\,p_i^{(k)}$$

and, for $y=0$, $(1-q_i^{(k)})^{n_i}(1-U_i)\,p_i^{(k)}+\{1-(1-U_i)\,p_i^{(k)}\}$. 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 whereas the second provides the probability that it is not present in the pixel.

We next turn to modelling $p_i^{(k)}$ 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 modelling has been widely used in the ecological literature. 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 equation (5) to an additive form in i and k which need not be appropriate. (See Section 6.) The Ψ_ks are defined below (Section 4.2) using species level attributes and an overall intercept. They are viewed as an intercept specification for each of the species. Hence, there is no intercept in $\boldsymbol{\beta}_k$. The ρ_is denote spatially associated random effects. In other words we believe that the potential probability of presence or absence of species k at pixel i is also affected by its direct neighbours. We expect pixels which are close together to behave in a similar fashion in terms of their distribution of species. We employ an intrinsic conditional autoregressive model (Besag, 1974) to capture the spatial association in the ρ_i. In this regard, Hoeting *et al.* (2000) employed a single-stage autologistic model to describe spatial association between the $X_i^{(k)}$ across i directly. To accommodate the untractable calculation of the normalizing constant arising under this model, they employed a pseudolikelihood approximation.

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 equation (6), $\tilde{\mathbf{w}}_i$ are location characteristics and $\tilde{\mathbf{z}}_k$ are species attributes which are expected to affect $q_i^{(k)}$.

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

$$\begin{aligned} l(\boldsymbol{\theta};\mathbf{Y}) = \sum_{i=1}^{N}\sum_{k=1}^{K} & \min(1, Y_{i+}^{(k)})[Y_{i+}^{(k)}(\tilde{\mathbf{w}}_i'\tilde{\boldsymbol{\beta}}_k + \tilde{\mathbf{z}}_k\tilde{\boldsymbol{\gamma}}) - n_i \log\{1+\exp(\tilde{\mathbf{w}}_i'\tilde{\boldsymbol{\beta}}_k + \tilde{\mathbf{z}}_k\tilde{\boldsymbol{\gamma}})\} \\ & + \log\{(1-U_i)p_i^{(k)}\}] + \{1-\min(1, Y_{i+}^{(k)})\}[\log\{(1-q_i^{(k)})^{n_i}(1-U_i)p_i^{(k)} \\ & + 1-(1-U_i)p_i^{(k)}\}]. \end{aligned} \tag{7}$$

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 equation (7) in a formal way for the likelihood even if $n_i = 0$. There will just be no contribution from the ith pixel. However, from equation (5), we can learn about $p_i^{(k)}$, i.e. $\mathbf{w}_i$ is known, we learn about $\boldsymbol{\beta}_k$ and Ψ_k from other pixels and, owing to the spatial modelling for ρ_i, we can still learn about it from its neighbours through $\rho_i \mid \rho_j$, $j \neq i$. The special case where $U_i = 1$ implies that $n_i = 0$. Hence our modelling 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

We must 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 centred at 0 and with large variance.

As previously noted, the Ψ_ks 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 equation (6), each species effect Ψ_k can be explained by an overall intercept plus, say, L species level attributes. We then assign a normal prior distribution to μ centred at 0 with a large variance, and also a normal prior distribution to $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)'$ centred at 0 with a large variance. For the variance of Ψ_k, σ^2_ψ, we assign an inverse gamma prior with infinite variance. We 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, centring the parameterization as above provides more stable computation. (See, for example, Gelfand *et al.* (1996) and Papaspiliopoulos *et al.* (1996).)

The prior distribution of the spatial random effects is described through a nearest neighbour Markov random-field model (Besag, 1974). 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

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

where δ_i denotes the neighbours of cell i and w_{i+} denotes the total number of cells which are neighbours of i, $w_{ij} = 1$ if sites i and j share the same boundary and $w_{ij} = 0$ otherwise. For the conditional variance of the Gaussian Markov random field, σ^2_ρ, we also assign an inverse gamma prior with infinite variance.

Inference for the resulting posterior is done through simulation-based model fitting using Gibbs sampling (Gelfand and Smith, 1990) to obtain samples from the posterior distribution. In implementing the Gibbs sampling we need to specify all the posterior full conditional distributions of all unknown quantities in the model. The parameter μ, the intercept of the species random effect, has 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. For all the remaining parameters we can use the adaptive rejection Metropolis sampling within Gibbs sampling that was introduced by Gilks *et al.* (1995).

5. Inference with regard to biodiversity

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

We begin with the range of species. The common presentation of the range of species is based on the extent of occupancy and range of occupancy. For the observed $\{Y^{(k)}_{ij}\}$, 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 cannot be present. It also fails to recognize edge effects in that presence or absence need not have a *hard* edge but perhaps a *soft* edge characterized by a 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 the 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, i.e. 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 that we have proposed can only be interpreted with respect to the domain of study.

Another important feature is the richness of species. The *observed* richness of species in pixel i is $\Sigma_{k=1}^{K} \mathbf{1}(Y_{i+}^{(k)} > 0)$ for pixels where $n_i > 0$ and $1 - U_i > 0$. Again, this is a purely descriptive summary. Regression models can be used to explain these observed richness values by using environmental features and enable interpolation to unobserved sites. Work of this sort has been mentioned in Section 1 and enables refined prediction of the richness of species. See Guisan and Zimmerman (2000) in this regard. Under our model, the analogue for pixel i is the posterior distribution of $\Sigma_{k=1}^{K} X_i^{(k)}|\mathbf{Y}$. This posterior speaks to potential richness, i.e. in the absence of human intervention, it is the *number* of species that we would expect to find in pixel i. Converting to the distribution of $(1-U_i)\ \Sigma_{k=1}^{K} X_i^{(k)}|\mathbf{Y}$ modifies to transformed richness, i.e. the number of species that 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 $E(\Sigma\ X_i^{(k)}|\mathbf{Y}) = \Sigma\ E(p_i^{(k)}|\mathbf{Y})$ *versus* i; similarly a posterior transformed richness surface can be obtained. These can be displayed in a fashion that is similar to that proposed above for the range of species. It is important to note that, under our modelling, the richness of species can only be inferred within the domain of study and is only relative to the set of species which have been modelled.

Since traditional modelling of richness of species 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 and low for the other. Indeed, this is why we work with species level coefficients. Expressed in different terms, when similar richness of species is observed at two different locations, the set of species 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 quite a different set of species?

In our setting we can offer the same clarification. Since $\log\{p_i^{(k)}/(1-p_i^{(k)})\}$ strictly increases in $p_i^{(k)}$, suppose that we look at

$$\sum_{k=1}^{K} E\left[\log\left(\frac{p_i^{(k)}}{1-p_i^{(k)}}\right)\middle|\mathbf{Y}\right]$$

rather than $\Sigma\, E(p_i^{(k)}|\mathbf{Y})$. With regard to environmental characteristics, the former involves $\mathbf{w}_i' E(\Sigma_{k=1}^{K} \boldsymbol{\beta}^{(k)}|\mathbf{Y})$. We see that $\Sigma\, \boldsymbol{\beta}^{(k)}$ plays the role of the coefficient vector when modelling richness of species directly. Thus we can see that for say the lth component of $\Sigma\, \boldsymbol{\beta}^{(k)}$ it can be that, for some k, $\Sigma\, \beta_l^{(k)}$ is significantly positive whereas for other k it may be significantly negative. In aggregate, we need not find significance. (To work on the same scale as the $\beta_l^{(k)}$s we might use the posterior of $K^{-1}\, \Sigma_{k=1}^{K} \boldsymbol{\beta}^{(k)}|\mathbf{Y}$). We explore the issues of explaining richness in detail in the data analysis of Section 6.

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

Finally, related to the foregoing discussion, we consider the issue of the turnover of species, i.e. not only do we expect similar richness in neighbouring pixels but also that it arises from

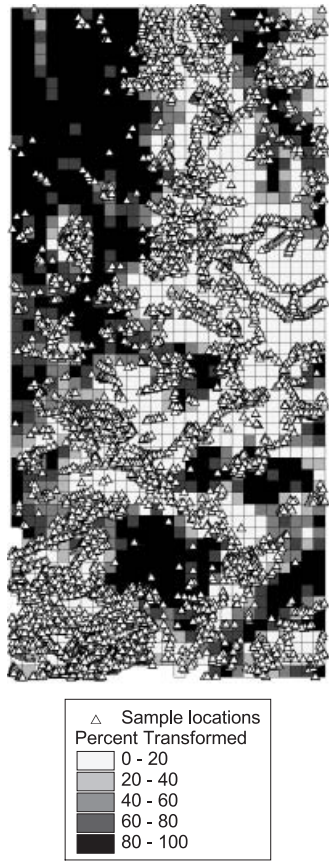

Fig. 1. Kogelberg–Hawequas subregion overlaid with the sampling locations

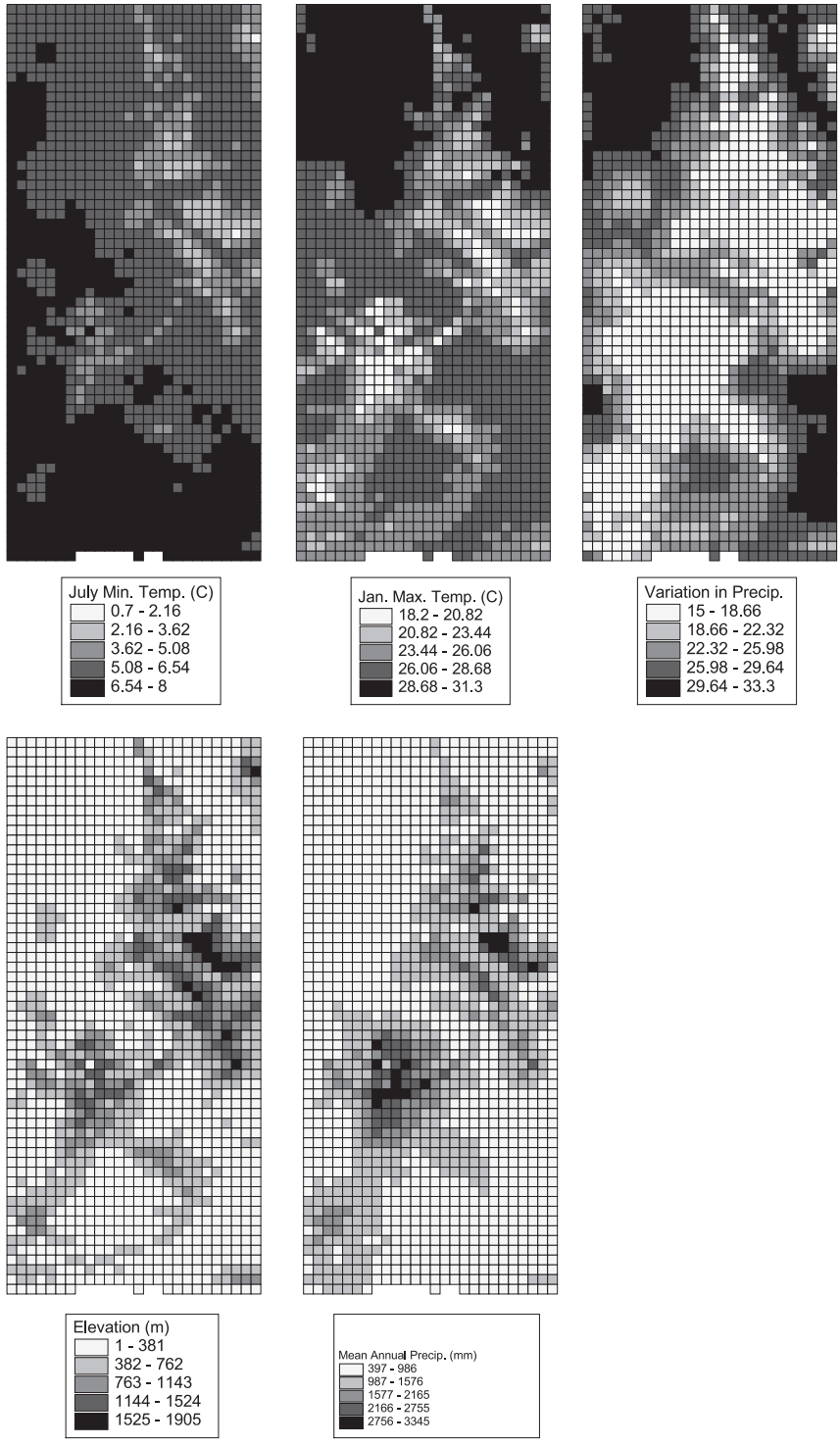

Fig. 2. Data layers of July minimum temperature, January maximum temperature, PPTCV, elevation and mean annual precipitation

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, as well, to 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 neighbourhood structure (first or second order), for each pixel i we propose to compute

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

For cell i, d_i yields an average similarity (first or second order) of cell i with its neighbours. 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 and where it is low.

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 CFR, occupying 1554 grid cells. Fig. 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). Cell level characteristics (the $\mathbf{w}_i$) include the July minimum temperature, January maximum temperature, intracell coefficient of variation in annual precipitation (PPTCV), altitude and mean annual precipitation. Grey scale maps of these data layers are supplied in Fig. 2. As can be observed, the

Table 1. List of species (in alphabetical order) with their attributes

Species	*Dispersal mechanism*	*Response to fire*	*Height (m)*	*Location population size*
Aulax pallasia Stapf	Wind (1)	Bole (1)	2.0	<1000
Aulax umbellata (Thunb.) R.Br.	Wind (1)	Killed (0)	2.0	>1000
Leucadendron corymbosum P.J. Bergius	Wind (1)	Killed (0)	1.5	<1000
Leucadendron daphnoides (Thunb.) Meisn.	Mammal (0)	Killed (0)	1.00	>1000
Leucadendron microcephalum (Grand.) Gand. & Schinz	Wind (1)	Killed (0)	1.25	>1000
Leucadendron salicifolium (Salisb.) I. Williams	Wind (1)	Killed (0)	2.00	<1000
Leucadendron salignum P.J. Bergius	Wind (1)	Bole (1)	0.50	>1000
Leucadendron sessile R.Br.	Mammal (0)	Killed (0)	1.0	>1000
Leucadendron spissifolium (Salisb. *ex* Knight) I. Williams	Wind (1)	Bole (1)	1.00	<1000
Leucadendron tinctum I. Williams	Mammal (0)	Killed (0)	0.75	<1000
Leucospermum bolusii Grand.	Ant (0)	Killed (0)	1.00	>1000
Leucospermum grandiflorum (Salisb.) R.Br.	Ant (0)	Killed (0)	1.5	<1000
Leucospermum oleifolium (P.J. Bergius) R. Br.	Ant (0)	Killed (0)	0.75	<1000
Mimetes arboreus Rourke	Ant (0)	Killed (0)	3.00	<1000
Mimetes 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
Protea nana (P.J. Bergius) Thunb.	Wind (1)	Killed (0)	1.00	<1000
Protea neriifolia R.Br.	Wind (1)	Killed (0)	2.50	>1000
Serruria elongata (P.J. Bergius) R.Br.	Ant (0)	Killed (0)	1.00	>1000
Serruria fasciflora Salisb. *ex* Knight	Ant (0)	Killed (0)	0.5	>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

July minimum temperature presents some high values in the south-west portion of the region, whereas the 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 centre portion of the subregion. This region presents some variability with respect to elevation, presenting higher elevation in its mid-east. Finally, the mean annual precipitation presents higher values towards the central portion of the subregion.

23 species were selected somewhat arbitrarily. They are listed alphabetically with 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 three locations (0.06%). Species level attributes (the $\mathbf{Z}_k$) include a dispersal

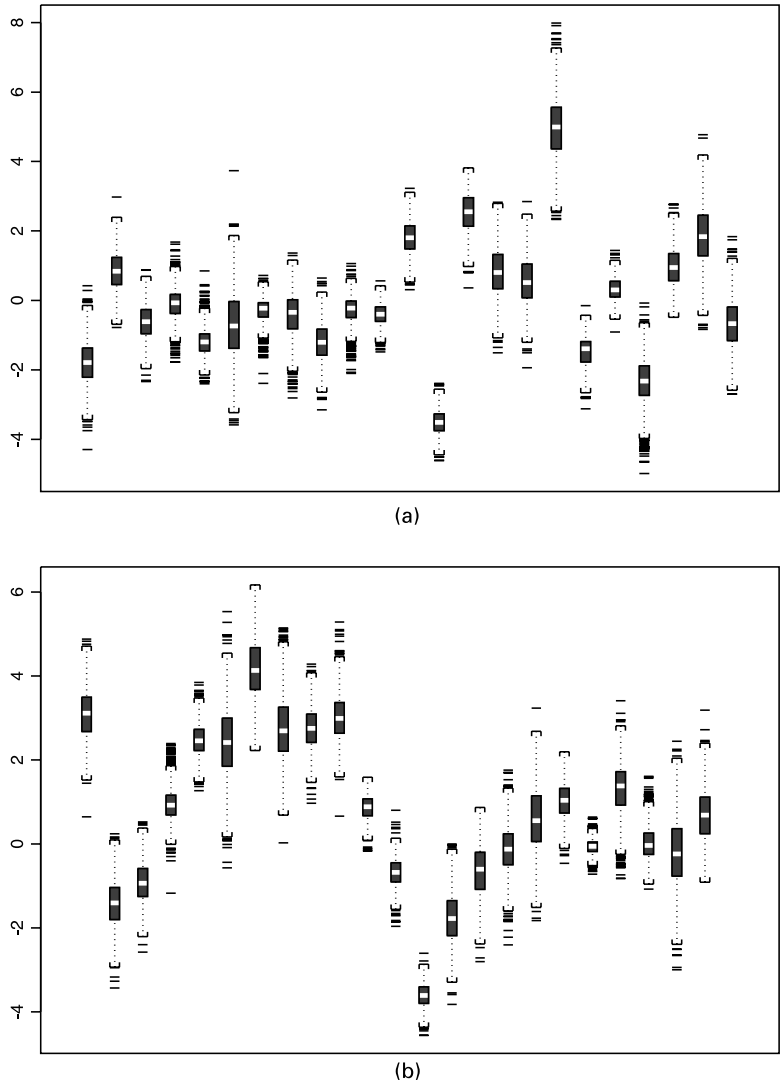

Fig. 3. Posterior summary of the coefficients of (a) January maximum temperature and (b) July minimum temperature for each of the 23 species (in alphabetical order as in Table 1)

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.

Illustratively, Fig. 3 shows posterior box plots for the $\beta_l^{(k)}$ associated with the January maximum temperature and July minimum temperature. It is clear that an increase in January maximum temperature decreases the probability of presence for most of the species. The July

Table 2. Posterior summary of the coefficients of the species level attributes (γs)

Covariate	*Mean*	*Values for the following percentiles:*		
		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 population size	1.14	−0.69	1.16	2.87

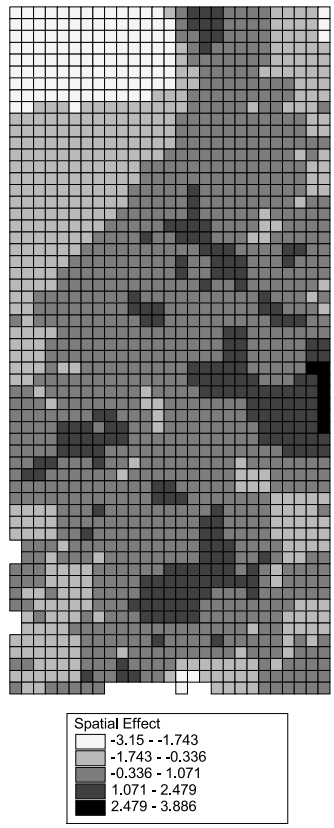

Fig. 4. Posterior mean of the spatial effects (ρ_is)

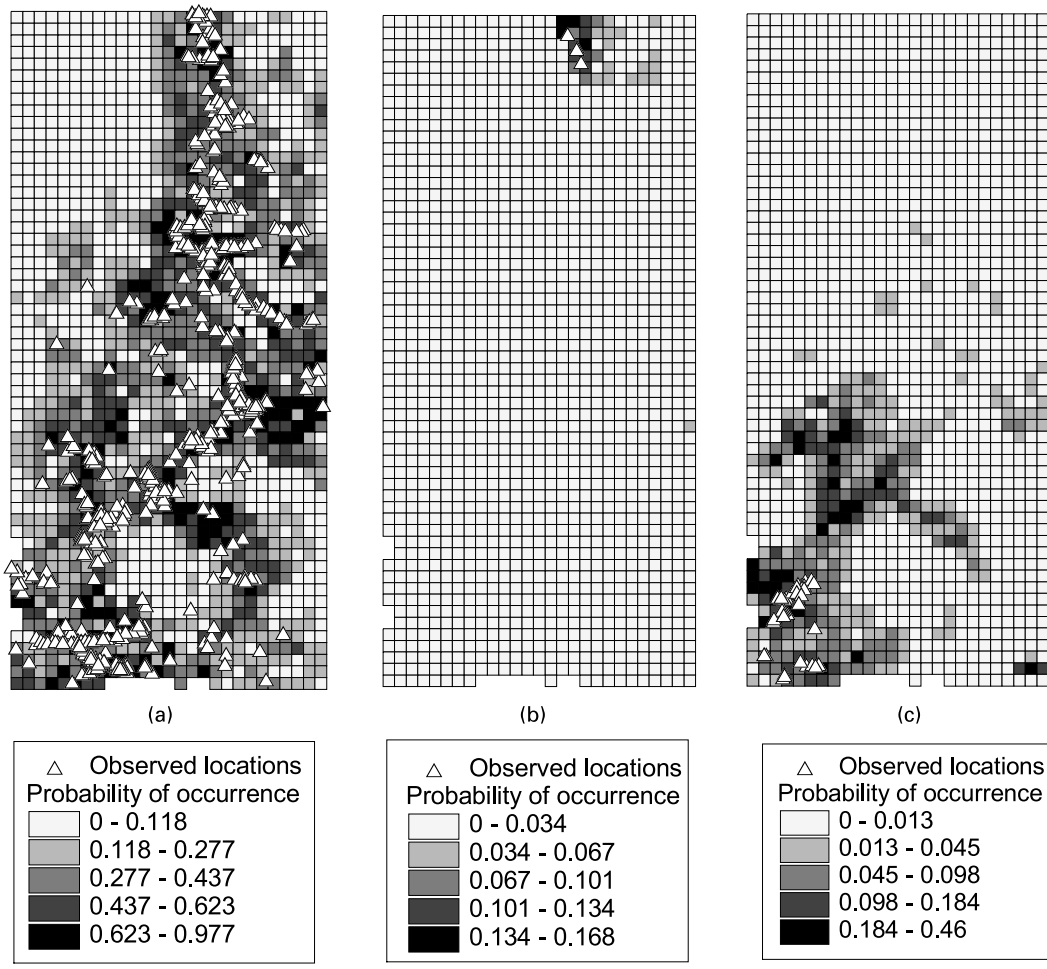

Fig. 5. Observed and adjusted ranges for (a) *Leucadendron salicifolium*, (b) *Sorocephalus imbricatus* and (c) *Mimetes arboreus*

minimum temperature reveals the importance of considering different coefficients for each of the species, as different species present considerably different behaviour in terms of July minimum temperature. For some species, an increase in the July minimum results in an increase in the chance of presence whereas, for others, the probability of presence decreases with higher July minimum temperature.

Table 2 provides a summary of the inference for the coefficients of the species level attributes (γs) implicitly within equation (5). Potential presence 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 towards larger local population size. It appears to be discouraged by increasing average plant height. Fig. 4 shows the spatial adjustment to the $p_i^{(k)}$ in equation (5) using the posterior means of the ρ_is. Spatial pattern, smoothed through the conditional autoregressive model, is evident. For instance, spatial effects are small in the north and west portion, and larger in the central east and south-east portion.

Next, we turn to the patterns of distributions and ranges of species described in the previous section. For the range of species we illustrate with three species, which are quite different from

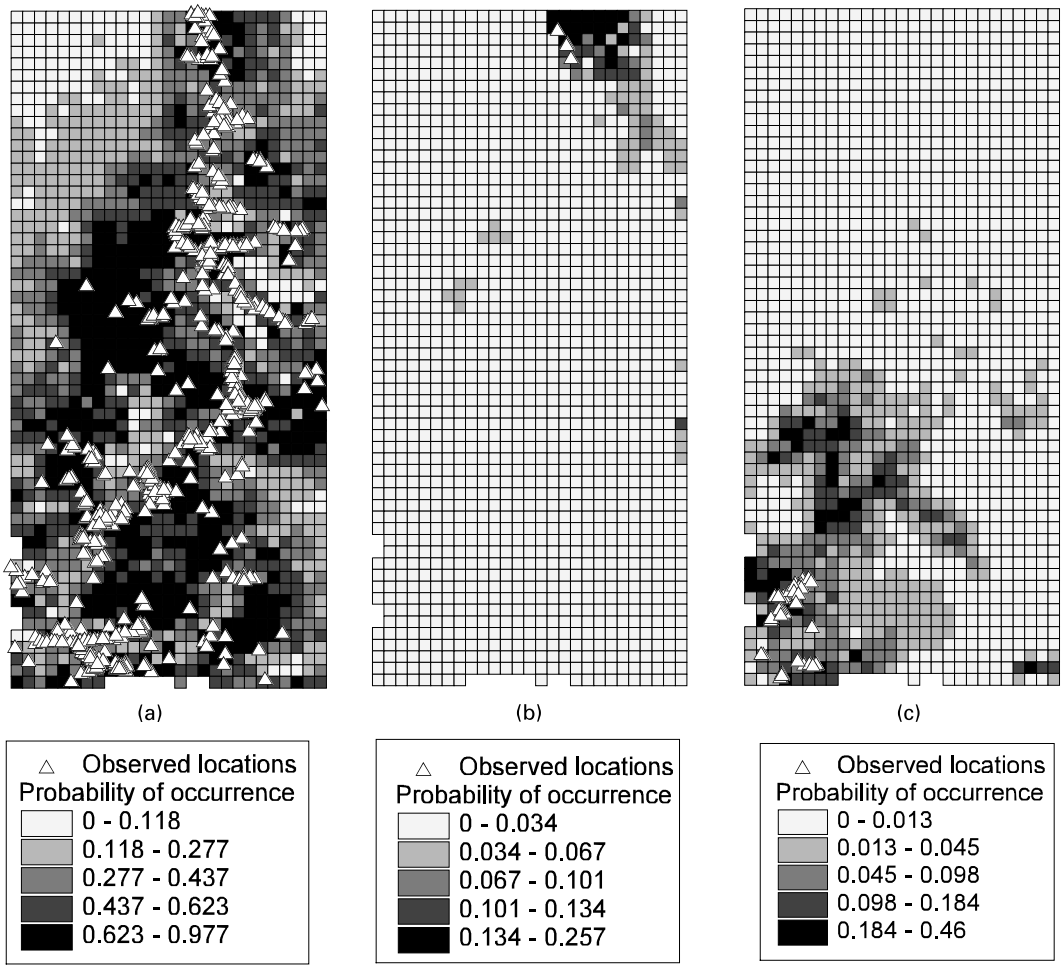

Fig. 6. Potential range for (a) *Leucadendron salicifolium*, (b) *Sorocephalus imbricatus* and (c) *Mimetes arboreus*

each other with regard to abundance and range, *Leucadendron 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 Fig. 5. Fig. 6 shows the potential ranges for these three species. They are larger than the corresponding ranges in Fig. 5 but a 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 Figs 5(a) and 5(b). Note that for the species in Fig. 5(c) 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.

Turning to richness of species, in Fig. 7 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 we compare the transformed richness with the observed richness, it is clear that the model can predict the richness quite well. Following the discussion in Section 5, with regard

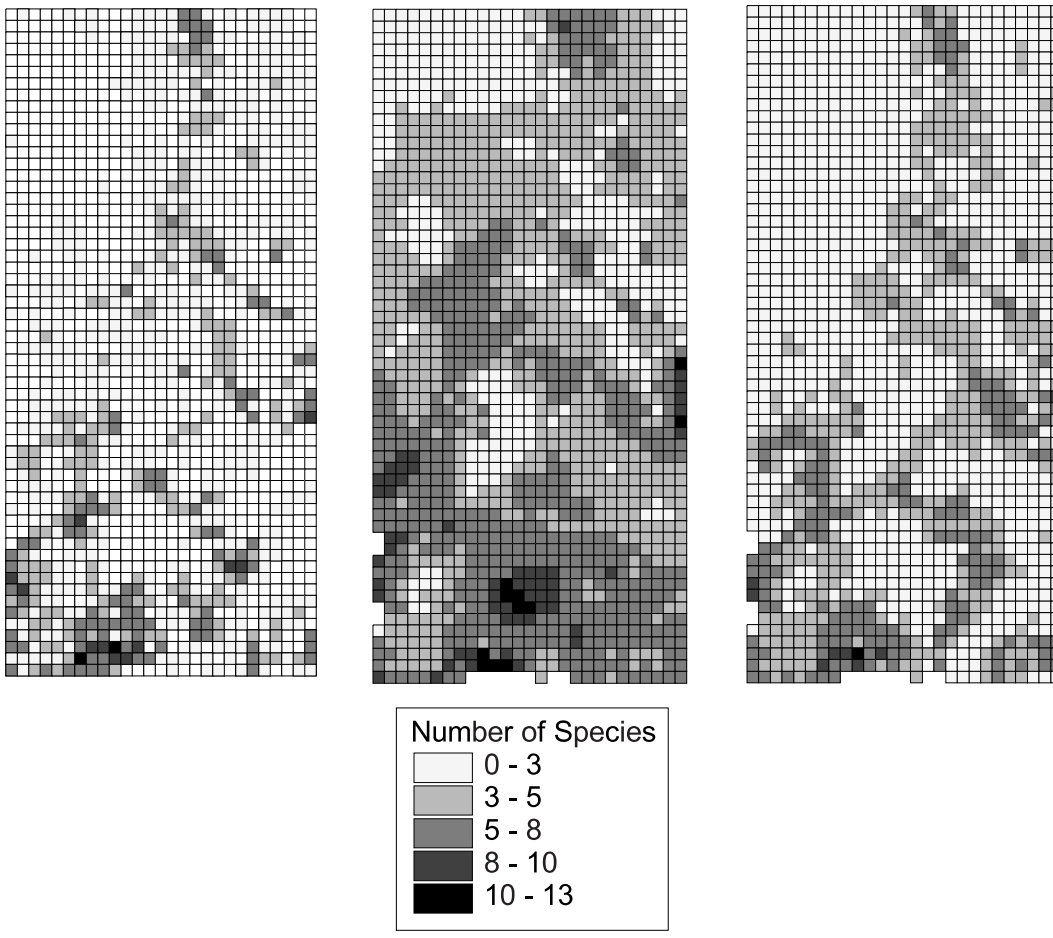

Fig. 7. Observed, potential and adjusted richness

to explaining richness, Table 3 summarizes the posterior distribution for the $\Sigma_{k=1}^{23} \beta_l^{(k)}$. Altitude is suggestively significant whereas January maximum temperature is not in explaining richness. An increase in July minimum temperature represents an increase in richness, whereas mean annual precipitation, PPTCV and elevation have a behaviour in the opposite direction.

Finally, to examine turnover, Fig. 8 summarizes the similarity of the $E(\mathbf{p}_i|\mathbf{Y})$ by using first- and second-order neighbours. 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. However, 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 model proposed marks a significant advance over previous efforts to model biogeographic patterns in distributions of species. The model form enables quantification

Table 3. Posterior summary of the area level attributes in terms of potential richness ($\Sigma_k \beta_l^{(k)}$)

Covariate	*Values for the following percentiles:*		
	2.5%	*50%*	*97.5%*
Mean annual precipitation	−26.73455	−19.43816	−12.70147
July minimum temperature	10.1223	16.58706	23.73233
January maximum temperature	−7.890144	−0.101283	6.829859
PPTCV	−13.45996	−7.281156	−1.342811
Elevation	−16.43027	−8.035193	0.4785387

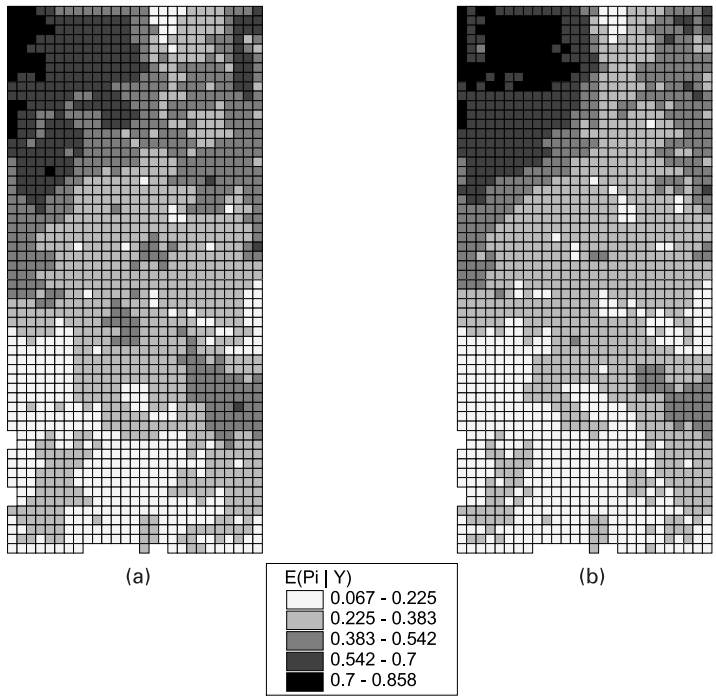

Fig. 8. Maps of the similarity of the $E(\mathbf{p}_i|\mathbf{Y})$ using (a) first- and (b) second-order neighbours

of uncertainty for all parameters. Moreover, the model proposed 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 model can handle variation in areal sampling intensity. Finally, predictions of presence or absence take into account spatial association, so 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 richness of species and the spatial

patterns in the turnover of species. 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 distribution of species as well as any declining probabilities of presence near distributional limits. Clearly this is a more realistic and practical means of specifying ranges of species than conventional solid polygons. This approach also provides testable predictions about the potential range of a species.

As with individual species range, the model proposed provides a rigorous way to predict the richness of species in a particular grid cell. Since our measure of richness is the sum of the probabilities of presence of individual species, it includes both an estimate and a quantification of uncertainty. The predicted turnover of species is similarly a summation of the differences of probability of presence for each species between a grid cell and its neighbourhood. 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 the edges of the range and patchiness within distributions and cannot directly suggest how individual species and site level characteristics may influence the rate of turnover.

Future work with this modelling approach would include more species and would incorporate other explanatory data layers. For example, geological information capturing fertility, texture and acidity of the soil would be very important in explaining the presence or absence of species. Another possibility would be to introduce phylogenetic information into the modelling to replace species attributes. The approach enables an assessment of the scale of resolution effects by fitting a given model at different resolutions. It also enables comparisons across regions with regard to biodiversity issues. Finally, the Bayesian approach for fitting such hierarchical models is clearly advantageous in enabling full inference. Moreover, there may not be another feasible approach for actually fitting the models that are proposed here.

The entire proposed hierarchical modelling and inference strategy can serve as a prototype for other biodiversity data analysis. For instance, in principle, the abundance of species could be studied replacing the binary $Y_{ij}^{(k)}$ with counts. However, the modelling will require some modification since we would want to associate abundance with an area rather than a point.

Acknowledgements

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 National Science Foundation grant DEB0089801 and by the National Center for Ecological Analysis and Synthesis. The authors thank Richard Cowling and Henri Laurie for valuable discussions. Finally, the authors gratefully acknowledge the insightful reports of the reviewers. In particular, their comments substantially improved the presentation in Section 4.

References

Aspinall, R. (1992) An inductive modeling procedure based on Bayes' Theorem for analysis of pattern in spatial data. *Int. J. Geogr. Inform. Syst.*, **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. *Photgram. Engng Remote Sens.*, **59**, 537–543.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *J. R. Statist. Soc.* B, **36**, 192–225.

Brzeziecki, B., Kienast, F. and Wildi, O. (1993) A simulated map of the potential natural forest vegetation of Switzerland. *J. Vegtn Sci.*, **4**, 499–508.

Colwell, R. K. and Lees, D. C. (2000) The mid-domain effect: geometric constraints on the geography of species richness. *Trends Ecol. Evoln*, **15**, 70–76.
Cressie, N. A. C. (1993) *Statistics for Spatial Data*, revised edn. New York: Wiley.
Currie, D. M. (1991) Energy and large-scale patterns of animal—and plant—species richness. *Am. Natur.*, **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*, 6th edn. London: Murray.
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. *Biodivers. Conservn*, **11**, 2309–2338.
Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996) Efficient parametrizations for generalized linear models. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 227–246. Oxford: Oxford University Press.
Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.
Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995) Adaptive rejection Metropolis sampling within Gibbs sampling. *Appl. Statist.*, **44**, 455–472.
Guisan, A., Edwards, Jr, T. C. and Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol. Modllng*, **157**, 89–100.
Guisan, A. and Zimmerman, N. E. (2000) Predictive habitat distribution models in ecology. *Ecol. Modllng*, **135**, 147–186.
Heikkinen, J. and Högmander, H. (1994) Fully Bayesian approach to image restoration with an application in biogeography. *Appl. Statist.*, **43**, 569–582.
Heikkinen, R. K. (1996) Predicting patterns of vascular plant species richness with composite variables: a meso-scale 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. *J. Agric. Biol. Environ. Statist.*, **5**, 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: Princeton University Press.
Huston, M. A. (1994) *Biological Diversity: the Coexistence of Species on Changing Landscapes*. Cambridge: Cambridge University Press.
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. *Ecol. Modllng*, **159**, 189–207.
Meyers, N., Mittermeier, R. A., Mittermeier, C. G., 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. *J. Biogeogr.*, **16**, 141–150.
Palmer, M. W. (1996) Variation in species richness: towards a unification of hypotheses. *Folia Geobot. Phytotax.*, **29**, 511–530.
Papaspiliopoulos, O., Roberts, G. O. and Skold, M. (2003) Noncentered parametrisations for hierarchical models and data augmentations. In *Bayesian Statistics 7* (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), pp. 307–326. Oxford: Oxford University Press.
Rebelo, A. G. (1991) *Protea Atlas Manual: Instruction Booklet to the Protea Atlas Project*. Cape Town: Protea Atlas Project.
Rebelo, A. G. (2001) *Proteas: a Field Guide to the Proteas of Southern Africa*, 2nd edn. Vlaeberg: Fernwood.
Rebelo, A. G. (2002a) The state of plants in the Cape Flora. In *The State of South Africa's Species* (eds G. H. Verdoorn and J. Le Roux), pp. 18–43. Parkview: Endangered Wildlife Trust.
Rebelo, A. G. (2002b) Are we really finished atlassing. *Protea Atlas Newslett.*, **51**. (Available from `http://protea.worldonline.co.za/pan51.htm`.)
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: Cambridge University Press.
Schultze, R. E. (1997) South African atlas of agrohydrology and climatology. *Technical Report TT82/96*. Water Research Commission, Pretoria.
Takhtajan, A. (1986) *Floristic Regions of the World*. Berkeley: University of California Press.