Bayesian Analysis (2006) **1**, Number 1, pp. 41–92

Explaining Species Distribution Patterns through Hierarchical Modeling

Alan E. Gelfand*, John A. Silander Jr†., Shanshan Wu‡, Andrew Latimer§, Paul O. Lewis¶, Anthony G. Rebelo‖ and Mark Holder**

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 analyze biodiversity patterns. Analysis for the region is developed at the spatial scale of one minute grid-cells ($\sim 37,000$ cells total for the region). We report on results for 40 species of a flowering plant family Proteaceae (of about 330 in the CFR) for a defined subregion.

Using a Bayesian framework, we develop a two stage, spatially explicit, hierarchical logistic regression. Stage one models the *suitability* or potential presence for each species at each cell, given species attributes along with grid cell (site-level) climate, precipitation, topography and geology data using species-level coefficients, and a spatial random effect. The second level of the hierarchy models, for each species, observed presence/absence at a sampling site through a conditional specification of the probability of presence at an arbitrary location in the grid cell given that the location is suitable. Because the atlas data are not evenly distributed across the landscape, grid cells contain variable numbers of sampling localities. Indeed, some grid cells are entirely unsampled; others have been transformed by human intervention (agriculture, urbanization) such that none of the species are there though some may have the potential to be present in the absence of disturbance. Thus the modeling 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.

In fact, a range of models can be examined incorporating different first and second stage specifications. This necessitates model comparison in a misaligned multilevel setting. All models are fitted using MCMC methods. A "best" model is selected. Parameter summaries offer considerable insight. In addition, results

*Institute of Statistics and Decision Sciences, Duke University, Durham, NC., http://www.stat.duke.edu/~alan/

†Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, http://www.eeb.uconn.edu/faculty/silander/silander.htm

‡ING Clarion, New York, http://merlot.stat.uconn.edu/~shanshan/

§Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, http://www.eeb.uconn.edu/grads/latimer/

¶Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, http://www.eeb.uconn.edu/faculty/plewis/plewis.htm

‖National Botanic Institute, Kirstenbosch, South Africa, zoot@nbict.nbi.ac.za

**School of Computational Science and Information Technology (CSIT), Florida State University, Tallahassee, FL, http://www.csit.fsu.edu/~mholder/

© 2006 International Society for Bayesian Analysis

are mapped as the model-estimated potential presence for each species across the domain. This probability surface 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. Other biodiversity measures emerge as model unknowns. A considerable range of inference is available. We illustrate with only a portion of the analyses we have conducted, noting that these initial results describe biogeographical patterns over the modeled region remarkably well.

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

1 Introduction

Ecologists increasingly use species distribution models to address theoretical and practical issues including predicting the response of species to climate change (Midgley et al. 2002), identifying and managing conservation areas (Austin and Meyers 1996a), and seeking evidence of competition among species (Leathwick 2002). In all of these contexts, the core problem is to use information about where a species occurs (and where it does not) and about the associated environment to predict how likely the species is to be present in unsampled locations.

The significant contribution to the ecology literature of this work is to clarify the value in modeling at the sampling site/species level. Through a hierarchical model, we are able to infer a spatial probability surface for potential presence as well as for transformed presence, along with associated uncertainty. Furthermore, through appropriate aggregation, inference regarding richness, prevalence, diversity, and turnover can be addressed. The broad scope of inference opportunity is evident.

The key modeling issues center on careful articulation of the definition of events and associated probabilities, the misalignment between the sampling for presence/absence (at sampling sites) and the available environmental data layers (at grid cells of roughly three square kilometers), the sparsity of observations in terms of the entire landscape (with uneven sampling intensity including "holes"), the occurrence of considerable human intervention (transformation) with regard to land use across this landscape, the need for explicit spatial modeling, and difficulties arising in model comparison.

The value of this work for the applied statistician is to raise prototypical issues in determining the scale for building a model; to build regression models with misaligned data layers; to recognize that, with regard to quantities of interest (in our case, biodiversity measures), rather than creating ad hoc descriptive statistics for them, they can be viewed as functions of the parameters in the model and thus, within a Bayesian framework, can be inferred about through their posterior distributions; that through such modeling we can assess uncertainty exactly rather than relying on (possibly inappropriate) asymptotics.

To begin, we might ask 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 few 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 in a site (rows) by species (columns) presence/absence matrix. A classification - then - modeling strategy gathers either the sites into groups containing similar species ("community assemblages") or the species into groups occurring at similar sites ("habitat assemblages"). Regression modeling follows, using environmental factors for the communities or species attributes for the assemblages. See, e.g., Ferrier et al. (2002a) 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 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 (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, as noted above, 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 in the modeling of suitability or potential presence rather than at the data stage. This contrasts with, e.g., Hoeting et al. (2000) as well as the GRASP approach.

Hence, we develop a fully model-based multilevel approach to illuminate biodiversity concepts such as species range, richness and turnover. 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 potential presence 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 sparse 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. agriculture, urban lands, dense alien plant infestations. This implies that there is a notion of *potential* presence as well as *transformed* (or adjusted) presence. These notions will be formally 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, essentially, at the point level.

Therefore, we envision a multilevel model. That is, we model potential presence, transformed presence (both *available* and *suitable*) given potential absence and observed presence/absence given suitability and availablility. 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 features to convey species range, to capture species richness, to explain species diversity, to study species turnover across the domain.

Section 2 provides the description of the dataset used to address our problem. Section

3 provides a brief review of spatial modeling in ecology. Section 4 develops our model along with prior and fitting details. Section 5 takes up the model choice questions. In Section 6 we propose a variety of useful biodiversity measures and displays. Section 7 offers a portion of the extensive analysis we have carried out. Section 8 extends our analysis to illustrate an approach to modeling inter-species dependence based upon evolutionary considerations. Section 9 provides summary and discussion.

2 The Cape Floristic Kingdom; Data Description

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). (See Figure A of the supplemental material.) 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 modeling. 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.

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 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 suitable for modeling biogeographic patterns. Data were collected at "record localities": relatively uniform, geo-referenced areas typically 50 to 100 m 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.

We used the following collection of environmental explanation variables: elevation (ELEV) and roughness of terrain (ROUGH). These are the mean and range for elevation measurements recorded within a grid cell, respectively. They provide the topographical information for the cell. These and the climate data we employ here were obtained from the South African Atlas of Hydrology and Climatology (Schultze 1997) and either downloaded or obtained on CD's 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. Other layers can be constructed from these. We used the following as explanatory variables: mean annual precipitation(MAP), inter-annual coefficient of variation in precipitation(PPTCV), July (winter) minimum temperature(JULMIN), January (summer) maximum temperature(JANMAX), potential evapotranspiration (POTEVT), frost duration (FROST), heat units (HEATU), rainfall concentration(PPTCON), summer soil moisture days (SUMSMD), winter soil moisture days(WINSMD) and an enhanced vegetation index(EVI). The remaining set of explanatory variables attempt to capture the geology associated with each grid cell. These were recorded as soil attribute categories from 1:200,000 scale digitized geological maps obtained from the South African Council for Geosciences. These are soil fertility, supplied in four ordinal classifications (FERT1 - FERT4) denoting increasing order of fertility; soil texture, again in four ordinal classifications (TEXT1 - TEXT 4) denoting transition from fine to coarse; soil pH in three ordinal classifications (PH1 - PH3) from low to high alkalinity. In fact, a grid cell does not consist entirely of one soil fertility, texture or pH. Rather, for each cell, we have a proportion of each classification. These proportions arise from the overlay of a vector map of geology onto our rasterized grid cell map. Since, in each cell, for each of these three ordinal variables, the sum of the proportions is 1, we can omit a classification for each variable. For fertility we chose "high" (FERT4), for texture we chose "medium fine" (TEXT 2) and for pH we chose "neutral" (PH2).

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 41m east and 107m north from Cape Hangklip. The region is shown in Figure 1. Further, we restricted the analysis to 40 species of Proteaceae out of roughly 150 found within this rectangular area. For each species we scored the following traits (attributes): height (continuous), local population size (ordinal), ability to resprout after fire (categorical), pollination mode (categorical) and dispersal mode (categorical). A list of the 40 species and their attributes is given in Table A of the supplementary material.

Transformed areas (by agriculture, afforestation, alien plants and urbanization) were obtained as a GIS data layer from R. Cowling (private communication). Approximately 1/3 of the Cape has been transformed, mainly in the lowlands on more fertile soils where rainfall is adequate (Rouget et al. 2003). Most of the transformation outside of these areas, on the infertile mountains, is due to dense alien invader species, which are currently a major threat to Fynbos vegetation and, in particular, to the Proteaceae.

3 Review of spatial modeling in ecology

At global and continental scales, ecologists have explored biodiversity (generally species density) as the phenomenon of interest and proceed by constructing diverse hypotheses that relate it to various explanatory variables. These hypotheses are then formalized by treating species richness as the response variable and exploring its relationship with one or more explanatory variables via correlation or regression. While many explanatory variables have been proposed, the core set can be approximately summarized as: 1) productivity, often expressed through simple proxies such as water/energy measures (e.g. potential evapo-transpiration, or mean temperature and precipitation), chlorophyll density (NDVI, akin to EVI mentioned above), etc., or modeled mechanistically (Woodward et al. 1995) or empirically via sets of observed parameters; 2) heterogeneity, including climatic and topographic heterogeneity; 3) disturbance history on some spatial and temporal scale; and 4) geographic constraints. Using regional or broader-scale databases of species diversity and coarse-scale geographic information, these models provide a rough means of assessing the validity and generality of the alternative proposed hypotheses. Recent work has recognized that the contribution of various explanatory factors may vary depending on spatial scale. For example, Rahbek and Graves (2001) showed that the contribution of areal effects, productivity, spatial heterogeneity and other variables to explaining avian species diversity in South America varied with the resolution of the spatial scale modeled.

At this point no pure "laws" appear to have yet been distilled, although some authors point to generalities that have emerged. There seems to be support for a combination of water and energy availability as providing significant influence on diversity at both small and large spatial scales. This finding is consistent with phenomenological predictions about diversity levels: since heat and radiant energy and water are limiting factors on net photosynthesis in most environments, their availability is a reasonable proxy for productivity, and all things being equal, greater productivity ought to support more diversity (Currie 1991; Whittaker et al. 2001). Spatial heterogeneity also emerges as an important correlate of diversity across scales, generally enhancing diversity as one would expect from classical competition theory (MacArthur et al. 1966). Temporal heterogeneity, with a long pedigree (Wallace 1895) has been formalized as an explanatory variable; on local scales it has been found to enhance diversity up to a point (i.e. intermediate disturbance) but this is more parsimoniously considered as a contributor to spatial heterogeneity. On larger scales temporal variability has been shown to have a negative relationship with diversity by inducing migration and local extinction (Dynesius and Jansson 2000; Jansson and Dynesius 2002). Other explanatory variables, including geometric constraints, have found some support in certain conditions (Colwell and Lees 2000).

3.1 Spatial prediction

Development of ecological predictive models for smaller-scale regions has exploded as increased computing power has made both GIS tools and statistical model implementa-

tions widely available. Guisan and Zimmerman provide a useful overview of the major modeling techniques and their applications (Guisan and Zimmerman 2000). The response variable may be presence/absence or abundance of individual species or species richness (Ferrier et al. 2002b).

Methodologically, predictive modeling at local and regional scales has been dominated by generalized linear models (GLMs), which provide a natural way of relating binary presence/absence data or abundance data to site-level explanatory variables (Austin et al. 1990), and generalized additive models (GAMs), which enable fitting of locally smoothed, nonlinear response forms (Yee and Mitchell 1991). GAMs generally tend to fit data better than GLMs, since they employ additional parameters to enable the response variables to assume highly nonlinear and even multimodal relationships with the data (Guisan et al. 2002). They also provide a qualitative picture of how species respond to explanatory variables (Austin and Meyers 1996b). The price of the flexibility of GAMs, however, is loss of simplicity in interpretation and approximation in quantifying uncertainty (Heegard 2002).

In search of more powerful prediction, researchers have tried many of the available tools of computer science, including neural networks, expert classification systems, discriminant analysis and artificial intelligence (AI) methods (Manel et al. 1999). For some problems, such approaches may provide important advantages, including accommodation of complex interactions among predictors, but they are prone to overfitting and they remain difficult to interpret since they do not present relationships between predictions and explanatory variables in a transparent way (Lehmann et al. 2002). The mulivariate techniques familiar to many ecologists, such as principal components analysis, correspondence analysis, canonical correspondence analysis, and multidimensional scaling offer few advantages relative to other methods available (Guisan and Zimmerman, 2000).

Currently, GAMs appear to be the most popular choice for spatial ecological modeling, since they are flexible and can easily be implemented using S-Plus or other statistical software (Venables and Ripley 1999). Just a few of the many recent examples of the use of GAM regressions in predictive spatial modeling of species distributions include Eucalyptus tree distributions (Austin and Meyers 1996b), Australian vascular plants, vertebrates, and ground-dwelling arthropods (Ferrier et al. 2002b), aquatic plants in Switzerland (Lehmann 1998), and New Zealand fern species (Zaniewski et al. 2002). Extensions of GAMs have also been developed, including the GRASP statistical package (http://www.cscf.ch/grasp), that use spatial association in the explanatory variables to make explicitly spatial predictions of individual species distributions and richness patterns (Lehmann et al. 2002). A significant constraint on GAMs in ecological contexts arises when one is faced with modeling a large ensemble of explanatory variables. This is further aggravated when one is faced with simultaneous prediction of multiple species.

3.2 Spatial dependence

The problem of spatial autocorrelation in data, including both species survey data and environmental data, is frequently ignored in ecological predictive modeling (Lehmann, et al. 2002). The expected dependence among spatially proximate data points means that formal statistical inference based on non-spatial models (e.g., confidence intervals for parameters, niche delineations for species) may be faulty. Even if prediction is the only goal, to ignore spatial relationships is to throw out information, leading to reduced predictive power, and model interpretability (Wikle and Royle 2002). An explicitly spatial model enables analysis of which part of residual variation is associated with spatial structure and provides an opportunity to interpret that spatial structure in terms of organismal or environmental characteristics not included in the model (Wikle 2002; Gelfand et al. 2003).

Modelers have adopted various ad-hoc approaches to spatial dependence, such as including latitude and longitude terms in the model (simple trend surface specification), thereby removing north-south and east-west trends in the residuals. While this may improve predictive performance, it is unsatisfactory because it is not sufficiently flexible and need not fully correct the problem of dependence. Similarly, breaking up the modeled region into sub-regions, may diagnose a problem with the model form, but need not remedy it (Osborne and Suarez-Seoane 2000).

Spatial association has been addressed explicitly by introducing an additional explanatory variable that reflects predictions based upon neighboring cells.. The resulting models use the predictions of a non-spatial model to generate a new explanatory variable in which each cell is assigned a weighted average of the predicted values of neighboring cells (Augustin et al. 1996; Lehmann et al. 2002). The models are then fitted again with the new variable included, and the spatial variable and model are iteratively updated until convergence. This method of introducing spatial association at the data stage of the model provides a picture of spatial relationships in the data, and produces predictions that better match observed levels of clumping (Augustin, *et al.*, 1996). Such iterative model fitting makes it difficult to attach variability to predictions. Moreover, formal conditional modeling to explain response in a particular cell given the responses of neighboring cells is available and provides the spatially explicit specification we propose below.

3.3 Bayesian hierarchical models

Bayesian modeling has been rarely applied to ecological spatial prediction. A few early applications have used Bayes' Theorem to combine relationships between observed data and individual predictive factors with prior probabilities of presence to produce probability surfaces for species (Aspinall 1992; Aspinall and Veitch 1993; Royle et al. 2002) or vegetation types (Fischer 1990). Since these approaches use a contingency table approach and carry over only point estimates from the data stage to the generation of predictions, they are not directly comparable to hierarchical regression models. Only very recently have Wikle and collaborators presented examples of full Bayesian hierarchical

modeling applied to individual plant or bird species(Wikle 2002, 2003; Wikle and Royle 2002). See also Clark et al. (2003) in this regard.

A Bayesian hierarchical model also allows the introduction of spatial dependence naturally into the model through random effects that capture spatial association not contained in the other covariates. Through marginalization, the spatial random effects are incorporated directly into the model likelihood, and are fitted simultaneously with the other model parameters. They are introduced into the mean (on a transformed scale) and, controlling for the other covariates in the mean, encourage mean behavior to be similar when cells are close to each other. Like random effects in general, they soak up the lack of explanation of the fixed component of the mean but in a spatial fashion. Thus, they account for omitted or unmeasured explanatory variables having spatial content. Like other parameters, the spatial random effects have fully specified probability distributions, providing information about both their magnitude and uncertainty. Their effect in explaining potential presence is explicitly specified, so their contribution to the model and to prediction may be rigorously investigated. Random effects may be introduced in different ways, e.g., through a conditional auto-regressive (CAR) model as implemented here (Besag et al., 1974) or via a matrix of spectral functions (Hooten et al. 2003). Finally, through the implicit dependence structure, spatial modeling for random effects allows learning about their contribution even for cells where there has been no sampling, accommodating gaps in sampling and irregular intensity in sampling.

Lastly, hierarchical modeling enables us to precisely capture the sampling scale used in the data collection while introducing latent dependent variables that reflect a notion of presence at a different spatial scale. This is critical in our case due to the misalignment between the point-based species sampling data and the raster-based GIS data layers which provide our environmental covariates.

4 Model Development, Prior Specification and Fitting

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. 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 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. Similarly, we work with the 1554 minute by minute areal units (pixels) in our study region (Figure 1). In this subregion the pixels are rectangular, approximately 1.85 km $\times$ 1.55 km. If we were to formalize potential presence 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 modeled using a latent binary process.

Suppose we let $X_i^{(k)}$ denote the event that a randomly selected location in cell i is suitable(1) or unsuitable(0) for species k and set $P(X_i^{(k)} = 1) = p_i^{(k)}$. $p_i^{(k)}$ is naturally conceptualized using a binary process. Let $\lambda^{(k)}(\mathbf{s})$ be a binary process over the region indicating the *suitability* (1) or not (0) of location $\mathbf{s}$ for species k 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. From (1), the interpretation is that the more locations in cell i where $\lambda^{(k)}(\mathbf{s}) = 1$, the more suitable cell i is for species k, i.e., the greater the chance of potential presence in cell i. The collection of $p_i^{(k)}$'s over i can be seen as representing the potential distribution of species k.

Let $V_i^{(k)}$ denote the event that a randomly selected location in cell i is suitable for species k in the presence of transormation of the landscape. 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 $\mathbf{s}$ is suitable under transformation, i.e., we need both suitability and availability. Therefore,

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

If we make the assumption that, for each pixel, availability is uncorrelated with suitability, then (2) simplifies 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} pixel 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 $q_i^{(k)}$, i.e., for a randomly selected location in cell i, $q_i^{(k)}$ is the probability of species k being present given the location is both suitable and available. 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.,

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

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 $0 \le q_i^{(k)} \le 1$. But also, $\tilde{\lambda}^{(k)}(\mathbf{s}) = 1$ implies $T(\mathbf{s}) = 0$, i.e., presence implies availability. So the numerator simplifies to $\int_{\text{cell i}} \mathbf{1}(\tilde{\lambda}^{(k)}(\mathbf{s}) = 1)\, d\mathbf{s}$, which, divided by $|A_i|$ is the expected probability of presence at a randomly selected location in cell 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". So, in modeling for the $p_i^{(k)}$ and $q_i^{(k)}$, we look for ecological variables or species attributes that are expected to affect the "number" of $\lambda^{(k)}(\mathbf{s})$ or $\tilde{\lambda}^{(k)}(\mathbf{s})$ turned "on" in unit i. 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 ($n_i \ge 1$) there will be a contribution to the likelihood and, in fact, we can marginalize over $V_i^{(k)}$ to give, for $y > 0$, $P(Y_{i+}^{(k)} = y) = \binom{n_i}{y} \left(q_i^{(k)}\right)^y \left(1 - q_i^{(k)}\right)^{n_i - y} (1 - U_i)p_i^{(k)}$, 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 explicit modeling for $p_i^{(k)}$ 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 grid cell level characteristics, and the $\boldsymbol{\beta}_k$'s are species level coefficients associated with the grid cell level covariates. Therefore, the model allows the flexibility of each species having a different coefficient for each grid cell 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. (See the discussion in Section 7 below.) The Ψ_k's are defined below (Section 4.2) using species level attributes and an overall intercept. The Ψ_k's are viewed as a random intercept specification for each of the species. Hence, there is no intercept in $\boldsymbol{\beta}_k$. 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. [1] We employ an intrinsic 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 intractable calculation of the normalizing constant arising under this model, they employ a pseudo-likelihood 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 (6), $\tilde{\mathbf{w}}_i$ are location characteristics and $\tilde{\mathbf{z}_k}$ are species attributes which are anticipated to affect $q_i^{(k)}$. In fact if we are modeling the joint distribution of the $Y_{ij}^{(k)}$ 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.

We note that model choice (Section 5) for us focuses entirely on (5). We consider inclusion or exclusion of Ψ_k (exclusion means no species attributes are included, just a species level random effect) as well as inclusion or exclusion of the ρ_i. It also addresses the nature of $\boldsymbol{w_i}$ partitioning it into three groups of variables: (i) topography (ii) climate and (iii) geology. Here inclusion or exclusion is with regard to the entire group. Ultimately, we wind up selecting the model which retains all components of (5) though retention of the Ψ_k is borderline. Indeed, Figure 2 provides a graphical model encompassing our full hierarchical specification. It is noteworthy that previous work in the literature using logistic regression modeling for species presence/absence data employs only the G,C, and T nodes in the figure to explain $Y_{ij}^{(k)}$ or perhaps richness, $\sum_k Y_{ij}^{(k)}$. Thus the figure offers another way of revealing the difference between our contribution and extant work; again, the X's, V's, and ρ's are not observed.

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

[1] Modeling of species random effects and spatial random effects need not be additive as in (5). Forms involving cell-species interaction can be introduced but are not presented here.

for $\mathbf{Y} = \{Y_{i+}^{(k)}\}$ as

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

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 (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

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 $\tilde{\boldsymbol{\beta}}_k$ and $\tilde{\boldsymbol{\gamma}}_k$. 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_ψ^2, the Ψ_k's are independent and identically distributed following a normal distribution with mean $\mu + \mathbf{z}_k' \boldsymbol{\gamma}$ and common variance σ_ψ^2. In other words, analogous to (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 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 σ_ψ^2, 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, centering the parametrization as above provides more stable computation (See, e.g., Gelfand et al. 1996; Papaspiliopoulos et al. 2003).

We are now left to assign the prior distribution of the spatial random effects. We presume there exists local spatially structured variation in the probability of 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 models, 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, ∂i 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

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

where w_{i+} denotes the total number of cells which are neighbors of i, $w_{ij} = 1$ if sites i and j share the same boundary and 0 otherwise. For the conditional variance of the Gaussian Markov random field, σ_ρ^2, 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) in order to obtain samples from the posterior distribution. In implementing the Gibbs sampling one needs to obtain the full conditional distributions of all unknown quantities in the model. Were we to condition on the $X_i^{(k)}$, the full conditional distributions for the β_k, γ_l, ρ_i, $\tilde{\beta}_k$ and $\tilde{\gamma}_l$ would be log concave and we could use the adaptive rejection method of Gilks and Wild (1992) to update these parameters. However, this would necessitate introducing and updating the entire set of $X_i^{(k)}$'s. To work with the marginalized likelihood in (7), we can use adaptive rejection Metropolis sampling within Gibbs introduced by Gilks et al. (1995). 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.

5 Model Choice

As noted in section 4, model choice is investigated with respect to the specification for $p_i^{(k)}$ in (5). Following the discussion there and using the notation of Figure 2, we investigate models which include or exclude the geological variables (G), the topographical variables (T), the climate and precipitation variables (C), the spatial random effects (ρ) and the species attributes (A). The only species attribute which was significant in the model for $q_i^{(k)}$, expression (6), was local population size. It was retained in all of the above choices.

Model selection here is not completely straightforward. From (7), the marginal likelihood will be difficult to compute. In fact, with the improper CAR prior on the ρ_i's, implemented by centering at the end of each iteration (following Besag et al. (1995) to ensure a proper posterior), such marginalization is impossible. Moreover, the $Y_{ij}^{(k)}$ are Bernoulli trails with $P(Y_{ij}^{(k)} = 1|\theta) = (1 - U_i)p_i^{(k)}q_i^{(k)}$ and $P(Y_{ij}^{(k)} = 0|\theta) = 1 - (1 - U_i)p_i^{(k)}q_i^{(k)}$. Evidently, there is no "canonical" parametrization for $P(Y_{ij}^{(k)} = 1|\theta)$. Also, $P(Y_{ij}^{(k)} = 1|\theta) \leq (1 - U_i)$ so it is difficult to introduce latent $Z_{ij}^{(k)}$ such that, say

$Y_{ij}^{(k)} = 1$ if and only if $Z_{ij}^{(k)} > 0$. The $Z_{ij}^{(k)}$ will have a sub-distribution for many i's.

In the end, we adopted the computationally convenient DIC criterion (Spiegelhalter et al. 2002). This criterion is sensitive to choice of parametrization (Section 8, Spiegelhalter et al. (2002)). We considered two parametrizations. One treats θ as the *natural* set of model parameters as in (7). The other treats the $P(Y_{ij}^{(k)} = 1)$ as the set of parameters. Other choices are possible including treating logit $P(Y_{ij}^{(k)} = 1)$ as the set of parameters, treating the set of $p_i^{(k)}$ and the set of $q_i^{(k)}$ as parameters or even treating logit $p_i^{(k)}$ and logit $q_i^{(k)}$ as parameters. All will give different answers for p_D, the effective degrees of freedom, but, at least for the two choices we tried, the DIC's were similar and the ordering of the models was the same (smaller is better).

Table 1 provides a summary of the model comparison using DIC. Of the single variable group models, C is clearly the best. Comparison of the full model with the model G, C, T, ρ shows that the species attributes add the least, once the other sets are included. Under either criterion, the full model emerges as best. We confine ourselves to analyses under this model for the remainder of this paper.

6 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 and displays 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 choropleth 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.

Species' prevalence is a familiar notion. For our data, the raw prevalence for species k is the number or (dividing by 1554) the raw proportion of grid cells in which the species was observed. In the absence of a model it is hard to attach uncertainty to such a statistic. Moreover, not all cells have been sampled and those that are sampled are not sampled with the same intensity. In our setting, the model based analogue is $\sum_i p_i^{(k)}/1554$ (or perhaps, $\sum_i (1 - U_i) p_i^{(k)}/1554$). A plot of posterior prevalence distribution summaries (say point and 95% interval estimate) versus species will be referred to as a prevalence plot. We can overlay the observed raw proportions on this plot. Note that, in aggregating across i, these plots are not spatial.

Another important feature is species richness. The *observed* species richness in pixel i is $\sum_{k=1}^{K} \mathbf{1}(Y_{i+}^{(k)} > 0)$ for pixels where $n_i > 0$ and $1 - U_i > 0$. Again, this is a purely descriptive summary. Regression models have been used to explain these observed richness values using environmental features and enable interpolation to unobserved sites. 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. Again, 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 features 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 the motivation for modeling 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 some clarification. Since $\log \frac{p_i^{(k)}}{1-p_i^{(k)}}$ strictly increases in $p_i^{(k)}$, suppose we look at $\sum_{k=1}^K E\left(\log\left(\frac{p_i^{(k)}}{1-p_i^{(k)}}\right)|\mathbf{Y}\right)$ 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, $\beta_l^{(k)}$ 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 as the $\boldsymbol{\beta}_l^{(k)}$'s, we might use the posterior of $K^{-1}\sum_{k=1}^K \beta^{(k)}$).

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.

There is a considerable literature on diversity measures. See the summary discussion in Kempton (2002). For illustration we work with a Shannon-Weiner form of index for each grid cell. In our case it takes the form

$$\exp\{-\sum_{k=1}^{L}\left(\frac{p_i^{(k)}}{\sum_k p_i^{(k)}}\log\frac{p_i^{(k)}}{\sum_k p_i^{(k)}}\right)\} \tag{9}$$

where L is the number of species. Note that (9) is maximized at $\frac{p_i^{(k)}}{\sum_k p_i^{(k)}} = \frac{1}{L}$ and equals L in this case. It is minimized if $\frac{p_i^{(k)}}{\sum_k p_i^{(k)}} \to 1$ for some k and tends to 1 in this case. Hence (9) is scaled to the number of species.

The interpretation which is attached to (9) is that it will be large if many species are equally likely to co-occur in grid cell *i*. It will be small if one or two species are much more likely to occur than the others. However, it is not a measure of richness since only the relative magnitudes of the $p_i^{(k)}$ matter, not the absolute ones. A suitable display would provide a map of, say, the posterior mean of (9) across *i*. We refer to this as a diversity plot.

Finally, related to the foregoing discussion, we consider the issue of beta diversity — how species composition changes with distance over our study region. 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 use the $E(p_i^{(k)}|\mathbf{Y})$ to investigate this as well. Defining $E(\mathbf{p}_i|\mathbf{Y})$ to be the $K \times 1$ vector whose entries are $E(p_i^{(k)}|\mathbf{Y})$, overlap (equivalently, turnover) is reflected by the similarity (difference) between these vectors. Using a neighborhood structure (say, first or second order), for

each pixel i, an illustrative measure computes

$$h_i = \exp\left(-\sum_{j\in\delta i}\frac{||E(\mathbf{p}_i|\mathbf{Y}) - E(\mathbf{p}_j|\mathbf{Y})||}{\text{number of neighbors of } i}\right). \tag{10}$$

where $||\cdot||$ denotes Euclidean distance. For cell i, h_i yields an average similarity (first or second order) of cell i with its neighbors. When h_i is large, high overlap is indicated; when h_i is small high turnover is indicated. A choropleth map of the h_i will reveal where in the region overlap is high, where it is low.

7 Analysis for the Kogelberg-Hawequas sub-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 as the sampling locations overlaid. There are a total of 7541 sampling locations within the region including null sites (sites where nothing was observed). The six most important environmental data layers (as a result of our modeling - see Table 2 below) are ROUGH, ELEV, JULMINT, PPTCON, EVI, TEXT3. Perspective plots for these six variables are shown in Figures B-G of the supplementary material. Spatial pattern arises for all six variables. Moreover, the patterns are different for each, mitigating concern with regard to multicollinearity.

Forty species were selected somewhat arbitrarily, but to provide a behaviorally diverse group. They are listed alphabetically with abbreviated versions of their full latin names in Table A of the supplementary material. The most frequently occurring, *Leucadendron salignum*, was found at 629 of the grid cells (40%). The least frequently occurring, *Mimetes stokoei* was found at 1 grid cell (0.06%). The species attribute classifications ($\mathbf{z_k}$) are given in Table A, as well.

Table 2 assembles a summary of the significance of each of the species level coefficients for the environmental variables. It is clear that some environmental factors are more important than others. For instance, rainfall concentration (PPTCON) and July minimum temperature (JULMINT) are significant for 21 of 40 species. By contrast, potential evapotranspiration (POTEVT) and soil texture class 3 (TEXT3) are only significant for three species. The frequently occurring species such as the various leucadendrons find significance on many of the variables. The rarest ones, like *Sorocephalus imbricatus* finds significance on none of them. Posterior box plots for the $\beta_l^{(k)}$ are more informative and are presented for the six layers above as Figures H-M in the supplementary material. Consistent with Table 2, the figures reveal considerable differences across species k, both positive and negative significance and that, generally, the width of the interval estimate reflects the frequency of occurrence of the species across the 7541 site records in our subregion.

Table 3 provides a summary of the inference for the coefficients of the species level attributes (γ's) implicit within (5). Not surprisingly, potential presence is encouraged by increasing local population size (>1000 is the baseline classification here). Also,

larger potential presence will be associated with resprouters. The remaining attributes do not show any significance.

Figure 3 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 presence/absence, 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.

Next, we turn to the patterns of species distributions and ranges described in the previous section. For species range we illustrate with six species, which are quite different from each other with regard to abundance and range, *Protea cynaroides*, *Leucadendron salignum*, *Aulax umbellata*, *Diastella myrtifolia*, *Protea restionifolia*, *Mimetes arboreus* and *Mimetes argenteus*. The map for *Protea cynaroides* is shown in Figure 4; the remaining five figures are shown as Figures N-R in the supplemental material. In each case we present the potential range, the adjusted range and variance of the potential range. The observed range, i.e., the locations where the species are observed are overlaid on the potential and adjusted range. The posterior predictions for each species distribution show remarkably tight agreement with the observed data points. The 6 species illustrated are fully representative of the suite of 40 species modeled. For *Protea cynaroides* (Figure 4) with a large number of observed locations across the region, but a narrow observed range, and with a large number of significant environmental explanatory variables (11), the fits are quite tight and the levels of uncertainty (variances) are quite low. Discussion of the five analogous figures is offered in the supplementary materials. A prevalence plot for the 40 species is presented in Figure 5. Clearly this shows a broad range of prevalence patterns among the 40 species modeled.

Turning to species richness, in Figure 6 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 6, with regard to explaining richness, Table 4 summarizes the posterior distribution for the $\sum_{k=1}^{40} \beta_l^{(k)}$. Here we see amplification of the discussion in Section 6. JULMINT and WINSMD emerge as significantly positive with PH3 significantly negative. Roughness, EVI, and FERT1 are suggestively significant. However, ELEV, for example, which was significant for 10 of 40 species is essentially centered around 0 here. This reflects the fact that 6 of those significance were positive, 4 negative with resultant cancellation in the sum. A diversity index (Shannon-Wiener) plot plot (see (9) above), and first and second order neighborhood similarity plots (see (T) above) are presented as Figures S and T in the supplementary materials.

8 Interspecies Dependence

So far in our modeling approach, presence or absence for species k in grid cell i is independent of that for species k' given ρ_i. This assumption facilitates writing and

computing the likelihood but may be called to questions in that it fails to reflect the possibility that, e.g., the presence of one species may diminish the chance of another being present. In fact, since presence or absence is viewed with respect to 1 min by 1 min grid cell, this is not likely to be a substantial concern. However, it does raise the notion of evolutionary constraints to the distribution of species. In fact, allopatric speciation is an evolutionary phenomenon which promotes separation of closely related species ranges within a particular domain.

Speciation is the process of divergence in which a single ancestral species becomes a pair of sister species. Though formal definition of species is difficult, the popular "biological" concept (Mayr 1942) holds that populations are separate species if they are reproductively isolated, and thus isolating mechanisms play a central role in proposed mechanisms of speciation (Grant 1981). Vicariance biogeography predicts that populations separated geographically by changes in river course, mountain building, etc., will over time evolve independently and become reproductively isolated and hence separate species (Wiley 1981). Allopatric speciation may also result from adaptation of a peripherally isolated population to local conditions, which may be near the limit of tolerance for the species as a whole. Regardless of the mechanism, this allopatric sister species pattern is common and expected for closely-related species, but over time becomes less predictable because of the potential for extinction of one of the two species, or futher adaptation that allows one or both species to expand into the range of the sister species, obscuring their allopatric origins. For simplicity, we refer to this allopatric speciation pattern as vicariance, regardless of the particular mechanism involved, and use the phrase "vicariance promoter" to describe means for encouraging the model to keep closely-related species allopatric.

The motivation for incorporating such predictions into our spatial model comes from the observation that closely-related sister species are often similar enough ecologically that, despite essentially disjoint ranges, each is predicted to occur in the other's range. We need to introduce some dependence between species at the grid cell level to remedy this deficiency in our predictions. The relative amount of time separating pairs of species can be independently determined through estimating phylogenies (genealogies that relate species rather than individuals) using DNA sequence data. The challenge has been to construct a model in which history places some constraints (which decay with time) on the predicted occurrence of species relative to their sister species in the phylogeny. A primary hurdle lies in the fact that phylogenetic information is naturally expressed in relative measures (i.e. the length of a path through the graph from one species to another species), whereas ecological data is naturally species-specific (species attributes) or area-specific (environmental variables).

A full discussion of modeling to accommodate allopatry is presented in Wu et al. (2004). Here we offer only a simple illustration for two sister species, *Mimetes arboreus* and *Mimetes argenteus*, which for simplicity we label as A and B respectively and show but one of many models which promote vicariance. In fact, for ease of interpretation we introduce two latent variables at each site. X_i^{AB} denotes the presence/absence state at grid cell i for the ancestral species of A and B (prior to speciation). We also introduce D_i^{AB} a "vicariance promoter" for grid cell i, i.e., $D_i^{AB} = 1$ encourages A to be present

at site i, $D_i^{AB} = -1$ encourages B to be present at cell i. That is, in (5) we add to the right side $+\eta D_i^{AB} X_i^{AB}$ for species A and $-\eta D_i^{AB} X_i^{AB}$ for species B with η an unknown coefficient. (Obviously, $D_i^{AB} X_i^{AB}$ can be written as T_i^{AB} a three-level indicator taking values -1, 0 and 1).

To complete the model specification, we introduce a Potts model for X_i^{AB} and also for D_i^{AB} (See Green and Richardson (2002)). That is,

$$P(D_i^{AB}|D_j^{AB}, j \neq i) = Ce^{\tau_D \sum_{j\in\delta_i} 1(D_i^{AB}=D_j^{AB})}$$
$$P(X_i^{AB}|X_j^{AB}, j \neq i) = Ce^{\tau_X \sum_{j\in\partial i} 1(X_i^{AB}=X_j^{AB})} \quad (11)$$

In (11), ∂i indicates the neighbors of grid cell i and τ_D and τ_X are two positive scaling parameters. Discrete priors on τ_D and τ_X, following the suggestions of Green and Richardson are used. In the supplementary materials we present three summary figures (Figures U-W) and related discussion for these two *Mimetes* species.

9 Discussion

How and why species are distributed across the landscape in the manner they are, has occupied the minds of many scientists and natural historians for hundreds of years; thousands of publications have been produced on this subject. Many feel (e.g., Gaston 2003) that we still have not come very far in addressing the fundamental problem of identifying: repeatable attributes of geographic distributions, variability in these attributes, and their determinants. This is true for particular species as well as collected species in general, which, in a spatial context, constitutes biodiversity. The study reported here provides a way to characterize geographic distributions of species across landscapes as potential presence surfaces, and to specify spatial uncertainty in these. We can do the same for suites of species considered jointly as biodiversity (i.e. species richness) surfaces. Moreover, for the extensive list of possible explanatory variables considered, we can attribute the contribution of each to explaining the distribution of individual species and jointly the biodiversity. We know of no other study that has accomplished this to date.

In addition to these more basic implications, there are a number of direct applications. Many of these are related to the effects of human activities altering the landscape and the consequences of this on patterns of biogeography and biodiversity. Of fundamental concern to many is the impact of human activities on biodiversity, the occurrence of threatened and endangered species, and consequently on ecosystem functions and services. The latter will certainly have the biggest direct impact on human well being. The specific applications from our study that are relevant here include: conservation planning for protected areas, predicting species extinctions or more generally reductions in the distributions of species, planning for landscape restoration or species re-introductions, assessing and predicting the effects of alien species invasions, and predicting biogeographic responses to climate change. The planning of protected areas for conservation is often based on finding the minimum spatial area needed to conserve the maximum number of species. Usually this is done without considering full

inference regarding species range, uncertainty in this inference, edge effects, biases, and other factors.

The methodology we have developed here has direct application to conservation planning. Moreover, the spatial predictions for potential species presence, as products from our model, have direct implications with regard to chances of extinction or reductions in geographical distributions of species. Here species prevalence or magnitude of occurrence probabilities across the landscape may give some insight to species extinction risks. The model provides potential presence across the landscape for sites at which species have not been observed (or not censused). Sites with sufficiently elevated potential and where the selected species appear to not be present, are obviously target regions for species reintroduction or restoration. For example, Figure N of the supplementary material shows broad areas where *Leucadendron salignum* does not now occur, but indicate high potential for success if the species was successfully planted or reintroduced.

The modeling we have developed for this study can also be applied to making predictions of species responses to global climate change. Certainly this remains one of the most important environmental concerns today. For a variety of different climate change scenarios (e.g. increases or decreases in local temperature, precipitation, etc.) one can easily project the altered distribution of modeled species and evaluate these predictions. In summary, biogeographic distributions are richly textured surfaces – complex topographies in species occurrences and biodiversity. These surfaces, unique to each species, ebb and flow spatially and temporally; in consequence they have proven difficult to visualize, understand, and predict. The modeling approach we have introduced here appears to provide some solution to this challenge.

Table 1: Model Comparison Using DIC

		$\theta = \{P(Y_{ij}^{(k)} = 1)\}$		$\theta = \{\text{model parameters}\}$	
models	#par*	p_D	DIC	p_D	DIC
T(Topography)	124	268.77	104564.86	399.16	104695.25
G(Geology)	364	490.79	101614.73	718.30	101842.24
C(Clima., Precip.)	484	624.04	91369.39	994.58	91739.93
G,C,T	884	943.14	88739.56	1487.35	89283.78
G,C,T,A(Attributes)	891	950.77	88861.54	1500.34	89411.10
ρ(Spatial effect)	1597	934.61	109070.14	1128.69	109264.22
T,ρ	1677	1021.68	103679.11	1392.69	104050.12
G,ρ	1917	1160.62	100620.46	1724.01	101183.85
G,T,ρ	1997	1216.52	96903.28	1917.91	97604.67
G,T,ρ,A	2004	1214.63	96882.00	1908.65	97576.02
C,ρ	2037	1167.91	90302.25	1904.65	91038.98
C,T,ρ	2117	1213.75	89532.33	2004.36	90322.93
C,T,ρ,A	2124	1217.49	89586.47	2013.98	90382.96
G,C,ρ	2357	1271.82	88195.33	2166.15	89089.66
G,C,ρ,A	2364	1269.62	88147.81	2156.59	89034.78
G,C,T,ρ	2437	1327.15	87788.84	2263.40	88725.09
G,C,T,ρ,A(Full)	2444	1323.41	87738.79	2243.82	88659.19

$*$: #par is the number of independent parameters in regression equation on $\text{logit}(p_i^{(k)})$ and $\text{logit}(q_i^{(k)})$

Table 2: Posterior Summary of Environmental Coefficients for Each Species (β's).

SPECIES (#pixel obs)	ROUGH	ELEV	POTEVT	PPTCV	FROST	HEATU	JANMAX	JULMIN	MAP	PPTCON	SUMSMD	WINSMD	EVI	FERT1	FERT2	FERT3	TEXT1	TEXT3	TEXT4	pH1	pH3	TOTALS
L. salignum (629)	+	−						−		+	(+)	−						+	(+)		(+)	6(3)
P. repens (541)	+							−		(+)		(−)										2(2)
H. sericea (525)	+		(+)					−		+	(+)					+		(+)				4(3)
P. cynaroides (371)		+			+			+	−	−		+	+					−		+	+	10(0)
L. spissifolium (324)	+	(+)				−			(−)				+								+	4(2)
M. cucullatus (289)								+	−	−			(+)	+			−					5(1)
L. salcifolium (271)	−		+		+	−		(+)					+		+		−	−	−	+	+	11(1)
L. rubrum (202)					−			−		+			(+)		−	−	+	+		−	−	9(1)
L. microcephalum (135)	−	(−)			+			+			−	+		(+)			−	−	(−)			7(3)
P. neriifolia (124)		(+)					−	+		+		+	+									5(1)
Se. elongata (114)						(−)		+		−	−						−					4(1)
L. oleifolium (110)		+		−		(+)		+		−												4(1)
A. umbellata (106)		−								−	−						−					4(0)
Se. fasciflora (103)					−	(+)	(−)	(+)		+			(+)			−	−	(+)			−	5(5)
A. pallasia (66)	+			−					(−)								+	(+)				3(2)
L. corymbosum (65)		(−)				+	+											+	−			4(1)
A. cancellata (53)	+						(−)	+		−	(−)											3(2)
P. grandiflora (52)		+											+						(−)		(+)	2(2)
L. daphnoides (48)								−				+	+					(+)				3(1)
L. sessile (42)	+							+		+							(+)				(−)	3(2)
P. nana (40)	(−)		+	−			+	−		−				+	+							7(1)
Sp. curvifolia (25)	+							+		−		(+)										3(1)
L. tinctum (22)	+					−				(−)							(−)					2(2)
P. lacticolor (21)		(+)											(+)									0(2)
P. mundii (19)	(+)									−									−			2(1)
M. arboreus (17)	+			(−)				+		−			+									4(1)
M. argenteus (15)								(+)					(+)									0(2)
P. punctata (14)	+							−			−		(−)									3(1)
L. grandiflorum (11)										+					+					(−)		2(1)
O. zeyheri (9)				(−)						−												1(1)
M. hottentoticus (8)							−			−			+									3(0)
D. myrtifolia (6)																						0
L. bolusii (6)								+					−				+					3(0)
P. rupicola (5)																						0
P. restionifolia (4)																						0
L. comosum (3)																						0
L. elimense (3)																						0
Se. zeyheri (3)																						0
So. imbricatus (3)																						0
M. stokoei (1)																						0
TOTAL + (95%)	11	3	2	0	3	1	2	11	0	7	0	4	8	2	3	1	3	3	0	2	3	
TOTAL (+) (90%)	1	3	1	0	0	2	0	3	0	1	2	1	5	1	0	0	1	4	1	0	2	
TOTAL − (95%)	2	2	0	3	2	3	2	7	2	12	4	1	1	0	2	2	5	3	3	1	2	
TOTAL (−) (90%)	1	2	0	2	0	1	2	0	2	1	1	1	1	0	0	0	1	0	2	1	1	
ALL +	12	6	3	0	3	3	2	14	0	8	2	5	13	3	3	1	4	7	1	2	5	
ALL −	3	4	0	5	2	4	4	7	4	13	5	2	2	0	2	2	6	3	5	2	3	

+ or - are positive or negative coefficients with 95% credible intervals that do not overlap 0.

(+) or (-) are positive or negative coefficients with 90% credible intervals that do not overlap 0.

Table 3: Posterior Summary of the Coefficients of the Species Level Attributes (γ's).

Covariate	Mean	2.5%	50.0%	97.5%
Height	-0.61	-2.14	-0.6	0.99
Locpop1(1-50)	-4.83	-7.58	-4.84	-1.95
Locpop2(50-1000)	-3.02	-5.67	-3.09	-0.05
Resprout (yes)	3	-0.04	3.01	5.93
Pollen1(bird)	-0.79	-3.39	-0.77	1.71
Pollen2(wind)	1.08	-2.45	1.06	4.75
Disp(wind)	1.95	-0.69	1.88	4.63

Table 4: Posterior Summary of the Area Level Attributes in terms of Potential Richness ($\sum_k \beta_l^{(k)}$).

Covariate	Mean	2.5%	50%	97.5%
ROUGH	8.01	-0.45	8.02	16.51
ELEV	0.21	-12.37	0.29	12.35
POTEVT	5.03	-10.84	5.13	21.15
PPTCV	-11.12	-23.88	-11.03	1.46
FROST	11.45	-22.37	-11.63	0.52
HEATU	-14.19	-27.17	-14.11	1.88
JANMAXT	-2.16	-15.13	-2.27	11.63
JULMINT	16.74	1.78	16.86	31.7
MAP	-19.61	-32.91	-19.69	-5.47
PPTCON	-12.26	-25.33	-12.29	1.18
SUMSMD	-15.59	-27.83	-15.44	-3.3
WINSMD	14.3	1.17	14.31	27.63
EVI	7.97	-0.49	7.88	16.45
FERT1	13.73	-0.9	13.72	27.29
FERT2	3.62	-7.56	3.49	15.6
FERT3	-4.44	-19.46	-4.11	8.79
TEXT1	-10.56	-24.42	-10.55	2.76
TEXT3	0.54	-10.45	0.49	11.87
TEXT4	-7.4	-19.28	-7.5	4.83
PH1	3.78	-10.62	3.42	17.85
PH3	-10.36	-21.18	-10.2	-0.18

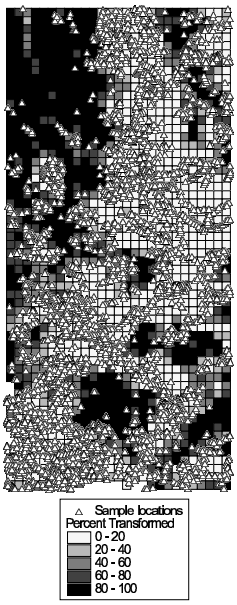

Figure 1: The Kogelberg-Hawequas sub-region used for our study overlaid with the sampling locations.

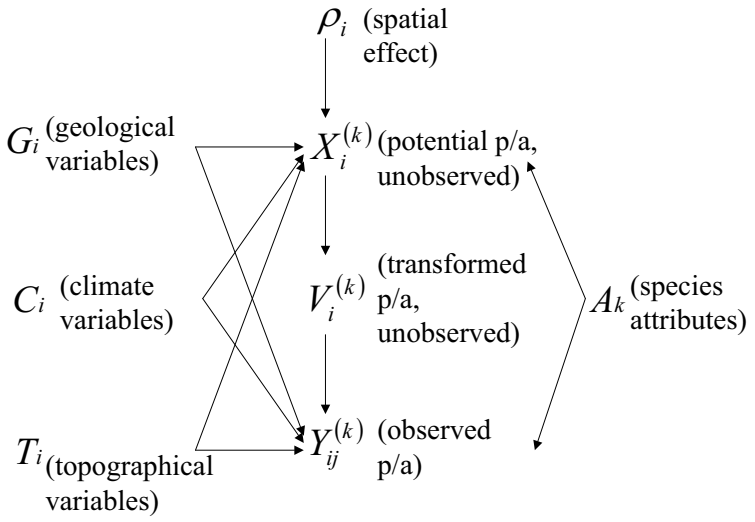

Figure 2: A graphical model for the hierarchical specification.

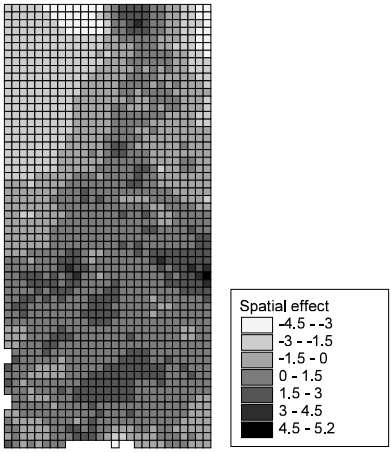

Figure 3: Posterior mean of the spatial effects ($\rho_i' s$).

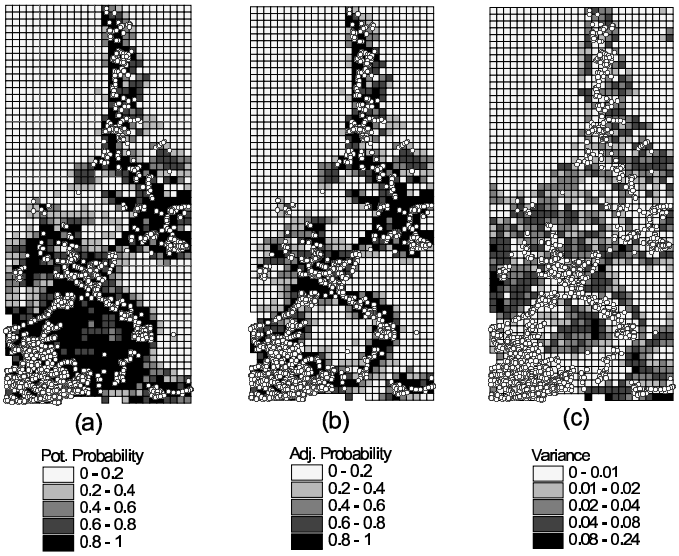

Figure 4: The potential, "adjusted" predicted range of *Protea cynaroides* and the variance in the potential range. ((a) is potential range, (b) is adjusted range, and (c) is variance in potential range.)

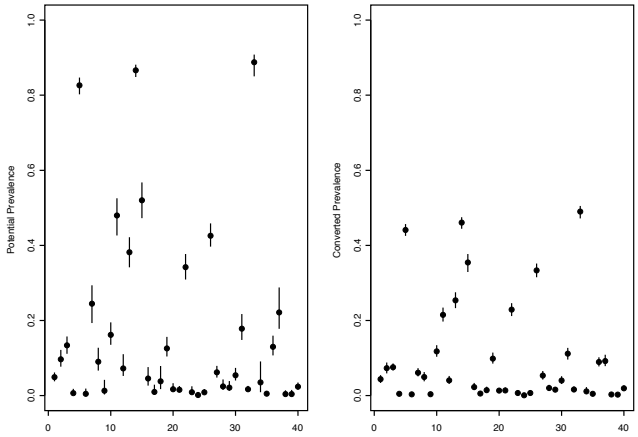

Figure 5: Potential and Transformed Prevalence Plot (95% credible set)(Species order is that of Table 2)

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. 44, 49

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. 44, 49

Augustin, N. H., Mugglestone, M. A., and Buckland, S. T. (1996). "The fate of clades in a world of recurrent climatic change: Milankovitch oscillations and evolution." *Journal of Applied Ecology*, 33:339–347. 49

Austin, M. P. and Meyers, J. A. (1996a). "Current Approaches to Modelling the Environmental Niche of Eucalypts: implication for management of forest biodiversity." *Forest Ecology and Management*, 85:95–106. 42

— (1996b). "Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity." *Forest Ecology and Management*, 85:95–106. 48

Austin, M. P., Nicholls, A. O., and Margules, C. R. (1990). "Measurement of the realized qualitative niche: environmental niches of five Eucalyptus species." *Ecological Monographs*, 60:161–177. 48

Besag, J. (1974). "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society, Series B*, 36:192–225. 54

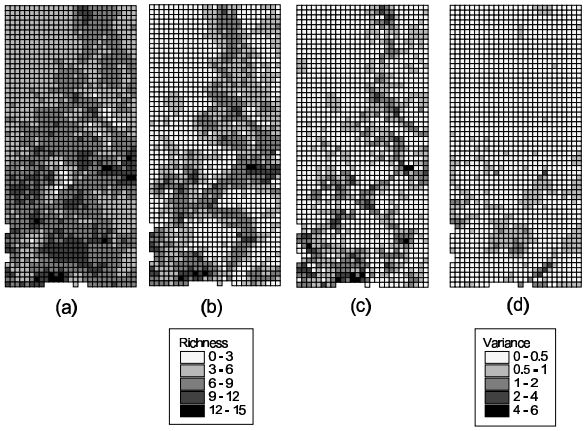

Figure 6: Observed (a), potential (b) and transformed (c) richness, and variance (d) in potential richness.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). "Bayesian Computation and Stochastic Systems (with discussion)." *Statistical Science*, 10:3–66. 55

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. 44

Clark, J. S., LaDeau, S., and Ibanez, I. (2003). "Fecundity of trees and the colonization-competition hypothesis." *Ecological Monographs*, (to appear). 50

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. 43, 47

Cressie, N. A. C. (1993). *Statistics for Spatial Data. Revised Edition.* John Wiley & Sons, Inc. 51

Currie, D. J. (1991). "Energy and Large-scale Patterns of Animal and Plant Species Richness." *American Naturalist*, 137:27–49. 43, 47

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. 43

Dynesius, M. and Jansson, R. (2000). "Evolutionary consequences of changes in species' geographic distributions driven by Milankovitch climate oscillations." In *Proceedings of the National Academy of Sciences*, 9115–9120. 47

Ferrier, S., Drielsma, M., Manion, G., and Watson, G. (2002a). "Extended Statistical Approaches to Modelling Spatial Pattern in Biodiversity in Northeast New South

Wales. II. Community-level modeling." *Biodiversity and Conservation*, 11:2309–2338. 43

Ferrier, S., Watson, G., Pearce, J., and Drielsma, M. (2002b). "Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modeling." *Biodiversity and Conservation*, 11(2275-2307):Biodiversity and Conservation. 48

Fischer, H. S. (1990). "Simulating the distribution of plant communities in an alpine landscape." *Coenoses*, 5:37–43. 49

Gaston, K. J. (2003). *The Structure and Dynamics of Geographic Ranges.*. Oxford University Press, Oxford, UK. 62

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1996). "Efficient Parametrizations for Generalised Linear Models." *Bayesian Statistics 5: 227-246, Eds: Bernardo, J.M. et al., Oxford University Press: Oxford*. 54

Gelfand, A. E., Schmidt, A. M., Wu, S., J. A. Silander, J., Latimer, A., and Rebelo, A. G. (2003). "Modelling Species Diversiy Through Species Level Hierarchical Modeling." *Applied Statistics. forthcoming.*. 49

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. 55

Gilks, W. R., Best, N., and Tan, K. K. C. (1995). "Adaptive Rejection Metropolis Sampling within Gibbs Sampling." *Applied Statistics*, 44:455–472. 55

Gilks, W. R. and Wild, P. (1992). "Adaptive Rejection Sampling for Gibbs Sampling." *Applied Statistics*, 41(2):337–348. 55

Grant, V. (1981). *Plant Speciation. 2nd. ed.*. Columbia University Press, New York. 61

Green, P. J. and Richardson, S. (2002). "Hidden Markov Models and Disease Mapping." *Journal of the American Statistical Association*, 94:1055–1070. 62

Guisan, A., Edwards, J., 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. 43, 48

Guisan, A. and Zimmerman, N. E. (2000). "Predictive Habitat Distribution Models in Ecology." *Ecological Modelling*, 135:147–186. 43, 48, 52, 57

Heegard, E. (2002). "The outer border and central border for species-environmental relationships estimated by non-parametric generalised additive models." *Ecological Modelling*, 157:131–139. 48

Heikkinen, J. and Högmander, H. (1994). "Fully Bayesian Approach to Image Restoration with an Application in Biogeography." *Applied Statistics*, 43:569–582. 56

Heikkinen, R. K. (1996). "Predicting Patterns of Vascular Plant Species Richness with Composite Variables: A Mesoscale Study in Finnish Lapland." *Vegetation*, 126:151–165. 43

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(1):102–114. 44, 53

Högmander, H. and Møller, J. (1995). "Estimating Distribution Maps from Atlas Data Using Methods of Statistical Image Analysis." *Biometrics*, 51:393–404. 56

Hooten, M. B., Larsen, D. R., and Wikle, C. K. (2003). *Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model.*. in review. 50

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, NJ, USA. 43

Huston, M. A. (1994). *Biological Diversity. The Coexistence of Species on Changing Landscapes*. Cambridge University Press, Cambridge, UK. 43

Jansson, R. and Dynesius, M. (2002). "The fate of clades in a world of recurrent climatic change: Milankovitch oscillations and evolution." *Annual Review of Ecology and Systematics*, 33:741–777. 47

Kempton, R. A. (2002). "Species Diversity." *In A.H. El-shaarwari and W.A. Piegorsch, Eds. Encyclopedia of Environmentircs*, 4:2086–2092. 58

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. 43

Leathwick, J. R. (2002). "Intra-generic Competition among Nothofagus in New Zealand's Primary Indiginous Forests." *Biodiversity and Conservation*, 11:2177–2187. 42

Lehmann, A. (1998). "GIS modeling of submerged macrophyte distribution using generalized additive models." *Plant Ecology*, 139:113–124. 48

Lehmann, A., Overton, J. M., and Leathwick, J. R. (2002). "GRASP: Generalized Regression Analysis and Spatial Prediction." *Ecological Modelling*, 159:189–207. 44, 48, 49

MacArthur, R. H., Recher, H. F., and Cody, M. (1966). "On the relation between habitat selection and species diversity." *American Naturalist*, 100:319–332. 47

Manel, S., Dias, J.-M., and Ormerod, S. J. (1999). "Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird." *Ecological Modelling*, 120:337–347. 48

Mayr, E. (1942). *Systematics and the Origin of Species.*. Columbia University Press, New York. 61

Meyers, N., Mittermeier, R., Mittermeier, C. G., de Fonesca G. A. B., and Kent, J. (2000). "Biodiversity Hotspots for Conservation Priorities." *Nature*, 403:853–858. 45

Midgley, G. F., Hannah, L., Millar, D., Rutherford, M. C., and Powrie, L. W. (2002). "Assessing the Vulnerability of Species Richness to Anthropogenic Climate Change in a Biodiversity Hotspot." *Global Ecology and Biogeography*, 11:445–451. 42

Osborne, P. E. and Suarez-Seoane, S. (2000). "Should data be partitioned spatially before building large-scale distribution models?" *Ecological Modelling*, 157:249–259. 49

Owen, J. G. (1989). "Patterns of Herpetofaunal Species Richness:Relation to Temperature, Precipitation and Variance in Elevation." *Journal of Biogeography*, 16:141–150. 43

Palmer, M. W. (1996). "Variation in Species Richness: Towards a Unification of Hypotheses." *Folia Geobotanica et Phytotaxonomica (Praha)*, 29:511–530. 43

Papaspiliopoulos, O., Roberts, G. O., and Skold, M. (2003). "Noncentered Parametrization for Hierarchical Models and Data Augmentation." *Bayesian Statistics 7: 307-326, Eds: Bernardo, J.M. et al., Oxford University Press: Oxford.* 54

Rahbek, C. and Graves, G. R. (2001). "Multiscale assessment of patterns of avian species richness." In *Proceedings of the National Academy of Sciences 89(8)*, 4534–4539. 47

Rebelo, A. G. (1991). *Protea Atlas Manual: Instruction Booklet to the Protea Atlas Project*. Protea Atlas Project, Cape Town. 45

— (2001). *Proteas: A Field Guide to the Proteas of Southern Africa.* Fernwood Press, Vlaeberg, South Africa (2nd Edition). 45

— (2002a). "The Protea Atlas Project." Technical report, Retrieved on-line 12 May, 2002 from: `http://protea.worldonline.co.za/default.htm`. 45

— (2002b). "The State of Plants in the Cape Flora." In *Proceedings of a conference held at the Rosebank Hotel in Johannesburg*, 18–43. G.H. Verdoorn and J. Le Roux (editors) The State of South Africa's Species. Endangered Wildlife Trust. 45

Ritchie, M. E. and Olff, H. (1999). "Spatial Scaling Laws Yield a Synthetic Theory of Biodiversity." *Nature*, 400:557–560. 43

Rohde, K. (1992). "Latitudinal Gradients in Species Diversity: the Search for the Primary Cause." *Oikos*, 65:514–527. 43

Rosenzweig, M. L. (1995). *Species Diversity in Space and Time*. Cambridge University Press, Cambridge, UK. 43

Rouget, M., Richardson, D. M., Cowling, R. M., Lloyd, J. W., and Lombard, A. T. (2003). "Current Patterns of Habitat Transformation and Future Threats to Biodiversity in Terrestrial Ecosystems of the Cape Floristic Region, South Africa." *Biological Conservation*, 112:63–83. 46

Royle, J. A., Link, W. A., and Sauer, J. R. (2002). *Statistical mapping of count survey data. In J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.Q. Wall and F.B. Samson (Eds.) Predicting Species Occurrences - Issues of Accuracy and Scale.*. Island Press, Washington, DC. 49

Schultze, R. E. (1997). "South African Atlas of Agrohyrdology and Climatology." Technical report, Report TT82/96. Water Research Commission, Pretoria, South Africa. 46

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society, Series B*, 64:1–34. 56

Takhtajan, A. (1986). *Floristic Regions of the World*. University of California Press, Berkeley, CA, USA. 45

Venables, W. and Ripley, B. (1999). *Modern Applied Statistics with S-PLUS (3rd edition).*. Springer-Verlag, New York. 48

Wallace, A. R. (1895). *Natural Selection and Tropical Nature: Essays on Descriptive and Theoretical Biology.*. Macmillan, London. 47

Whittaker, R. J., Willis, K. J., and Field, R. (2001). "Scale and species richness: towards a general, hierachical theory of species diversity." *Journal of Biogeography*, 28:453–470. 47

Wikle, C. K. (2002). *Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains. In: A. Lawson and D. Denison (eds). Spatial Cluster Modelling.*. CRC Press, Boca Raton, FL. 49, 50

— (2003). "Hierarchical Bayesian models for predicting the spread of ecological processes." *Ecology (to appear)*. 50

Wikle, C. K. and Royle, J. A. (2002). *Spatial statistical modeling in biology. In: Encyclopedia of Life Support Systems.*. Publishers, Oxford, UK. 49, 50

Wiley, E. O. (1981). *Phylogenetics: the theory and practice of phylogenetic systematics.*. John Wiley and Sons, New York. 61

Woodward, F. I., Smith, T. M., and Emanuel, W. R. (1995). "A global land primary productivity and phytogeography model." *Global Biogeochemical Cycles*, 9(4):471–490. 47

Wu, S., Lewis, P., Holder, M., Silander, J. J., and Gelfand, A. E. (2004). "A Hierarchical Allopatry Model for Interspecies Range Dependence." *Submitted*. 61

Yee, T. W. and Mitchell, N. D. (1991). "Generalised additive models in plant ecology." *Journal of Vegetation Science*, 2:587–602. 48

Zaniewski, A. E., Lehmann, A., and Overton, J. M. (2002). "Predicting species spatial distributions using presence-only data: a case study of New Zealand ferns." *Ecological Modelling*, 157:261–280. 48

Supplementary Material

Table A: List of Species in the Study and their Attributes.

SPECIES	ABBREV.	HEIGHT	LOCAL POP.	RES-PROUT	POLLI-NATION	DIS-PERSAL
Aulax cancellata	AUCANC	1500	1 - 50	no	insect	wind
Aulax pallasia	AUPALL	2000	51 - 1000	yes	insect	wind
Aulax umbellata	AUUMBE	2000	>1000	no	insect	wind
Diastella myrtifolia	DIMYRT	750	1 - 50	no	insect	ant
Hakea sericea	HASERI	3500	>1000	no	insect	wind
Leucadendron comosum	LDCOMOH	1500	51 - 1000	no	insect	wind
Leucadendron corymbosum	LDCORY	1500	51 - 1000	no	insect	wind
Leucadendron daphnoides	LDDAPH	1000	>1000	no	insect	ant or rodent
Leucadendron elimense	LDELIMS	1000	1 - 50	no	insect	wind
Leucadendron microcephalum	LDMICR	1250	>1000	no	insect	wind
Leucadendron rubrum	LDRUBR	1500	>1000	no	wind	wind
Leucadendron sessile	LDSESS	1000	>1000	no	insect	ant or rodent
Leucadendron salcifolium	LDSFLM	2000	>1000	no	wind	wind
Leucadendron salignum	LDSGNM	500	>1000	yes	insect	wind
Leucadendron spissifolium	LDSPISS	1000	51 - 1000	yes	insect	wind
Leucadendron tinctum	LDTINC	750	51 - 1000	no	insect	ant or rodent
Leucospermum bolusii	LSBOLU	1000	>1000	no	insect	ant
Luecospermum grandiflorum	LSGRAN	1500	1 - 50	no	bird	ant
Leucospermum oleifolium	LSOLEI	750	51 - 1000	no	bird	ant
Mimetes arboreus	MIARBO	3000	1 - 50	yes	bird	ant
Mimetes argenteus	MIARGE	2500	1 - 50	no	bird	ant
Mimetes cucullatus	MICUCU	1000	1 - 50	yes	bird	ant
Mimetes hottentoticus	MIHOTT	2000	1 - 50	no	bird	ant
Mimetes stokoei	MISTOK	1500	1 - 50	no	bird	ant
Orothamnus zeyheri	ORZEYH	2900	1 - 50	no	insect	ant
Protea cynaroides	PRCYNA	1000	1 - 50	yes	bird	wind
Protea grandiflora	PRGRAN	2000	51 - 1000	no	bird	wind
Protea lacticolor	PRLACT	4000	>1000	no	bird	wind
Protea mundii	PRMUND	4000	51 - 1000	no	bird	wind
Protea nana	PRNANA	1000	51 - 1000	no	bird	wind
Protea neriifolia	PRNERI	2500	>1000	no	bird	wind
Protea punctata	PRPUNC	3000	>1000	no	bird	wind
Protea repens	PRREPE	2500	>1000	no	bird	wind
Protea restionifolia	PRREST	300	51 - 1000	yes	bird	wind
Protea rupicola	PRRUPI	1000	1 - 50	no	bird	wind
Serruria elongata	SEELON	1000	>1000	no	insect	ant
Serruria fasciflora	SEFASC	500	>1000	no	insect	ant
Serruria zeyheri	SEZEYH	400	1 - 50	no	insect	ant
Sorocephalus imbricatus	SOIMBR	1200	1 - 50	no	insect	ant
Spatalla curvifolia	SPCURV	650	51 - 1000	no	insect	ant

NOTE: Local Pop. >1000 set to 0; Non-resprouting set to 0; Insect pollination set to 0; Ant and/or rodent dispersal set to 0.

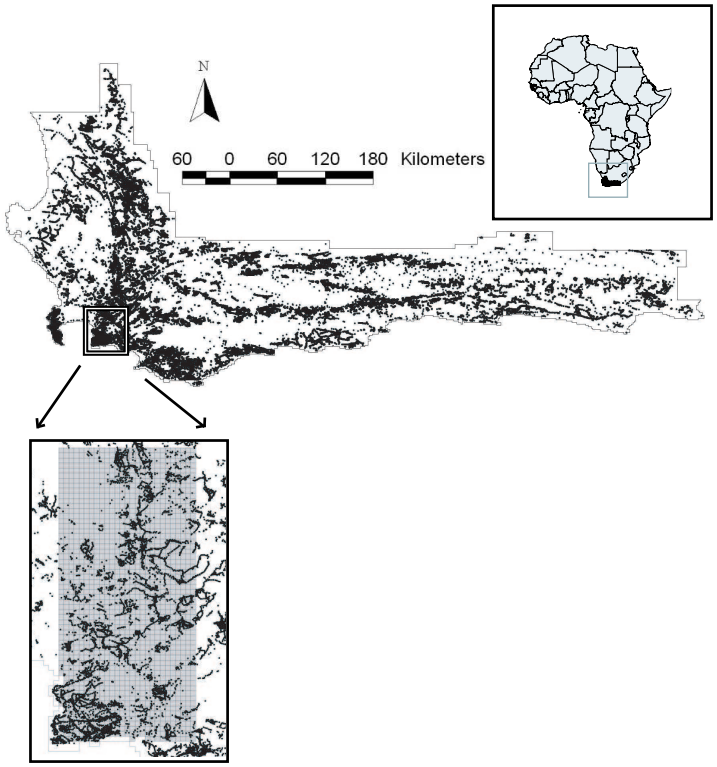

Figure A: The CFR and associated sampling locations.

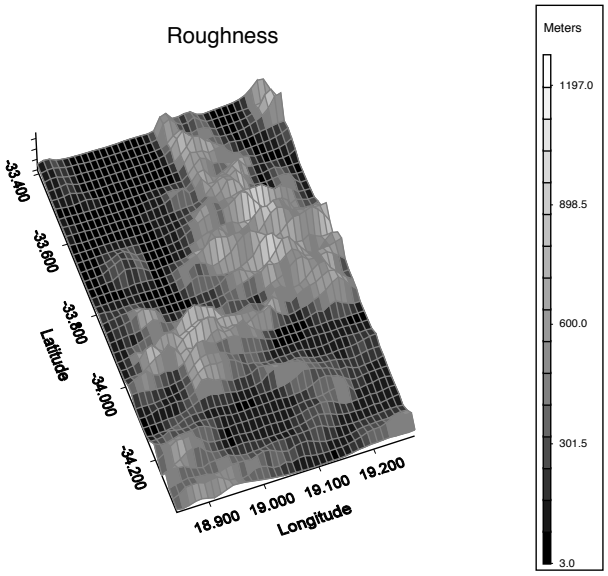

Figure B: Data layer of Roughness.

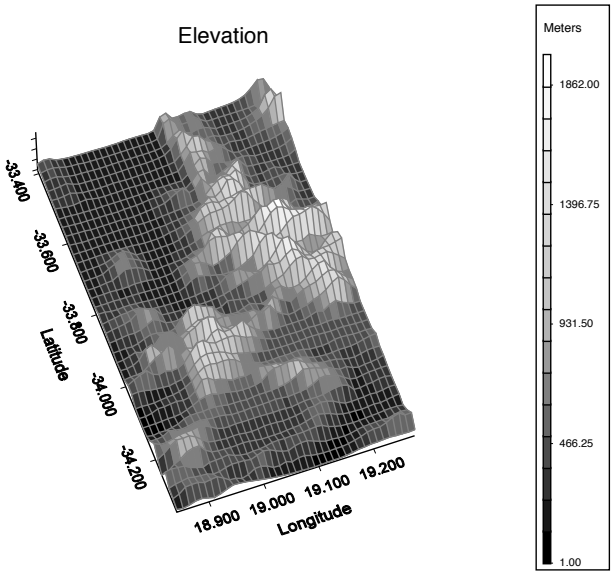

Figure C: Data layer of Elevation.

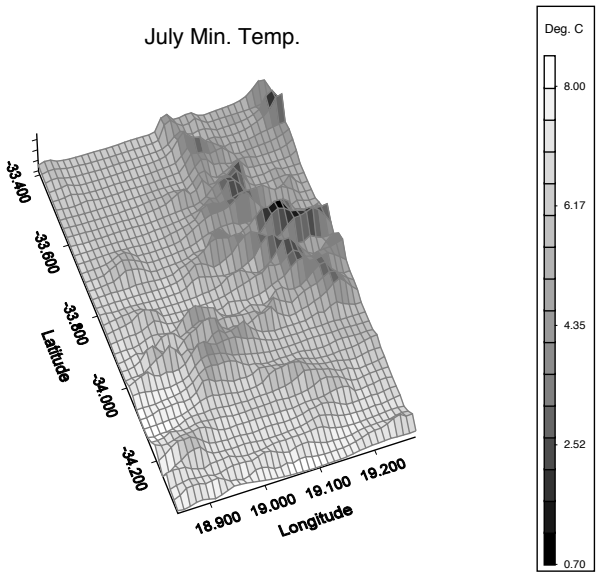

Figure D: Data layer of July Minimum Temperature.

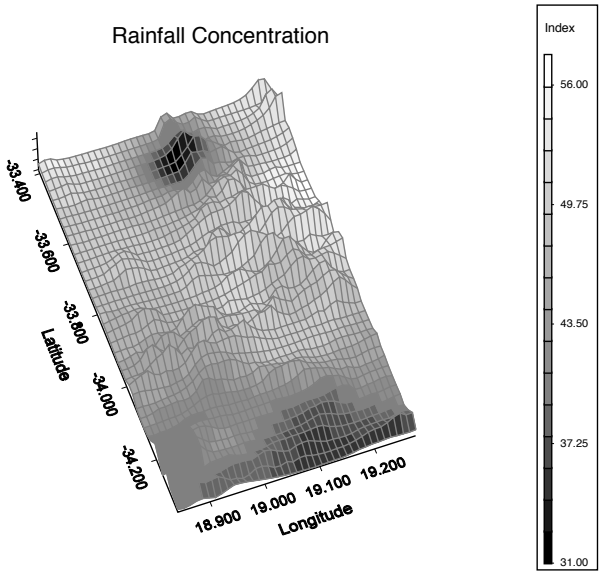

Figure E: Data layer of rainfall concentration.

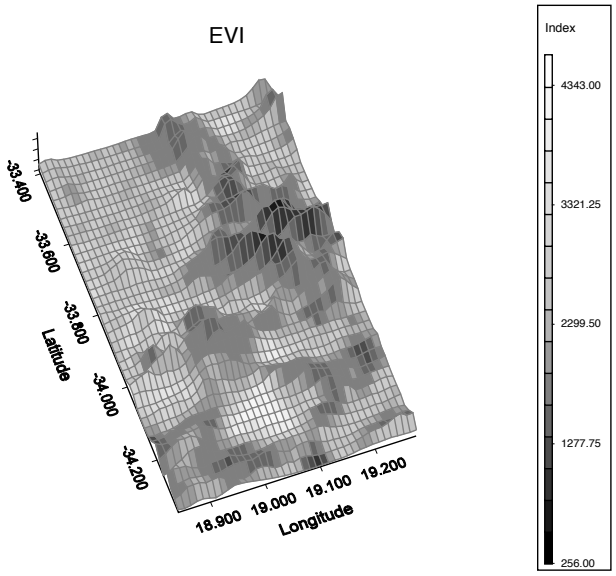

Figure F: Data layer of enhanced vegetation index.

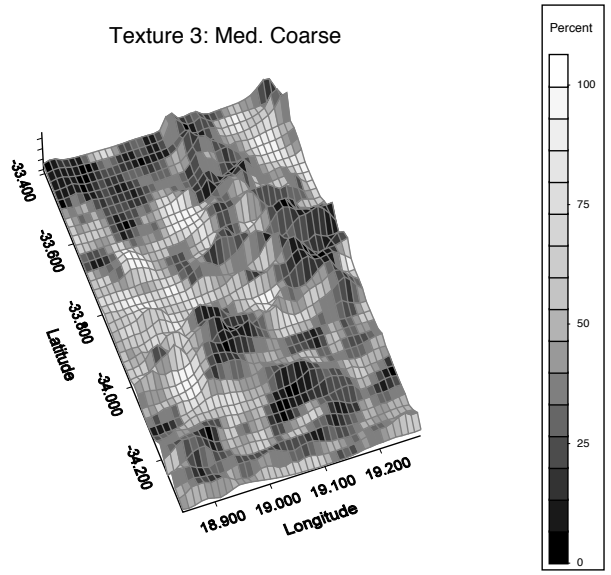

Figure G: Data layer of Text3.

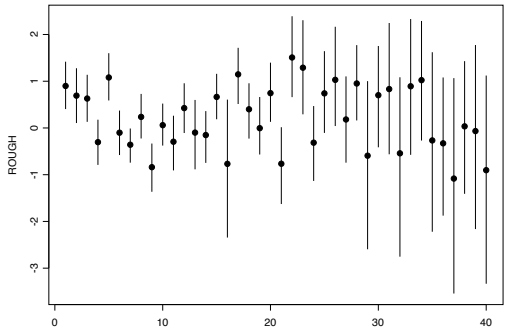

Figure H: Posterior summary of the coefficients of Roughness for each of the 40 species.

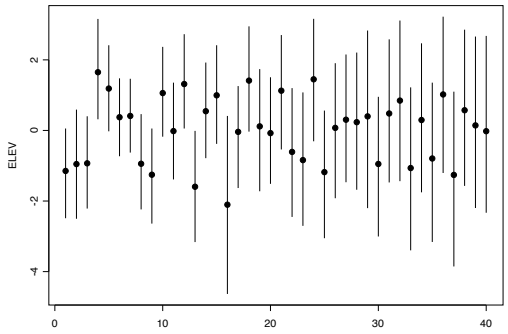

Figure I: Posterior summary of the coefficients of Elevation for each of the 40 species.

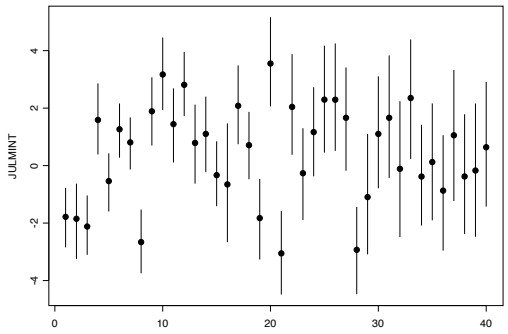

Figure J: Posterior summary of the coefficients of July Minimum Temperature for each of the 40 species.

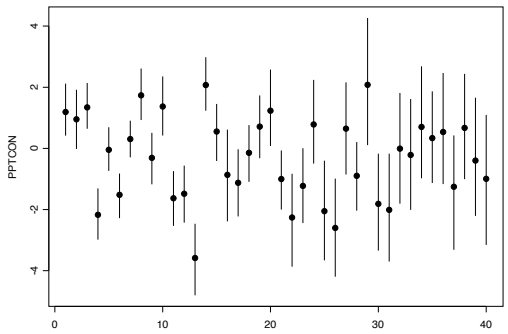

Figure K: Posterior summary of the coefficients of rainfall concentration for each of the 40 species.

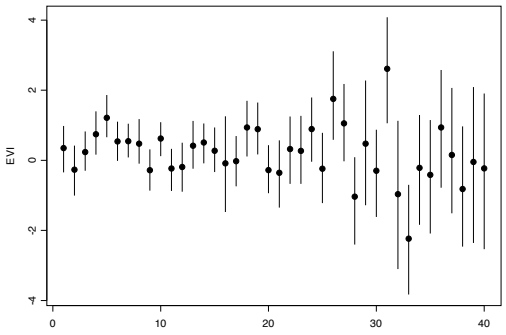

Figure L: Posterior summary of the coefficients of enhanced vegetation index for each of the 40 species.

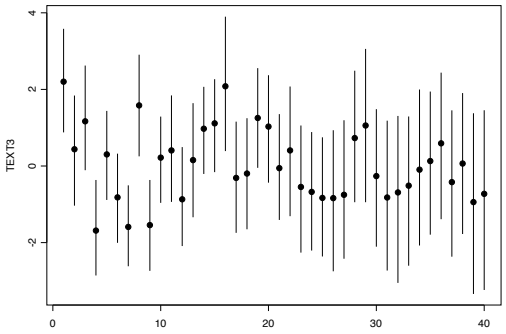

Figure M: Posterior summary of the coefficients of Text3 for each of the 40 species.

Discussion to Figures N-R

For a species like *Leucadendron salignum* (Figure N), which is just as commonly encountered in samples across the region, but is very widespread, the predictions are for a much broader prevalence across the region. Note that for *P. repens*, only 2 environmental explanatory variables figure prominently, and that the level of elevated uncertainty appears over a somewhat broader region than is the case for *P. cynaroides*. *Aulax umbellata* as well as *Mimetes arboreus* and *Mimetes argenteus* are species restricted to only a small portion of the landscape (Figures O and R). Yet even with a much smaller set of observations, the predicted fits still appear quite good. Four of the environmental explanatory variables figure prominently in predicting distribution of both species, but note that levels of uncertainty are more elevated in areas surrounding cells with high probabilities of occurrence. In the case of *M. arboreus* elevated probabilities of occurrence appear to the north and east of the observed sites, also associated with elevated variances. It turns out that these sites coincide roughly with known occurrences of a closely related sister species, *M. argenteus*, and the spatial pattern observed may well reflect similar ecological responses to the same explanatory variables (See Section 8). It is remarkable that one can obtain good fits with very rare species such as shown in Figure P. Here *Diastella myrtifolia* was only found in 3 of 7541 sample locations. But also note that these are 3 tightly clustered samples.

Spatial effects obviously play an important role even when none of the environmental explanatory variables are prominent. Elevated uncertainty is associated with the cells in the immediate vicinity of the sample points but is remarkably low or widely scattered elsewhere in the landscape. Not surprisingly, for some species that are rarely encountered at only widely scattered sites, such as observed for *Protea restionifolia* (Figure Q), the predicted distribution patterns are quite coarse and the level of uncertainty is quite high throughout the landscape.

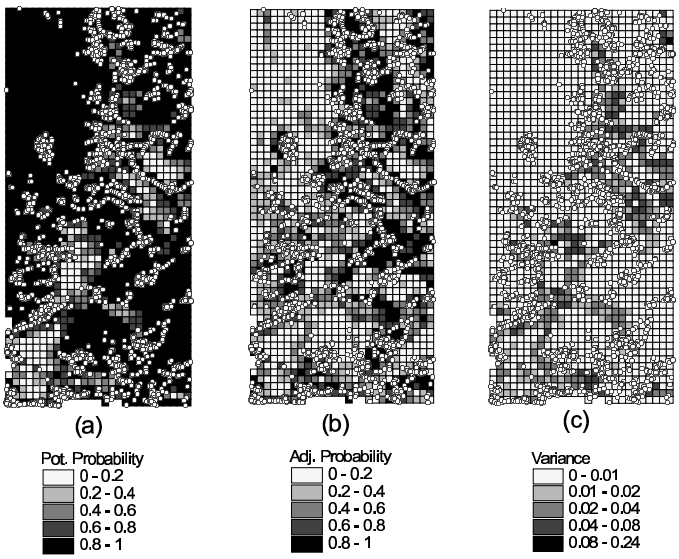

Figure N: The potential, "adjusted" predicted range of *Leucadendron salignum* and the variance in the potential range. ((a) is potential range, (b) is adjusted range, and (c) is variance in potential range.)

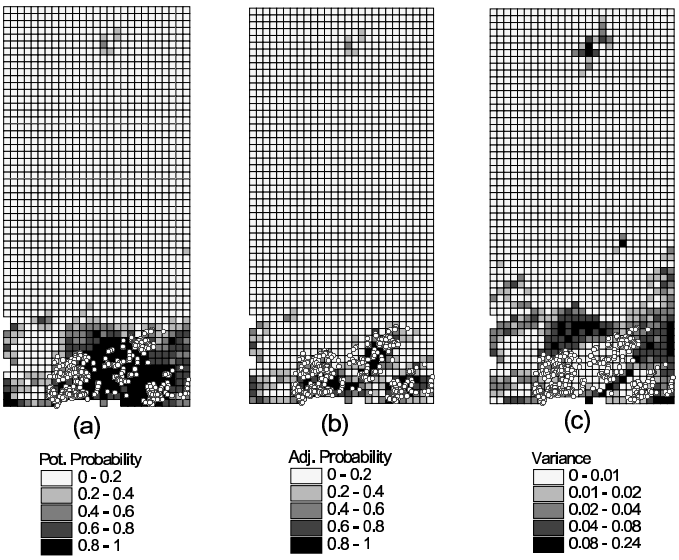

Figure O: The potential, "adjusted" predicted range of *Aulax umbellata* and the variance in the potential range. ((a) is potential range, (b) is adjusted range, and (c) is variance in potential range.)

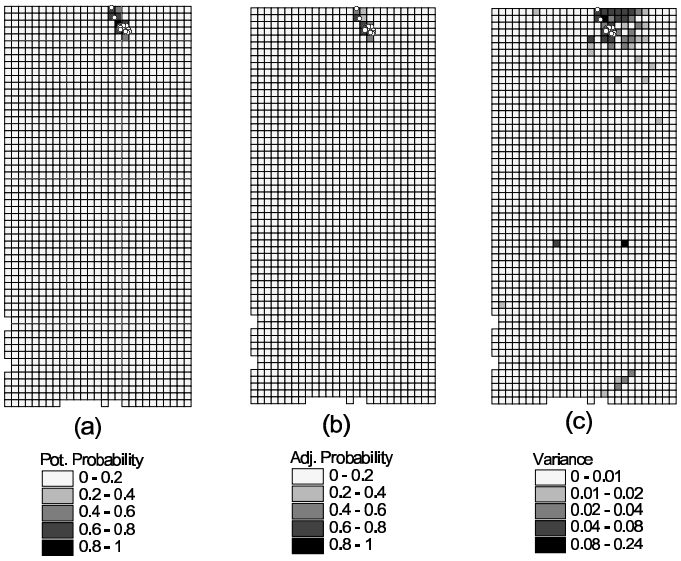

Figure P: The potential, "adjusted" predicted range of *Diastella myrtifolia* and the variance in the potential range. ((a) is potential range, (b) is adjusted range, and (c) is variance in potential range.)

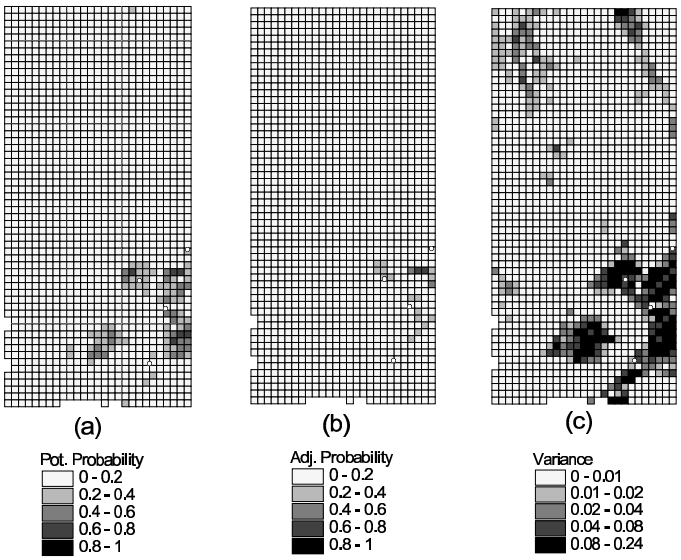

Figure Q: The potential, "adjusted" predicted range of *Protea restionifolia* and the variance in the potential range. ((a) is potential range, (b) is adjusted range, and (c) is variance in potential range.)

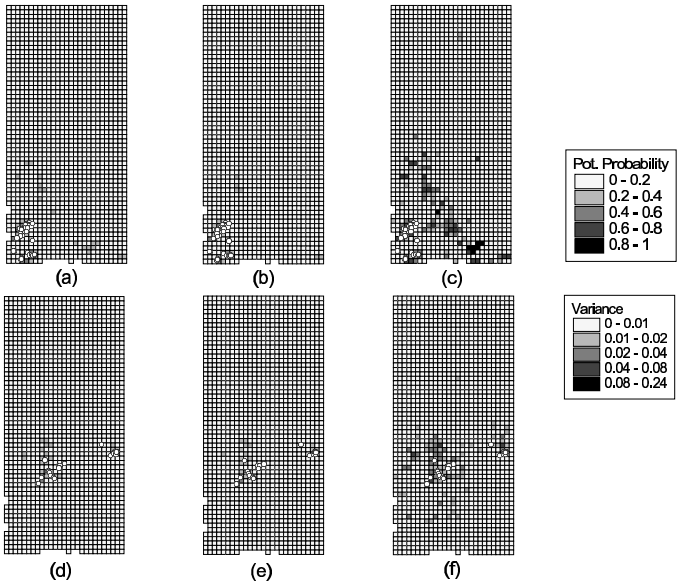

Figure R: The potential, "adjusted" predicted range of *Mimetes arboreus* (a)-(c) and *M. argenteus* (d)-(f) and the variance in the potential range. ((a) and (d) are potential range, (b) and (e) are adjusted range, and (c) and (f) are variance in potential range.)

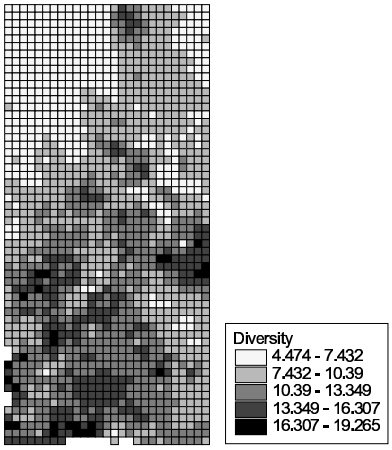

Figure S: Diversity Plot.

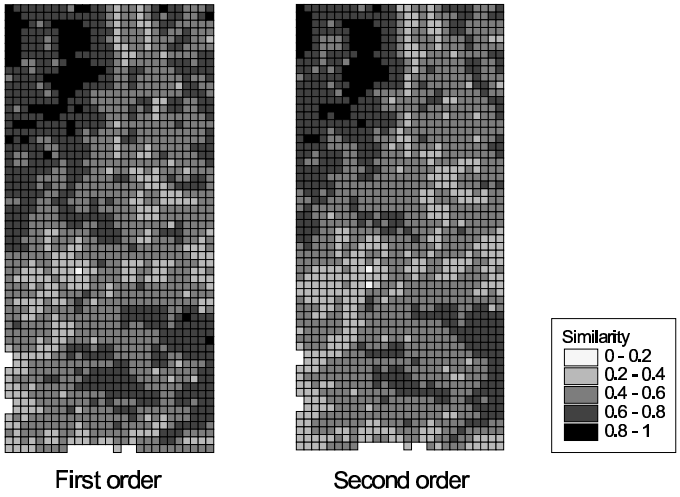

Figure T: Maps of the similarity of the $E(\mathbf{p}_i|\mathbf{Y})$ using first order neighbors.

Discussion to Figures U-W

Figure U shows the adjusted predicted range for *Mimetes arboreus* (U a) and *Mimetes argenteus* (U b) under the model augmented with (11). This follows nicely the discussion of Figure R above, where we mentioned that the predicted distribution of *Mimetes arboreus* picked up hints of the distribution of *Mimetes argenteus* to its west and north, a sort of predictive ecological ghosting for a closely related species. Figure V shows the presence/absence of the ancestor to species A and B. Figure W provides a vicariance promotion surface which shows the encouragement for species A and B respectively given AB was present.

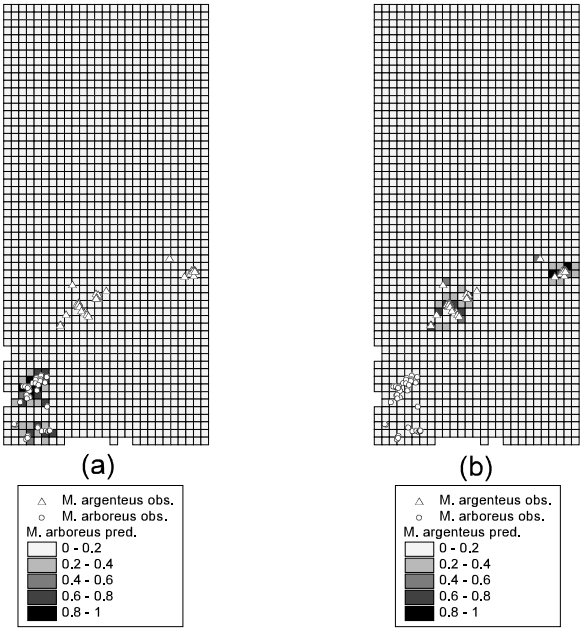

Figure U: Adjusted predicted ranges for *Mimetes arboreus* (a) and *M. argenteus* (b) using the augmented model (11).

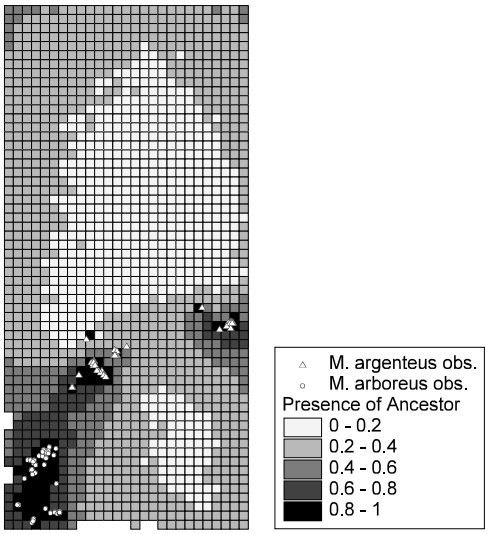

Figure V: Adjusted predicted ranges for ancestor of *Mimetes arboreus* and *M. argenteus*.

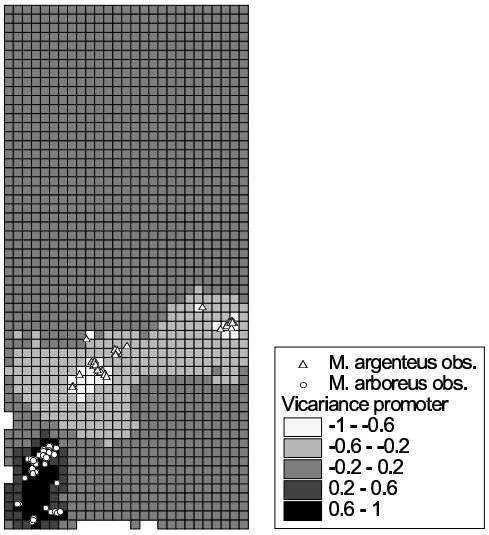

Figure W: Posterior mean vicariance promotion surface for Mimetes. See text for details.

Acknowledgments

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, Henri Laurie and Alexandra M. Schmidt for valuable discussions.