Journal of the
Royal Statistical Society

SERIES C
Applied
Statistics

Appl. Statist. (2011)
60, *Part* 5, *pp.* 757–776

Point pattern modelling for degraded presence-only data over large regions

Avishek Chakraborty,

Texas A&M University, College Station, USA

Alan E. Gelfand,

Duke University, Durham, USA

Adam M. Wilson,

University of Connecticut, Storrs, USA

Andrew M. Latimer

University of California at Davis, USA

and John A. Silander

University of Connecticut, Storrs, USA

[Received April 2010. Revised March 2011]

Summary. Explaining the distribution of a species by using local environmental features is a long-standing ecological problem. Often, available data are collected as a set of presence locations only, thus precluding the possibility of a desired presence–absence analysis. We propose that it is natural to view presence-only data as a point pattern over a region and to use local environmental features to explain the intensity driving this point pattern. We use a hierarchical model to treat the presence data as a realization of a spatial point process, whose intensity is governed by the set of environmental covariates. Spatial dependence in the intensity levels is modelled with random effects involving a zero-mean Gaussian process. We augment the model to capture highly variable and typically sparse sampling effort as well as land transformation, both of which degrade the point pattern. The Cape Floristic Region in South Africa provides an extensive class of such species data. The potential (i.e. non-degraded) presence surfaces over the entire area are of interest from a conservation and policy perspective. The region is divided into about 37000 grid cells. To work with a Gaussian process over a very large number of cells we use a predictive spatial process approximation. Bias correction by adding a heteroscedastic error component has also been implemented. We illustrate with modelling for six different species. Also, a comparison is made with the now popular Maxent approach though it is limited with regard to inference. The resultant patterns are important on their own but also enable a comparative view, for example, to investigate whether a pair of species are potentially competing in the same area. An additional feature of our modelling is the opportunity to infer about biodiversity through species richness, i.e. the number of distinct species in an areal unit. Such an investigation immediately follows within our modelling framework.

Keywords: Gaussian process; Intensity surface; Land use transformation; Non-homogeneous Poisson process; Predictive spatial process; Sampling bias; Spatial point pattern

Address for correspondence: Avishek Chakraborty, Department of Statistics, Texas A&M University, TAMU 3143, College Station, TX 77843, USA.
E-mail: avishekc@stat.tamu.edu

© 2011 Royal Statistical Society 0035–9254/11/60757

1. Introduction

Learning about distributions of species is a long-standing issue in ecology with, by now, an enormous literature. Useful review papers that organize and compare model approaches include Elith *et al.* (2006), Wisz *et al.* (2008) and references therein. Our focus here is on model-based approaches to study this problem. A substantial proportion of the model-based work focuses on modelling presence or absence where the data are available as a presence (1) or absence (0) at a collection of sampling locations. The goal is to explain the probability of presence at a location given the environmental conditions that are present there. The natural approach is to build a generalized additive binary regression model, with say a logistic link, where the covariates can be introduced linearly or as smoothly varying functions. Such generalized additive models tend to fit data well since they employ additional parameters to enable the response variables to assume non-linear and multimodal relationships with the data (Guisan *et al.*, 2002; Elith *et al.*, 2006). They can also provide a qualitative picture of how species respond to environmental variables. The price is that generalized additive models lose simplicity in interpretation and risk overfitting with poor out-of-sample prediction.

Much of this work is *non-spatial* in the sense that, though it includes spatial covariate information, it does not model anticipated spatial dependence in presence–absence probabilities. Accounting for the latter seems critical since causal ecological explanations such as localized dispersal as well as omitted (unobserved) explanatory variables with spatial pattern such as local smoothness of soil or topographic features suggest that, at sufficiently high resolution, occurrence of a species at one location will be associated with its occurrence at neighbouring locations (Ver Hoef *et al.*, 2001). In particular, such dependence structure, which is introduced through spatial random effects, facilitates learning about presence or absence for portions of a study region that have not been sampled, accommodating gaps in sampling and irregular sampling intensity. For point level categorical responses, Higgs and Hoeting (2010) used a Gaussian process (GP) prior for these spatial effects. For areal level count data, Markov random-field priors (Besag, 1974; Banerjee *et al.*, 2004) have been used in Augustin *et al.* (1996) and later incorporated into a hierarchical Bayesian model setting by Gelfand *et al.* (2005a, b) and Chakraborty *et al.* (2010). See also Latimer *et al.* (2006) in this regard.

The focus of the work here is on the so-called *presence-only* setting. Analysis of presence-only data has seen growing popularity in recent years due to increased availability of such records from museum databases and other non-systematic surveys; see Graham *et al.* (2004). A noteworthy point is that presence-only data are not *inferior* to presence–absence data. In fact, it is the converse; in principle, presence-only data offer a complete census whereas presence–absence data, since confined to a specified set of sampling sites, contain less information. One model-based strategy for presence-only data has attempted to implement a presence–absence approach. All of this work depends on drawing so-called *background samples*, which are random samples of locations in the region with known environmental features. Early work here characterized these samples as pseudoabsences (Engler *et al.*, 2004; Ferrier *et al.*, 2002) and fitted a logistic regression to the observed presences and these pseudoabsences. Since presence or absence is unknown for these samples, recent work (Pearce and Boyce, 2006; Ward *et al.*, 2009) shows how to adjust the resulting logistic regression to account for this. Additionally, all of this work is non-spatial in the sense of the previous paragraph. Perhaps, most importantly, as we argue below, this approach conditions in the wrong direction. We assert that the observed presences can be viewed as a *marked point pattern* with the mark indicating presence (see the recent work of Warton and Shepherd (2010) in this regard). We do not have a point pattern of absences; pseudoabsences create an unobserved and artificial pattern of absences.

Alternative algorithmic approaches include the genetic algorithm for rule-set prediction approach (Peterson and Kluza, 2003) and the maximum entropy approach Maxent; see, for example, Phillips *et al.* (2006, 2009). The genetic algorithm for rule-set prediction is based on an artificial intelligence framework to produce a set of positive and negative rules that, together, give a binary prediction. Rules are favoured according to their effectiveness (compared with random prediction) on the basis of a sample of background data and presence data. Maxent is a constrained optimization method which finds the optimal species density (closest to a uniform) subject to moment constraints. Maxent predictions have usually been found to have higher predictive accuracy on average than the genetic algorithm for rule-set prediction (Elith *et al.*, 2006). Moreover, with the availability of a fairly recent attractive software package (`http://www.cs.princeton.edu/schapire/maxent`), Maxent is now becoming the standard approach for presence-only data analysis. The point pattern analysis approach that we develop provides an appealing alternative in that it is fully model based, allowing full inference with associated uncertainty everywhere in the region.

We model presence-only data as a point pattern with associated intensity specified in terms of the available environments across the region, as in Warton and Sherpherd (2010). We do this through typical regression modelling, enabling a natural interpretation for the coefficients. We employ a hierarchical model to introduce spatial structure for the intensity surface through spatial random effects. We do not assume any background or pseudoabsence samples; rather, we assume that the covariates that we employ are available as surfaces over the region to interpolate an intensity over the entire region. We acknowledge that the observed point pattern is biased through anthropogenic processes, e.g. human intervention to transform the landscape, and non-uniform (in fact, often very irregular) sampling effort. This requires adjusting the *potential* species intensity to a *realized* intensity which we treat as a *degradation* of the former. The implications of such bias and the need for bias correction by using any of the presence-only analysis approaches has been discussed in the literature. See, for example, Phillips *et al.* (2009), Veloz (2009) and references therein. Lastly, an attractive by-product of our modelling is the opportunity to study species richness, i.e. the expected number of distinct species in a specified region. In particular, we can obtain potential and observed richness surfaces.

We work with presence-only data collected from the Cape Floristic Region (CFR) in South Africa (Fig. 1). The region is divided into approximately 37000 grid cells, each $1' \times 1'$ (roughly $1.55\text{ km} \times 1.85\text{ km}$). Covariate information is available only at grid cell level so we model the intensity as a tiled surface over these cells. We provide potential and degraded intensities for six species as well as richness distributions with respect to them (Section 4).

The format of the paper is as follows. Section 2 reviews the common issues that arise in modelling species distribution data sets. Section 3 develops a point process model for the presence-only data sets. Section 4 shows how we can study richness. Section 5 details the computational and inferential issues that are related to high dimensional spatial data. Section 6 compares our approach with the Maxent (Phillips and Dudík, 2008) method for synthetic data sets. In Section 7, we present the data analysis with interpretation and conclude with some discussion and future extensions in Section 8.

2. Basic issues and existing approaches

The simplest approaches to predicting species distributions based on presence-only data are based directly on the environmental envelope that is associated with observed occurrences. In these approaches, one summarizes the suite of environmental attributes of species site occurrences and extrapolates potential presence to other sites with similar attributes; there is no

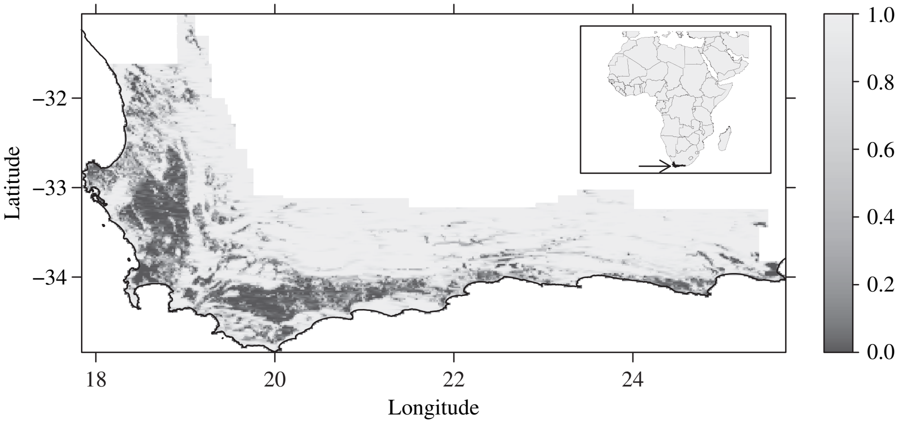

Fig. 1. CFR of South Africa: the inset shows the location of the CFR within the African continent; the 90 000 km^2 region was divided into 36 907 1′ cells; the proportion of untransformed land at the grid cell level is shown as well

spatial component to the prediction. The BIOCLIM and DOMAIN models use variations on this approach (Elith *et al.*, 2006). Typically, one draws convex hulls around sites of occurrence (perhaps with extrapolations) to identify a geographical region which is interpreted as the range for the species. This raises the question of whether we should require spatial contiguity for a species geographic range. Furthermore, from a stochastic perspective, do we insist on hard edges for ranges or, by thresholding probabilities, do we prefer *soft* ranges? Various alternative algorithmic approaches include the genetic algorithm for rule-set prediction (Stockwell and Peters, 1999), artificial neural nets (Lek *et al.*, 1996), boosted regression trees (also called stochastic gradient boosting; Elith *et al.* (2008)), random forests (Breiman, 2001) and climate envelope models (Heikkinen *et al.*, 2006).

Maxent is an algorithmic tool that produces a probability density surface which maximizes entropy given constraints that are imposed by the collection of vectors of environmental variable values at the sites at which the species has been observed. These constraints require that the average of each of the environmental covariates under this distribution *essentially* agrees with the empirical average for this covariate on the basis of samples over the region. The constrained optimization introduces regularization weights, one for each moment constraint. The optimization is solved only approximately, i.e. each constraint is satisfied within a specified precision to avoid overfitting. As an optimization strategy rather than a stochastic modelling approach, Maxent cannot attach any uncertainty to resulting optimized estimates. The resultant surface is interpreted as providing the relative probability of observing a species at a given location compared with other locations in the region. However, Maxent cannot provide an absolute intensity; hence, we cannot determine the expected number of individuals in a specified region. The approach that we propose below addresses all of these issues. However, in Section 5 we make comparison between Maxent and our approach within the limitations of Maxent. For the CFR data set, we present only the analysis under our approach.

As noted Section 1, a much different strategy introduces pseudoabsences to fit a binary regression model, most commonly a logistic regression, modelling the probability of presence given environmental covariates; see for example Ferrier *et al.* (2002) and Engler *et al.* (2004). More recent work (Pearce and Boyce, 2006; Ward *et al.*, 2009) acknowledges that presence or absence is unknown for these background samples and attempts to adjust the resulting logistic regression to account for this. Alas, this requires specifying the overall population prevalence of the species, a notion which, for a given region, is not well defined. In any event, both Pearce and Boyce (2006) and Ward *et al.* (2009) noted that this *marginal* prevalence will not be known and the latter clearly argued that estimating it from presence-only data is not feasible in practice. Furthermore, how are such background samples developed? Ideally, we seek a *random sample* of the available environment. It is not clear how to do this but it is clear that we do not want to draw locations uniformly from the study region. Moreover, it is obvious that inference depends on the number of background samples drawn, an arbitrary choice which can substantially influence the resultant inference (Pearce and Boyce (2006), page 407). Finally, despite the natural expectation that there should be spatial dependence in the presence–absence probabilities, none of this work employs spatial dependence structure.

Most of these approaches fail to address bias that may exist in sampling occurrences. Yet such bias in sampling is a common problem; see for example Loiselle *et al.* (2007) and references therein. Recently Diggle *et al.* (2010) addressed this issue, referring to it as preferential sampling. Variation in site access is one of the factors that influence the likelihood of the site to be sampled. For example, sites that are adjacent to roads or along paths, near urban areas, with public ownership or with flat topography are likely to be oversampled relatively to more inaccessible sites. When bias implies that only a portion of the region is sampled, perhaps only a portion of the overall point pattern is observed. In addition, there may be temporal bias in sampling. For example, as one learns more about the ecology of the species of interest sample site selection may change (Lobo *et al.*, 2007).

Land use, as a result of human intervention, affects *availability* of locations, and hence inference about the intensity. As a result of human intervention, areas within the study region are not available for a species. Also, agricultural transformation and dense stands of alien invasive species preclude availability. Fig. 1 shows the extent of transformation across the CFR at the grid cell level. Transformed areas are not sampled and so this information must be included in the modelling. Altogether, sampling is sparse and irregular. In fact, only 10158 of the 36907 (28%) grid cells were sampled (Fig. 2). It is unlikely that we have collected a random sample of available environments.

Detection can affect inference regarding the intensity, i.e. we may incorrectly identify a species as present which is actually absent (false presence) or fail to detect a species that is actually present (false absence) (Reese *et al.*, 2005). Evidently, the prevalence of these false records will affect

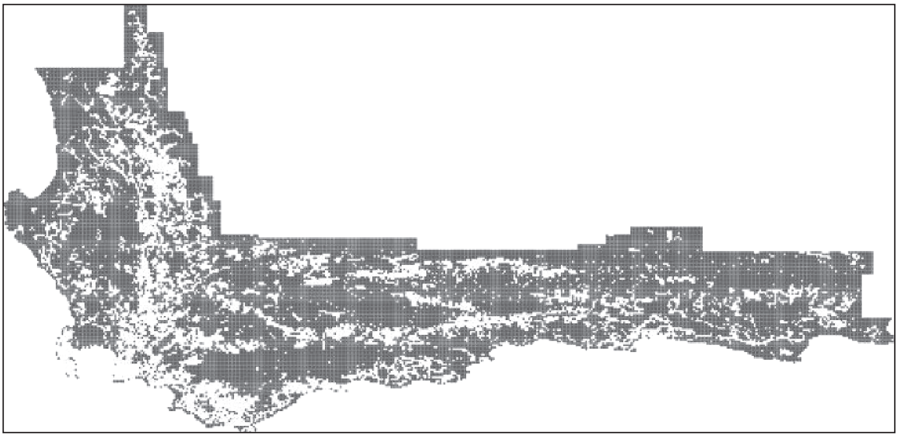

Fig. 2. Cells within the CFR that have at least one observation from the *Proteas Atlas* data set (□) and cells with no observations (■)

the attempt of an explanatory model on environmental features (Tyre *et al.*, 2003). Modelling for these errors can be attempted but requires information beyond the observed presence data.

3. Point process modelling

We view the observed presence-only data as a point pattern subject to degradation. In Section 3.1 we detail a general point process specification for this problem. In Section 3.2 we formalize the likelihood and posterior and propose a grid cell level approximation.

3.1. Probability model for presence locations

As is customary for point patterns, we assume a non-homogeneous Poisson process (NHPP) model (Van Lieshout, 2000), which was also referred to as an inhomogeneous Poisson process in Diggle (2003), for the set of presence locations. We must introduce degradation caused by sampling bias as well as by land transformation. As a result, we conceptualize a *potential* intensity, i.e. the intensity in the absence of degradation, as well as a *realized* (or effective) intensity that operates in the presence of degradation. Further, we tile the intensity to reflect our inability to explain it at spatial resolution finer than our grid cells.

We begin by imagining three surfaces over D. Let $\lambda(s)$ be the 'potential' intensity surface, i.e. a positive function which is integrable over D. $\lambda(s)$ is the intensity in the absence of degradation. Let $\int_D \lambda(s)\,\mathrm{d}s = \lambda(D)$. Then, $g(s) = \lambda(s)/\lambda(D)$ gives the potential density over D. Modelling for $\lambda(s)$ will be provided in Section 3.2. Next, we envision an availability surface $U(s)$, which is a binary surface over D such that $U(s) = 1$ or $U(s) = 0$ according to whether location s is untransformed by land use or not, i.e., assuming no sampling bias, $\lambda(s)\,U(s)$ can be only $\lambda(s)$ or 0 according to whether s is available or not. Let A_i denote the geographical region corresponding to cell i. Then, if we average $U(s)$ over A_i, we obtain $u_i = \int_{A_i} U(s)\,\mathrm{d}s/|A_i|$, where u_i is the proportion of cell i that is transformed and $|A_i|$ is the area of cell i. In our setting u_i is known, through remote sensing, for all grid cells. Similarly, we envision a sampling effort surface over D which we denote as $T(s)$. $T(s)$ is also a binary surface and $T(s)\,U(s) = 1$ indicates that location s is both available and sampled. Now, we can set $q_i = \int_{A_i} T(s)\,U(s)\,\mathrm{d}s/|A_i|$ and interpret q_i as the probability that a randomly selected location in A_i was available and sampled. Thus we can capture availability and sampling effort at areal unit scale.

Hence, $\lambda(s)\,U(s)\,T(s)$ becomes the degradation at location s. This implies that, in regions where no locations were sampled, the operating intensity for the species is 0. In this regard, we note that we do *not* envision a probability density surface for sampling effort as in the Maxent literature (Phillips and Dudík, 2008). Rather, $\int_{A_i} T(s)\,\mathrm{d}s/|A_i|$ can be viewed as the sampling probability that is associated with cell i. Then, if $T(s)$ is viewed as random, the expectation of the integral would yield $\int_{A_i} p(s)\,\mathrm{d}s/|A_i|$ where, now, $p(s) = P\{T(s) = 1\} \in [0, 1]$. Clearly, $p(s)$ gives the local probabilities of sampling: not a probability density over D.

To go forward, we assume that $\lambda(s)$ is independent of $T(s)\,U(s)$, i.e. the potential intensity for a species is independent of the degradation process. Then, omitting the details, we can write $\int_{A_i} \lambda(s)\,T(S)\,U(s)\,\mathrm{d}s = \lambda_i q_i$ where $\lambda_i = \int_{A_i} \lambda(s)\,\mathrm{d}s$ is the cumulative intensity that is associated with cell A_i and, again,

$$q_i = \frac{1}{|A_i|} \int_{A_i} T(s)\,U(s)\,\mathrm{d}s.$$

It is not sensible to imagine that sampling effort is independent of land transformation. In fact, we might expect less sampling attention to be paid to more transformed areas (Reese *et al.*, 2005; Veloz, 2009). More directly, if $U(s) = 0$ then $T(s) = 0$. Hence, if we define $q_i = u_i p_i$, then

$$p_i = \int_{A_i} T(s)U(s)\mathrm{d}s \bigg/ \int_{A_i} U(s)\mathrm{d}s,$$

i.e. p_i is the conditional probability that a randomly selected location in cell i is sampled given that it is available. In our application below we set p_i equal to 1 or 0 which we interpret as $T(s)=U(s)\ \forall s \in A_i$ or $T(s)=0\ \forall s \in A_i$ respectively. In particular, we set $p_i=1$ if cell i was sampled for any species in our data set; otherwise, we set $p_i=0$. For the CFR, this sets $p_i=1$ for the 10158 grid cells that have been visited.

3.2. Likelihood and posterior

Now we turn to modelling of the potential intensity surface $\lambda(\cdot)$. We employ a GP prior, which results in a log-Gaussian Cox process model (Møller and Waagepetersen (2004), section 5.6) for the observed data. We expect the environmental covariates, say $\mathbf{x}(s)$, to influence the intensity and model the mean of the GP as a linear combination of them. Then for any location $s \in D$, we have

$$\log\{\lambda(s)\} = \mathbf{x}^{\mathrm{T}}(s)\beta + w(s) \tag{1}$$

with $w(\cdot)$, a zero-mean stationary, isotropic GP over D, to capture residual spatial association in the λ-surface across grid cells. The Matérn family of covariance functions provides a flexible class of isotropic dependence; in what follows we use the special case of the exponential covariance function.

As above, suppose that we have n_i presence locations $(s_{i,1}, s_{i,2}, \ldots, s_{i,n_i})$ within cell i for $i=1,2,\ldots,I$. Following the discussion in Section 3.1, $U(s_{i,j})\ T(s_{i,j}) \equiv 1,\ 0 \leqslant j \leqslant n_i, 1 \leqslant i \leqslant I$. Then the likelihood function corresponding to NHPP$\{\lambda(\cdot)\}$ becomes

$$L[\lambda(\cdot); \{s_{i,j}\}] \propto \exp\left\{-\int_D \lambda(s)U(s)T(s)\mathrm{d}s\right\} \prod_{i=1}^{I} \prod_{j=1}^{n_i} \lambda(s_{i,j}). \tag{2}$$

Although we have only finitely many presence locations, the integral term in L involves the uncountable random field $\{\lambda(s): s \in D\}$. Fortunately, we have a natural approximation by recalling that the data set is gathered at the scale of grid cells in the CFR, i.e. though we have geocoded locations for the observed sites, with covariate information at grid cell level, we attempt to explain only the point pattern at grid cell level. In particular, let D denote our CFR study domain where D is divided into $I=36907$ grid cells of equal area. For each cell $i=1,2,3,\ldots,I$, we are given information on l covariates as $x_i=(x_{i1}, x_{i2}, \ldots, x_{il})$. We also have cell level information about availability of land across D, as a proportion of the area of the cell (Fig. 1). Following the previous subsection, we denote this by u_i. For many cells $n_i=0$ primarily because 72% were actually unsampled. Additionally, a computational advantage accrues to working at grid cell level; we can work with a product Poisson likelihood approximation rather than the point pattern likelihood in expression (2), i.e. we assume that $\lambda(\cdot)$ is a tiled surface such that, for cell i, the height is $\Delta\lambda(s_i)$ where Δ is the area of the cell and s_i is the centroid. Then, given the set $\{\lambda(s_i), i=1,2,\ldots,I\}$, the n_i are independent and $n_i \sim \mathrm{Po}\{\Delta\lambda(s_i)q_i\}$. Approximation of the point pattern likelihood by using a *tiled* surface over a lattice embedding the region was discussed in Beneš *et al.* (2003). The approximation was justified in the sense that the resulting approximate posterior converges to the true posterior as the partition becomes finer.

Note that, for any cell with $q_i=0$ (which can happen if either $p_i=0$ or $u_i=0$), there is no contribution from A_i in the product Poisson likelihood. Since, from equation (1), $\log\{\lambda(s)\}$ follows a GP, the posterior distribution takes the form

$$\pi\{\lambda(\mathbf{s}_{1:m}),\beta,\theta|\mathbf{n},\mathbf{x},\mathbf{u},\mathbf{q}\}\propto\exp\left\{-\sum_{i=1}^{I}\lambda(s_i)\Delta_i q_i\right\}\prod_{i=1}^{m}\lambda^{n_i}(s_i)$$
$$\times\phi_m[\log\{\lambda(\mathbf{s}_{1:m})\}|\beta,\mathbf{x},\theta]\,\pi(\beta)\,\pi(\theta) \qquad (3)$$

where ϕ_m denotes the m-dimensional Gaussian density and θ the parameters in the covariance function of $w(\cdot)$ in equation (1). Sampling from expression (3) by using Markov chain Monte Carlo (MCMC) methods is discussed in Section 5.2 and in Appendix A.

We conclude this section by noting potential interest in establishing the consistency of the posterior (3), i.e. convergence of the modelled posterior to the true posterior. This involves proving that, under a log-Gaussian process model, $\lambda(s)\rightarrow\lambda_0(s)$, the true intensity, in supremum norm. Conditions and the associated argument are mentioned in a file of supplementary information associated with the paper on the publisher's Web site.

4. Studying richness with presence-only data

Recall the definition of species richness for a specified region. Relative to a specified set of species, the observed richness is the number of distinct species that are found in that region. Here, we show how our modelling above provides a parametric function for expected richness. By comparison, an often-used approach with Maxent is merely to sum over the individual species densities (Newbold *et al.*, 2009). The interpretation of such a sum as a richness when integrated over a subregion is possibly unsatisfying and, in any case, no uncertainty can be attached to any estimates that are made by using this sum.

Under the presence-only setting, we imagine that data arrive in the form $(s_j, l(s_j))$, $j=1,2,\ldots,n$, i.e. a random location and a species label associated with that location. Suppose that we use the foregoing modelling to create a species intensity $\lambda_l(s)$ for species $l=1,2,\ldots,L$. For a set A within the study region, we define the richness for A to be the expected number of distinct species in A. Under this definition, we expect more species as A grows larger and no species as the area of A goes to 0.

Let $n(A)$ be the total number of observations in A, i.e. the total number of locations in A where a 'presence-only' observation of any species was recorded. Let $n_l(A)$ be the number of locations in A where species l was observed. Finally, let $r(A)=\Sigma_l\mathbf{1}\{n_l(A)>0\}$, where $\mathbf{1}(\cdot)$ is the indicator function. Then $r(A)$ is the 'realized' richness in A. Thus, the quantity that we seek to infer about is $E\{r(A)\}$. Note that $E[\mathbf{1}\{n_l(A)>0\}]=1-\exp\{-\lambda_l(A)\}$ since $n_l(A)\sim\text{Po}\{\lambda_l(A)\}$. Hence,

$$E\{r(A)\}=\sum_l[1-\exp\{-\lambda_l(A)\}].$$

Evidently, richness is not additive, i.e.

$$E\{r(A_1)\cup r(A_2)\}\neq E\{r(A_1)\}+E\{r(A_2)\}.$$

With model fitting for each $\lambda_l(s)$, we can obtain posterior samples of $E\{r(A)\}$ for any A by obtaining posterior samples for each $\lambda_l(A)$. Such samples are obtained through appropriate integration (summation for the Poisson approximation version) of $\lambda_l(s)$ over A. If we work with the collection of grid cells A_i, we can supply a richness surface for the entire CFR. Adjustment for transformation and sampling intensity can be introduced as above to distinguish a potential and degraded surface. We illustrate this in Section 7.

5. Computation and inference

We fit the models of the previous section by using MCMC sampling. The primary computational challenge in working with the CFR is handling the model fitting for 37 000 grid cells, which is the familiar 'large *n*' problem for GPs; see Banerjee *et al.* (2004). We employ the predictive process approximation for Gaussian random fields (Banerjee *et al.*, 2008). With grid cells, an alternative is a parallelization scheme in conjunction with a conditional auto-regressive model as described in Chakraborty *et al.* (2010).

5.1. *Predictive process approximation*

In the context of MCMC sampling, employing a GP on a large collection of locations is computationally demanding, because of the necessary repeated inversion of the covariance matrix arising from the process. There are several approximation techniques in the literature, such as process convolution (Higdon, 2002), approximate likelihood (Stein *et al.*, 2004) and fixed rank kriging (Cressie and Johannesson, 2008). The *predictive process* method (Banerjee *et al.*, 2008) accommodates a high dimensional GP as follows. If $w(\cdot)$ is the zero-mean GP under consideration, and our data consist of locations $\mathbf{s}_{1:I}=(s_1,s_2,\ldots,s_I)$ where I is large, then the method proceeds by first choosing r locations $\mathbf{s}^0_{1:r}=(s^0_1,s^0_2,\ldots,s^0_r)$ from the region, called *knots*, and then replaces $w(\mathbf{s}_{1:I})$ in the model equation by $\tilde{w}(\mathbf{s}_{1:I})=E\{w(\mathbf{s}_{1:I})|w(\mathbf{s}^0_{1:r})\}=L\,w(\mathbf{s}^0_{1:r})$ where the matrix L is calculated from the dependence structure of $w(\cdot)$. L depends on correlation parameters but not on the process variance. In our setting, we apply this approximation to the $\{\lambda(s_j)\}$ in distribution (3) through the $\{w(s_j)\}$.

We introduce bias correction, which is a modification that was discussed in Finley *et al.* (2009). Since $\text{var}\{w(\mathbf{s}_j)\}\geqslant\text{var}\{\tilde{w}(\mathbf{s}_j)\}$ for each j, the predictive process is expected to underestimate the spatial variance and to increase the variance of the nugget term. The correction introduces a heteroscedastic error ε^* with $\text{var}(\varepsilon^*_j)=\text{var}\{w(\mathbf{s}_j)\}-\text{var}\{\tilde{w}(\mathbf{s}_j)\}$. No additional parameters are brought in with this correction so we retain the benefit of a lower dimensional spatial association structure. The computational advantage of this method is illustrated in Section 5.3 and Appendix A.

5.2. *Markov chain Monte Carlo sampling*

In Section 3.2, we supplied the likelihood and posterior under our model. In what follows, we approximate expression (3), employing the predictive process technique that was discussed above. The joint set of locations $(\mathbf{s}_{1:I},\mathbf{s}^0_{1:r})$ partition the spatial covariance matrix as

$$\sigma^2 R_{n+r}(\phi)=\sigma^2\begin{pmatrix} R_I(\phi) & R_{r,I}(\phi)\\ R_{I,r}(\phi) & R_r(\phi)\end{pmatrix},$$

where the entries of R_{r+I} are exponential correlation terms with decay parameter ϕ. We rewrite $\Lambda_{0,i}=\lambda(s_i)\Delta$, which denotes the expected species count in cell i under potential prevalence. Now the hierarchical model looks like

$$\left.\begin{aligned} n_i|\Lambda_{0,i}&\overset{\text{ind}}{\sim}\text{Poi}(\Lambda_{0,i}q_i), \qquad\qquad i=1,2,\ldots,I,\\ \log\{\lambda(\mathbf{s}_i)\}&=x_i^{\mathrm{T}}\beta+\tilde{w}(\mathbf{s}_i)+\varepsilon^*_i,\\ \tilde{w}(\mathbf{s}_{1:I})&=R_{I,r}(\phi)\,R_r^{-1}(\phi)\,w(\mathbf{s}^0_{1:r}),\\ w(\mathbf{s}^0_{1:r})&\sim N_r\{\mathbf{0}_r,R_r(\phi)\},\\ \varepsilon^*_i&\overset{\text{ind}}{\sim}N[0,\sigma^2\{1-(R_{I,r}(\phi)\,R_r^{-1}(\phi)\,R_{r,I}(\phi))_{ii}\}],\\ \pi(\beta,\phi,\sigma^2)&=\pi(\beta)\,\pi(\phi)\,\pi(\sigma^2).\end{aligned}\right\}\tag{4}$$

From Section 5.1, $L(\phi)=R_{I,r}(\phi)R_r^{-1}(\phi)$. In the absence of prior knowledge, we can use weak prior distributions for β and σ^2 as Gaussian and inverse gamma respectively. A common issue in spatial modelling is to identify σ^2 and ϕ simultaneously (Zhang, 2004). For the exponential covariance function, we have $\phi \approx 3/d$, where d is the spatial range, i.e. the distance after which spatial association between pairs of sites falls below 0.05. With a vague prior for σ^2, we must be informative about the support of d. In implementing the MCMC algorithm, β, σ^2 and $w(\mathbf{s}^0_{1:r})$ were updated by using Gibbs steps whereas, for $\lambda(\mathbf{s}_{1:I})$ and ϕ, Metropolis–Hastings steps need to be used. Alternatively, $\lambda(\mathbf{s}_{1:I})$ can be updated by using slice sampling. The posterior uncertainty of any λ_i goes down as q_i approaches 1. The acceptance rate for ϕ merits attention, since a rate above 40% or so pushes the posterior mode to one end of the prior support (indicating the need to readjust the support), whereas a rate less than 10% suggests that the proposal interval is too wide. So, tuning is needed but, once an effective choice has been made, the data could identify a unimodal posterior for ϕ within that support. To compute the Metropolis–Hastings ratio for ϕ, we can use the Sherman–Woodbury–Morrison formula (Harville, 1997) to invert the I-dimensional covariance matrix, utilizing the r-dimensional dependence in the spatial part. Computational details as well as the posterior full conditionals for model parameters are provided in Appendix A.

5.3. Posterior inference

As mentioned before, the two principal objectives of this data analysis are to understand the effect of environmental variables on species distribution and, more importantly, to construct maps of the potential and realized intensities over the entire study region. Posterior samples of β help us to infer whether a particular factor has a significant effect (positive or negative) on species intensity. The ϕ-parameter indicates the strength of spatial association between neighbouring cells after adjusting for covariate effects. A large value of ϕ implies rapid decay in such association, whereas a small value generally highlights the usefulness of a spatial random effect in the intensity function. This association may arise because some potentially important covariates are not available or because the effect of the covariate is not well captured by using a linear form. However, since Gaussian processes can capture a wide range of dependences, using them in a hierarchical setting enhances predictive performance for the model.

If, out of I cells, only m cells were sampled and contributed to the model fitting then inference for the remaining $I-m$ cells is done from their posterior predictive distributions. The foregoing predictive process approximation yields

$$\begin{aligned} \log\{\lambda(\mathbf{s}_{m+1:I})\} &= \mathbf{x}(\mathbf{s}_{m+1:I})\beta + R_{I-m,r}(\phi)\, R_r^{-1}(\phi) w(\mathbf{s}^0_{1:r}) + \varepsilon^*_{m+1:I}, \\ \varepsilon^*_i &\overset{\text{ind}}{\sim} \phi_{I-m}[0, \sigma^2\{1-(R_{I-m,r}(\phi) R_r^{-1}(\phi)\, R_{r,I-m}(\phi))_{ii}\}], \qquad m < i \leqslant I. \end{aligned} \tag{5}$$

So, conditional on posterior samples of β, ϕ and $w(\mathbf{s}^0_{1:r})$, we can draw samples from the posterior predictive distribution of $\log\{\lambda(\mathbf{s}_{m+1:I})\}$, independent of $\log\{\lambda(\mathbf{s}_{1:m})\}$, owing to the independence between ε^*s across sites. This is computationally very efficient, as independence also across components of $\lambda(\mathbf{s}_{m+1:I})$ (conditional on process parameters) ensures that we do not have to draw from a high dimensional multivariate Gaussian distribution, even if we want to predict the intensity surface at thousands of unsampled sites.

With regard to displays of intensity surfaces, since, in our CFR application, $p_i = 1$ (i.e. $T(s) = U(s)$ for all s in cell i) or $p_i = 0$ (i.e. $T(s) = 0$ for all s in cell i) and since only 28% of cells were sampled, the $\lambda_i p_i$-surface will be 0 for 72% of cells, primarily capturing the (lack of) sampling effort. The $\lambda_i u_i$-surface reveals the effect of transformation. Since few cells are completely transformed, most $\lambda_i u_i > 0$. Of course, the $\lambda(s)$ surface is most interesting since it offers insight into the expected pattern of presences over all of D. Posterior draws of $\lambda_{1:I}$ can be used to infer about the potential intensity, displaying say the posterior mean surface. We can also learn about the potential density g (Section 3.1) in this discretized setting as $g_i = \lambda_i/\Sigma_{k=1}^{I}\lambda_k$, and the corresponding density under transformation as $g_{u,i} = \lambda_i u_i/\Sigma_{k=1}^{I}\lambda_k u_k$. Corresponding posterior summaries for the CFR application are shown in Section 7.

6. Comparison with Maxent

Since Maxent has emerged as the current approach of choice for handling presence-only data, we attempt a comparison of performance with our methodology. Owing to the limitations in inference that are available under Maxent, which were detailed in Section 2, we confine ourselves to a comparison within the capabilities of Maxent.

We offer a simulation study comparing the performance of Maxent and our method under several scenarios. First, we do an *error analysis* under both models with the data generated by using a log-GP point pattern. We compare performance under varying σ^2 and ϕ by using three different values for each (Table 1). We first experimented with a moderate-size data set on 225 grid cells such that exact GP computation can be performed. To match the number of cells for the application in Section 7, we also worked with a simulated data set on 5625 cells, employing the predictive process approximation of Section 5.1. The goodness of fit was measured by

(a) the usual mean-square error

$$\frac{1}{n}\sum(g_{\text{true},i} - g_{\text{est},i})^2$$

(where the g_is were defined in Section 5.3) and by

(b) the commonly used loss function for probabilities,

$$\frac{1}{n}\sum\frac{(g_{\text{true},i} - g_{\text{est},i})^2}{g_{\text{true},i}(1 - g_{\text{true},i})}.$$

We shall refer to these as ‘loss 1’ and ‘loss 2’ respectively in the following tables.

To describe the details of simulation, let us start with a rectangle $D = [0, 3.4] \times [0, 3.4]$ as the event region. On D, we construct three covariate surfaces x_1, x_2 and x_3 such that at a location $s = (s_1, s_2) \in D$

$$\begin{aligned} x_1(s) &= -1 + 0.7s_1 - 0.5\exp(-0.5s_2) - 2\sin(3s_1) + 0.1\,N(0, 1),\\ x_2(s) &= -1 - 0.55s_1 s_2 + 0.4\,N(0, 1),\\ x_3(s) &= -2 - 0.7s_1^{1.5} + \log(0.5 + s_1^{0.8}) + 0.68s_2 + 0.2\,N(0, 1). \end{aligned}$$

For well-behavedness we rescale all covariate surfaces to [0, 1]. Then, over D, we generate a zero-mean spatial random field w with scale σ^2 and range ϕ for the exponential covariance function varying as described above. The within-cell homogeneous intensity surface using $\lambda(\cdot)$ evaluated at the cell centres was created as in equation (1). We simulated a point pattern realization with the intensity surface $\lambda(s)$ each time. The results are summarized in Table 1.

Next, to enrich our comparison, we analyse relative performance under *misspecification* of the model. Typically, environmental factors will be influencing the presence intensity which we do not know or cannot observe. So, the set of covariates that is used for model fitting excludes such factors. For the 225 grid cells setting, we employ three covariate surfaces x_1, x_2 and x_3, but do *not* use any spatial random effects (ws) in the simulation.

For analysis, we consider six models, excluding the complete and the null specifications. In each case, we do two types of performance analysis using both Maxent and our approach. First, we assume that all the cells are exhaustively sampled, and we use the full set of points for

Table 1. Comparison between Maxent and GP regression under two loss functions

Spatial effect parameters		*Results for the following sample sizes:*							
		$n = 225$				$n = 5625$			
Scale	*Range*	*Loss 1*		*Loss 2*		*Loss 1*		*Loss 2*	
		GP	*Maxent*	*GP*	*Maxent*	*GP*	*Maxent*	*GP*	*Maxent*
0.09	0.6	3.388×10^{-7}	2.280×10^{-6}	7.161×10^{-5}	3.452×10^{-4}	1.922×10^{-9}	3.414×10^{-9}	8.480×10^{-6}	1.639×10^{-5}
	2.8	3.142×10^{-7}	1.108×10^{-6}	6.159×10^{-5}	1.791×10^{-4}	7.457×10^{-10}	1.517×10^{-9}	3.412×10^{-6}	7.937×10^{-6}
	4.5	2.890×10^{-7}	7.838×10^{-7}	5.601×10^{-5}	1.331×10^{-4}	5.741×10^{-10}	1.076×10^{-9}	2.673×10^{-6}	5.592×10^{-6}
1.6	0.6	3.681×10^{-7}	9.407×10^{-5}	7.235×10^{-5}	1.249×10^{-2}	6.682×10^{-9}	8.973×10^{-8}	2.838×10^{-5}	6.780×10^{-4}
	2.8	1.234×10^{-6}	2.947×10^{-5}	2.802×10^{-4}	4.278×10^{-3}	5.972×10^{-9}	3.414×10^{-8}	1.909×10^{-5}	1.837×10^{-4}
	4.5	1.663×10^{-6}	1.506×10^{-5}	3.020×10^{-4}	2.665×10^{-3}	4.277×10^{-9}	2.361×10^{-8}	1.380×10^{-5}	1.080×10^{-4}
3.6	0.6	9.333×10^{-8}	3.730×10^{-4}	3.569×10^{-5}	9.227×10^{-2}	2.987×10^{-9}	3.250×10^{-7}	1.679×10^{-5}	6.241×10^{-3}
	2.8	1.609×10^{-6}	1.161×10^{-4}	3.013×10^{-4}	1.578×10^{-2}	6.298×10^{-9}	1.039×10^{-7}	1.883×10^{-5}	6.730×10^{-4}
	4.5	4.471×10^{-6}	5.214×10^{-5}	7.093×10^{-4}	8.675×10^{-3}	6.655×10^{-9}	7.017×10^{-8}	1.841×10^{-5}	2.977×10^{-4}

Table 2. Comparison between Maxent and GP regression for models with different subsets of covariates

Variable subset	*Results for exhaustive sampling*				*Results for biased sampling*			
	Loss 1		*Loss 2*		*Loss 1*		*Loss 2*	
	GP	*Maxent*	*GP*	*Maxent*	*GP*	*Maxent*	*GP*	*Maxent*
x_1	2.428×10^{-6}	5.919×10^{-6}	4.756×10^{-4}	2.133×10^{-3}	7.725×10^{-7}	5.006×10^{-6}	1.489×10^{-4}	1.738×10^{-3}
x_2	2.637×10^{-6}	5.999×10^{-6}	5.214×10^{-4}	1.449×10^{-3}	1.017×10^{-6}	5.649×10^{-6}	2.233×10^{-4}	1.402×10^{-3}
x_3	2.633×10^{-6}	5.143×10^{-6}	5.395×10^{-4}	1.443×10^{-3}	1.335×10^{-6}	4.678×10^{-6}	2.902×10^{-4}	1.196×10^{-3}
x_2, x_3	2.548×10^{-6}	4.428×10^{-6}	5.127×10^{-4}	9.389×10^{-4}	9.029×10^{-7}	4.194×10^{-6}	2.077×10^{-4}	8.904×10^{-4}
x_1, x_3	2.304×10^{-6}	2.976×10^{-6}	4.539×10^{-4}	1.041×10^{-3}	6.554×10^{-7}	2.864×10^{-6}	1.246×10^{-4}	8.678×10^{-4}
x_1, x_2	2.296×10^{-6}	3.570×10^{-6}	4.377×10^{-4}	9.480×10^{-4}	4.690×10^{-7}	3.616×10^{-6}	1.135×10^{-4}	9.990×10^{-4}

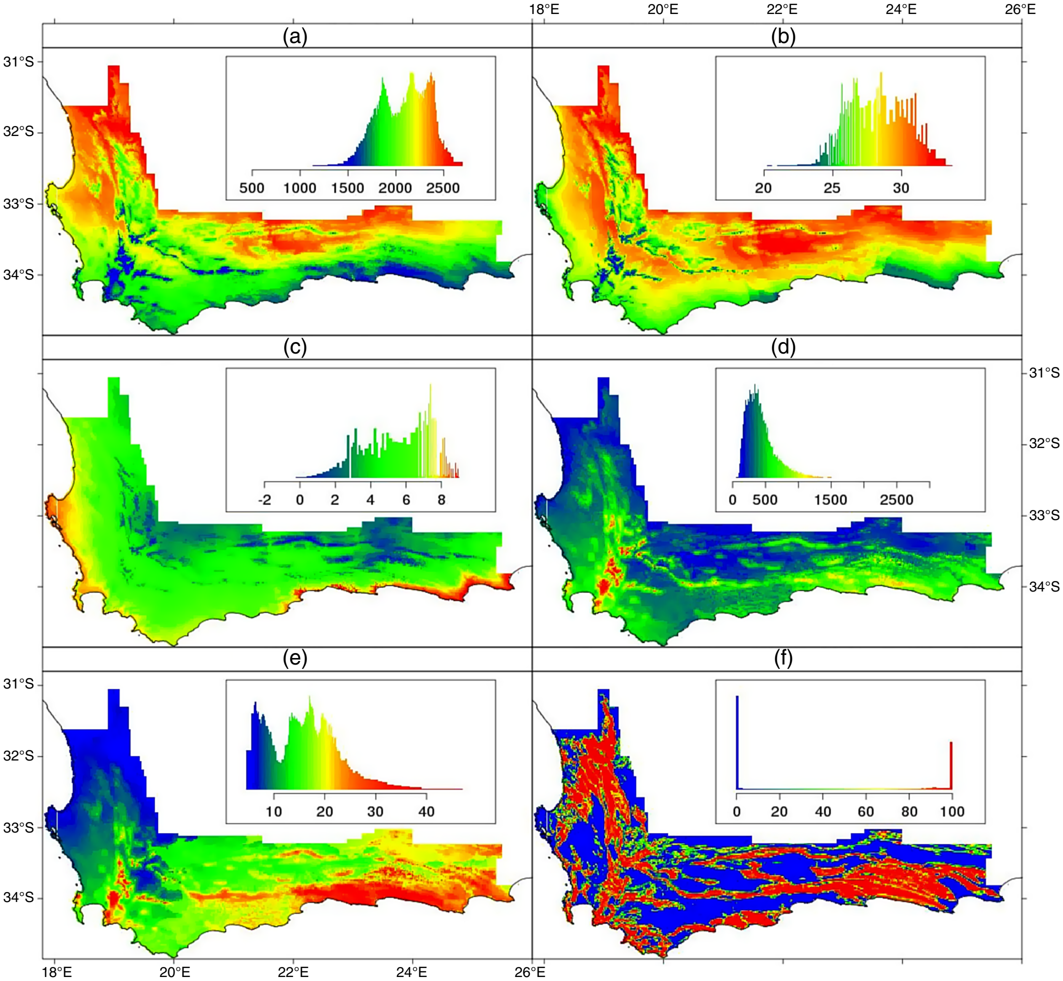

Fig. 3. Maps for all six covariate surfaces over the CFR: (a) mean annual precipitation; (b) July (winter) minimum temperature; (c) January (summer) maximum temperature; (d) potential evapotranspiration; (e) summer soil moisture days; (f) percentage of the grid cell with low fertility soil

inference and comparison. Then, we use a biased sampling approach assuming a known 0–1 sampling scheme, with 75 out of 225 cells unsampled and locations from those cells not considered in model fitting. In this case, we compare the performance under loss 1 and loss 2 only for the prediction for unsampled cells.

Table 2 summarizes the results for models employing various subsets of variables. Again, we see substantial improvement in predictive performance by using our modelling approach. With loss 1, at the least, we see a gain of 45% and in the best case more than 85%. With loss 2 the gains are even greater. We do relatively better with the biased sampling setting than the exhaustive sampling case. Finally, we do relatively better with smaller models than with larger models.

7. Analysing the Cape Floristic Region data

The CFR is the smallest of the world's six floral kingdoms, encompassing a small region of

Table 3. Posterior mean of covariate effects with central 95% credible interval in parentheses

Species	*EVAP*	*MAX01*	*MIN07*	*MAP*	*SMDSUM*	*FERT*
PRAURE	−4.909	2.702	−0.301	−1.222	−0.049	0.501
	(−6.506,−3.057)	(1.574,3.678)	(−0.967,0.425)	(−1.975,−0.423)	(−0.816,0.711)	(0.034,0.967)
PRCYNA	−2.447	1.268	−1.032	−0.833	0.552	0.802
	(−2.981,−1.859)	(0.853,1.619)	(−1.314,−0.469)	(−1.021,−0.626)	(0.255,0.830)	(0.642,0.957)
LDSG	0.721	−0.420	0.137	−0.376	0.488	0.099
	(0.373,1.085)	(−0.658,−0.181)	(−0.011,0.295)	(−0.513,−0.237)	(0.304,0.673)	(0.045,0.152)
PRMUND	−0.219	0.028	−0.609	−0.199	1.082	0.507
	(−2.724,1.429)	(−1.163,1.702)	(−1.039,−0.055)	(−1.024,0.510)	(0.277,1.809)	(0.101,0.929)
PRPUNC	2.076	−1.590	−1.722	0.363	0.535	0.186
	(1.031,3.096)	(−2.290,−0.921)	(−2.048,−1.409)	(0.082,0.662)	(0.052,1.079)	(−0.014,0.381)
PRREPE	1.690	−1.205	−0.275	0.124	0.094	0.224
	(1.243,2.124)	(−1.498,−0.907)	(−0.431,−0.110)	(−0.011,0.278)	(−0.112,0.320)	(0.152,0.295)

south-western South Africa, about 90000 km^2, including the Cape of Good Hope. It offers high levels of plant species diversity (9000 plant species) and endemism (69% found nowhere else). The plant diversity in the CFR is concentrated in relatively few groups, like the icon flowering plant family of South Africa, the *Proteaceae*. We consider six species within this family. Our point pattern for each species is drawn from the *Protea Atlas* data set (Rebelo, 2002). They are *Protea aurea*, PRAURE, at 603 locations, *Protea cynaroides*, PRCYNA, at 8172 locations, *Leucadendron salignum*, LDSG, at 22 949 locations, *Protea mundii*, PRMUND, at 764 locations, *Protea punctata*, PRPUNC, at 2148 locations and *Protea repens*, PRREPE, at 14 574 locations.

In earlier work (Gelfand *et al.*, 2005a, b) 18 environmental explanatory variables were considered, which were available at a minimum pixel resolution of 1′ latitude by 1′ longitude. On the basis of these analyses we have chosen the six most important as covariates for our intensity function. They are mean annual precipitation MAP, July (winter) minimum temperature MIN07, January (summer) maximum temperature MAX01, potential evapotranspiration EVAP, summer soil moisture days SMDSUM, and percentage of the grid cell with low fertility soil, FERT, with associated surfaces in Fig. 3.

We implemented the modelling in Section 3 on the CFR presence-only data for the six species above over the whole CFR. We centred and scaled all the xs. The very large data sets were handled efficiently by using C++ with the Intel mathematics kernel library (`http://software.intel.com/en-us/intel-mkl/`). The outputs that are presented below are created by first running 15000 MCMC iterations, discarding the initial 5000 samples and thinning the rest at every fifth sample. Summary output from the model fitting is presented through the following table and diagrams. Table 3 provides the posterior mean covariate effects for all species along with the associated 95% equal tail credible intervals in parentheses. Most of the coefficients are significantly different from 0 and, also, the direction of significance varies with species.

Together, Figs 4 and 5 show the posterior mean intensity surfaces (potential and transformed) for the six species. Evidently there is strong spatial pattern and the pattern varies with species, i.e. the nature of local adjustment to the regression is species dependent. A comparison between the transformed and potential for each species is illuminating. Differentials of multiple orders of magnitude in expected cell counts are seen across many grid cells. Finally, Fig. 6 shows the transformed and potential richness surfaces at grid cell level utilizing the six species. Admittedly, these displays are primarily illustrative; in practice, we would investigate richness with regard

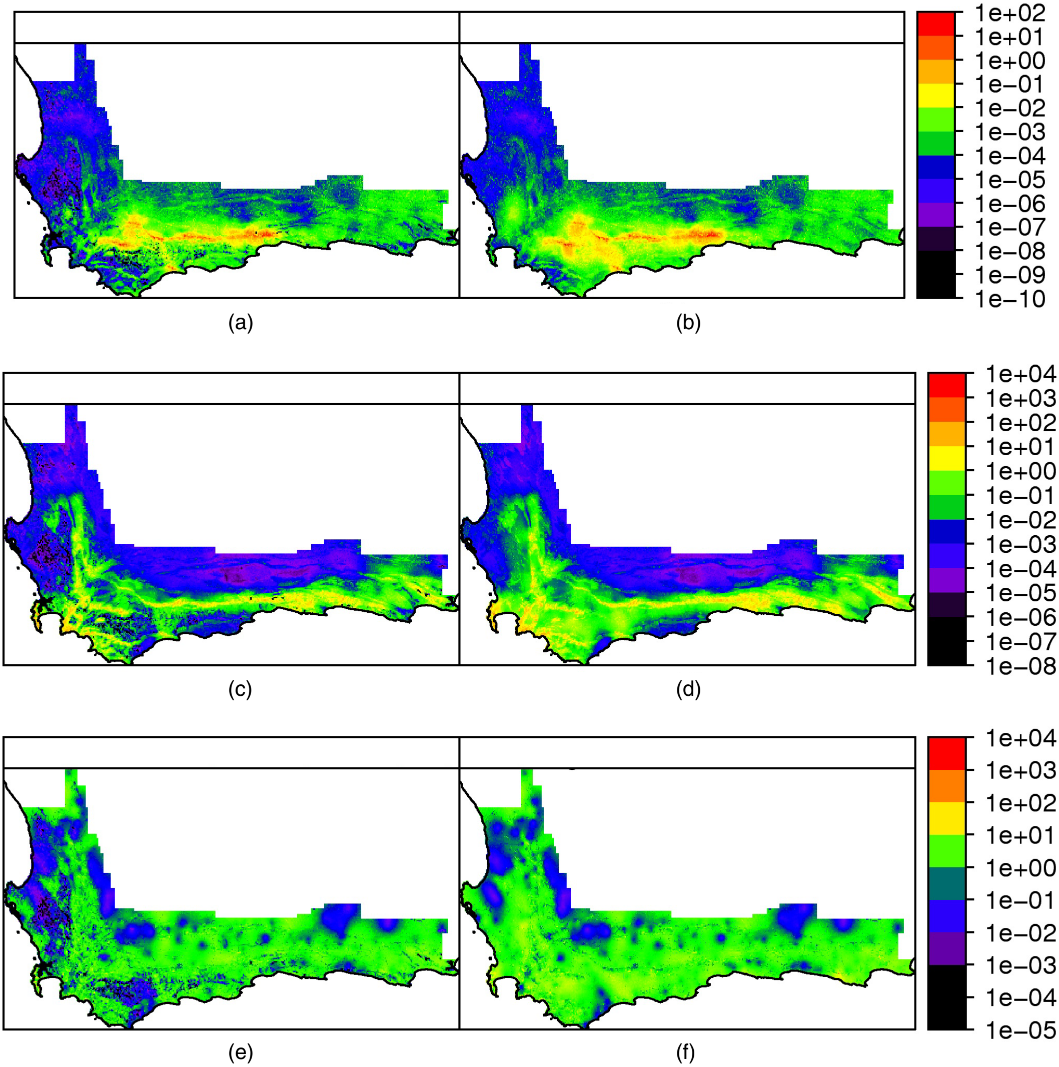

Fig. 4. Maps of estimated species intensities over the CFR for (a), (b) *Protea aurea*, (c), (d) *Protea cynaroides* and (e), (f) *Leucadendron salignum*: (a), (c), (e) potential; (b), (d), (f) transformed

to a much larger set of species. Still, the variation and pattern in potential biodiversity across the CFR is noteworthy.

8. Discussion

We have developed a multilevel point pattern model to explain species distribution by using presence-only data. Our fully model-based approach, though computationally more demanding, provides inference beyond the capabilities of the now widely used Maxent and substantially improves on it in terms of predictive performance. Our approach also avoids the problematic assumptions that are needed in converting the problem to a presence–absence analysis using

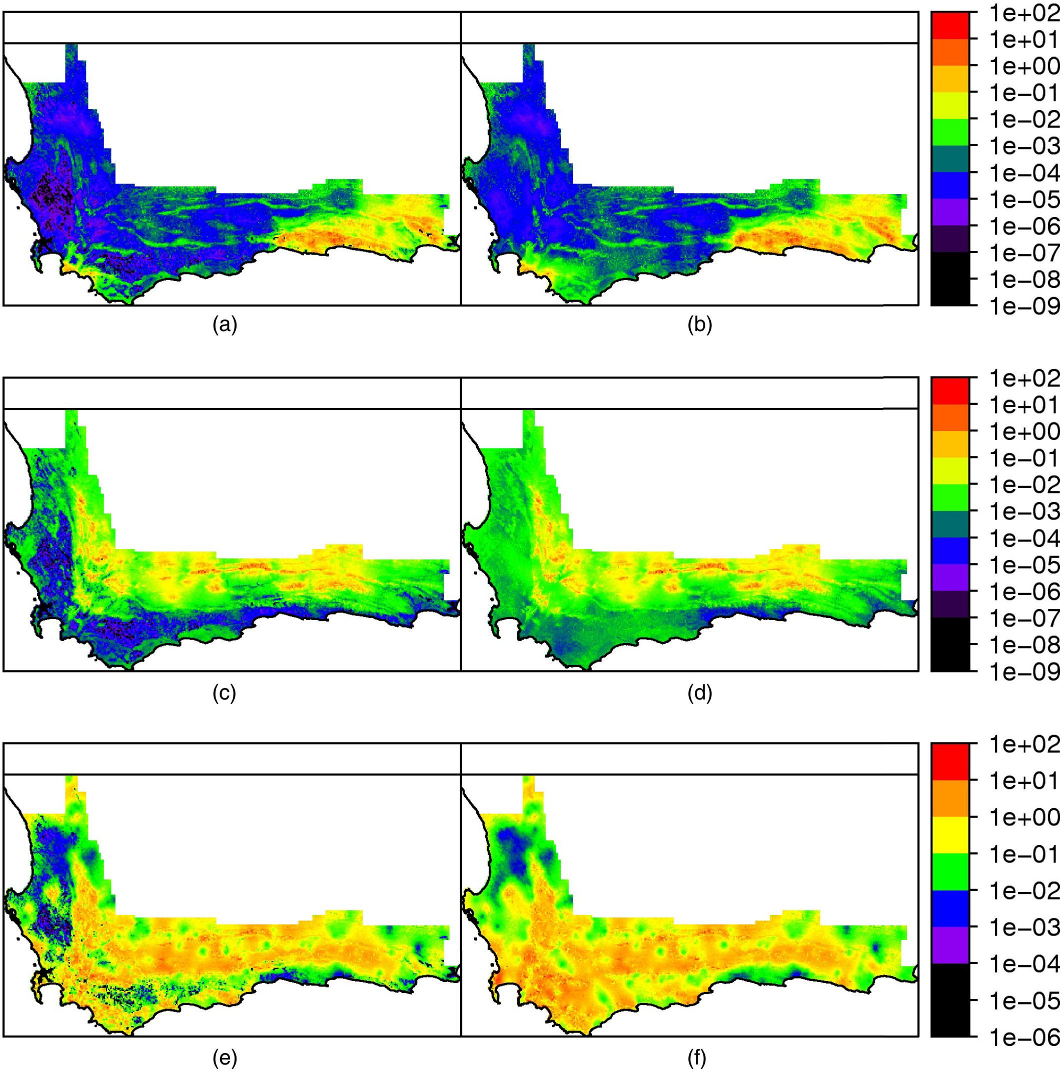

Fig. 5. Maps of estimated species intensities over the CFR for (a), (b) *Protea mundii*, (c), (d) *Protea punctata* and (e), (f) *Protea repens*: (a), (c), (e) potential; (b), (d), (f) transformed

background samples. As with Maxent, our approach accommodates covariate information provided that it is available (at some suitable resolution) for the entire region and it introduces spatial random effects to provide spatially smooth local adjustment to the intensity surface.

Future work could see a dynamic investigation, possibly to assess the response in terms of species distribution to climate change. Another possibility is to model the intensities *jointly* through multivariate GP models for the random effects. Arguably, the most serious challenge is to study the entire ensemble of more than 8000 species in the CFR for which we have point patterns. Evidently, it is computationally infeasible to do this at the individual level. We are exploring clustering strategies (a data-driven taxonomy) through extensions of Dirichlet process models (Chakraborty (2010), chapter 6).

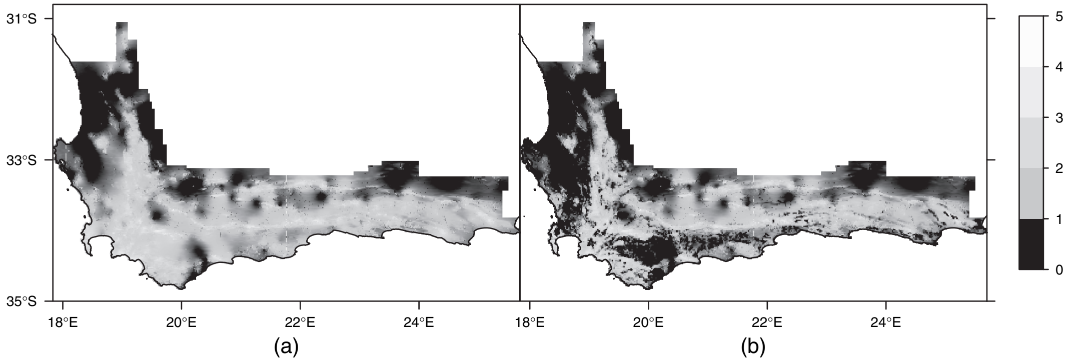

Fig. 6. (a) Potential and (b) transformation-adjusted richness distribution over the CFR with respect to the six species

Acknowledgements

The authors thank Guy Midgley and Anthony Rebelo for useful discussions. In implementing the extensive computation for the CFR data sets, Andrew Finley supplied both resources and advice. Natesh Pillai offered valuable feedback regarding modification of results from his work (Pillai, 2008). We also thank Corey Merow and Debdeep Pati for their useful suggestions. The work of the authors was supported in part by National Science Foundation grants DEB 056320 and DEB 0516198.

Appendix A: Details of Markov chain Monte Carlo algorithm

Inside the MCMC algorithm, the *appropriate* conditional distributions for sampling the parameters can be chosen in several ways to make the run efficient. In what follows, we list steps from one such scheme, for estimation and prediction by using expressions (4) and (5).

(a) Denote $L_1(\phi) = R_{m,r}(\phi) R_r^{-1}(\phi)$, $L_2(\phi) = R_{I-m,r}(\phi) R_r^{-1}(\phi)$, $M_1(\phi) = I_m - \text{diag}\{L_1(\phi) R_{r,m}(\phi)\}$, $M_2(\phi) = I_{I-m} - \text{diag}\{L_2(\phi) R_{r,I-m}(\phi)\}$, $X_1 = x(\mathbf{s}_{1:m})$ and $X_2 = x(\mathbf{s}_{m+1:I})$.

(b) With $\pi(\beta) = N(\beta_0, \Sigma_0)$, draw $\beta|\cdots \sim t(\mu_\beta, \Sigma_\beta)$, where $\Sigma_\beta^{-1} = \Sigma_0^{-1} + \sigma^{-2} X_1^{\mathrm{T}} \{L_1(\phi) R_{r,m}(\phi) + M_1(\phi)\}^{-1}(\phi) X_1$ and $\mu_\beta = \Sigma_\beta[\Sigma_0^{-1}\beta_0 + \sigma^{-2} X_1^{\mathrm{T}} \{L_1(\phi) R_{r,m}(\phi) + M_1(\phi)\}^{-1}(\phi) \log(\lambda_{1:m})]$.

(c) With $\pi(\sigma^2) = \text{IG}(a_0, b_0)$, draw $\sigma^2|\cdots \sim \text{IG}(a_{\sigma^2}, b_{\sigma^2})$ where $a_{\sigma^2} = a_0 + m/2$ and $b_{\sigma^2} = b_0 + e^{\mathrm{T}}\{L_1(\phi) \times R_{r,m}(\phi) + M_1(\phi)\}^{-1} e/2$ for $e = \log(\lambda_{1:m}) - X_1\beta$.

(d) Draw $w(\mathbf{s}_{1:r}^0) \sim N_r(\mu_w, \Sigma_w)$, where $\Sigma_w^{-1} = R_r(\phi)^{-1} + \sigma^{-2} L_1^{\mathrm{T}}(\phi) M_1^{-1}(\phi) L_1(\phi)$ and $\Sigma_w^{-1}\mu_w = \sigma^{-2} L_1^{\mathrm{T}}(\phi) \times M_1^{-1}(\phi)\{\log(\lambda_{1:m}) - X_1\beta\}$.

(e) Use $\pi(\phi) \sim U(\phi_0, \phi_1)$. In equations (4), marginalizing over $w(\mathbf{s}_{1:r}^0)$, the expression involving ϕ becomes

$$
\begin{aligned}
S(\phi) = &-\log\{|L_1(\phi) R_{r,m}(\phi) + M_1(\phi)|\}/2 \\
&- (\log(\lambda_{1:m}) - X_1\beta)^{\mathrm{T}} \{L_1(\phi) R_{r,m}(\phi) + M_1(\phi)\}^{-1} \{\log(\lambda_{1:m}) - X_1\beta\}/2\sigma^2.
\end{aligned}
$$

The inverse and determinant that are involved in the above expression can be calculated efficiently (as $M_1(\phi)$ is diagonal) by using the Sherman–Woodbury–Morrison formula as in Banerjee *et al.* (2008). So, one can use a random walk or independent sampler Metropolis–Hasting update for ϕ. Then update each of $L_i(\phi)$ and $M_i(\phi)$, $i = 1, 2$.

(f) For $1 \leqslant i \leqslant m$, $\pi(\lambda_i|\cdots) \sim^{\text{ind}} \text{LN}(\lambda_i; x_i^{\mathrm{T}}\beta + [L_1(\phi) w(\mathbf{s}_{1:r}^0)]_i, \sigma^2[M_1(\phi)]_{ii}) \times \text{Poisson}(y_i|\lambda_i, q_i)$. One can either do a slice sampling (introduce an auxiliary variable u such that $u|\lambda \sim \exp(1)\mathbf{1}[u > \lambda]$) or a Metropolis–Hastings sampler.

(g) For prediction of $\lambda_{m+1:I}$, draw $\log(\lambda_{m+i}) \sim X_2[i,]\beta + [L_2(\phi) w(\mathbf{s}_{1:r}^0)]_i + \sigma\sqrt{}[M_2(\phi)]_{ii} z$, where $z \sim N(0, 1)$.

References

Augustin, N. H., Mugglestone, M. A. and Buckland, S. T. (1996) An autologistic model for the spatial distribution of wildlife. *J. Appl. Ecol.*, **33**, 339–347.

Banerjee, S., Carlin, B. and Gelfand, A. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall–CRC.

Banerjee, S., Gelfand, A., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc.* B, **70**, 825–848.

Beneš, V., Bodlák, K., Møller, J. and Waagepetersen, R. (2003) Application of log Gaussian Cox processes in disease mapping. In *Proc. Int. Conf. Environmental Statistics and Health, Santiago de Compostela*, pp. 95–105. Santiago de Compostela: Universidade Santiago de Compostela.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Chakraborty, A. (2010) Modeling point patterns, measurement error and abundance for exploring species distributions. *PhD Thesis*. Duke University, Durham.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M. and Silander, J. A. (2010) Modeling large scale species abundance with latent spatial processes. *Ann. Appl. Statist.*, **4**, 1403–1429.

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc.* B, **70**, 209–226.

Diggle, P. J. (2003) *Statistical Analysis of Spatial Point Patterns*, 2nd edn. London: Arnold.

Diggle, P. J., Menezes, R. and Su, T. (2010) Geostatistical inference under preferential sampling (with discussion). *Appl. Statist.*, **59**, 191–232.

Elith, J., Graham, C., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K. S., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S. and Zimmermann, N. E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J., Leathwick, J. R. and Hastie, T. (2008) A working guide to boosted regression trees. *J. Anim. Ecol.*, **77**, 802–813.

Engler, R., Guisan, A. and Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *J. Appl. Ecol.*, **41**, 263–274.

Ferrier, S., Watson, G., Pearce, J. and Drielsma, M. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales: I, species-level modelling. *Biodivrsty Conservn*, **11**, 2275–2307.

Finley, A., Sang, H., Banerjee, S. and Gelfand, A. (2009) Improving the performance of predictive process modeling for large datasets. *Computnl Statist. Data Anal.*, **53**, 2873–2884.

Gelfand, A. E., Schmidt, A. M., Wu, S., Silander, Jr, J. A., Latimer, A. and Rebelo, A. G. (2005a) Explaining species diversity through species level hierarchical modelling. *Appl. Statist.*, **54**, 1–20.

Gelfand, A., Silander, Jr, J., Wu, S., Latimer, A., Lewis, P., Rebelo, A. and Holder, M. (2005b) Explaining species distribution patterns through hierarchical modeling. *Baysn Anal.*, **1**, 41–92.

Graham, C., Ferrier, S., Huettman, F., Moritz, C. and Peterson, A. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol. Evoln*, **19**, 497–503.

Guisan, A., Edwards, T. and Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol. Modllng*, **157**, 89–100.

Harville, D. A. (1997) *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Heikkinen, R. K., Luoto, M., Araújo, M. B., Virkkala, R., Thuiller, W. and Sykes, M. T. (2006) Methods and uncertainties in bioclimatic envelope modelling under climate change. *Prog. Phys. Geog.*, **30**, 751–777.

Higdon, D. (2002) Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* (eds C. Anderson, V. Barnett, P. C. Chatwin and A. H. El-Shaarawi), pp. 37–56. London: Springer.

Higgs, M. D. and Hoeting, J. A. (2010) A clipped latent variable model for spatially correlated ordered categorical data. *Computnl Statist. Data Anal.*, **54**, 1999–2011.

Latimer, A. M., Wu, S., Gelfand, A. E. and Silander, Jr, J. A. (2006) Building statistical models to analyze species distributions. *Ecol. Applic.*, **16**, 33–50.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J. and Aulagnier, S. (1996) Application of neural networks to modelling nonlinear relationships in ecology. *Ecol. Modllng*, **90**, 39–52.

Lobo, J. M., Baselga, A., Hortal, J., Jiménez-Valverde, A. and Gómez, J. (2007) How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? *Divrsty Distribns*, **13**, 772–780.

Loiselle, B., Jørgensen, P., Consiglio, T., Jiménez, I., Blake, J., Lohmann, L. and Montiel, O. (2007) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *J. Biogeog.*, **35**, 105–116.

Møller, J. and Waagepetersen, R. P. (2004) *Statistical Inference and Simulation for Spatial Point Processes*, 1st edn. Boca Raton: Chapman and Hall.

Newbold, T., Gilbert, F., Zalat, S., El-Gabbas, A. and Reader, T. (2009) Climate-based models of spatial patterns of species richness in Egypt's butterfly and mammal fauna. *J. Biogeog.*, **36**, 2085–2095.

Pearce, J. L. and Boyce, M. S. (2006) Modelling distribution and abundance with presence-only data. *J. Appl. Ecol.*, **43**, 405–412.

Peterson, A. T. and Kluza, D. (2003) New distributional modelling approaches for gap analysis. *Anim. Conservn*, **6**, 47–54.

Phillips, S., Anderson, R. and Schapire, R. (2006) Maximum entropy modeling of species geographic distributions. *Ecol. Modllng*, **190**, 231–259.

Phillips, S. and Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Phillips, S., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. and Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Applic.*, **19**, 181–197.

Pillai, N. S. (2008) Posterior consistency of nonparametric Poisson regression models. *PhD Thesis*, pp. 66–77. Duke University, Durham.

Rebelo, A. G. (2002) The state of plants in the Cape Flora. In *The State of South Africa Species* (eds G. H. Verdoorn and J. L. Roux). Johannesburg: Endangered Wildlife Trust.

Reese, G. C., Wilson, K. R., Hoeting, J. A. and Flather, C. H. (2005) Factors affecting species distribution predictions: a simulation modeling experiment. *Ecol. Applic.*, **15**, 554–564.

Stein, M. L., Chi, Z. and Welty, L. J. (2004) Approximating likelihoods for large spatial data sets. *J. R. Statist. Soc.* B, **66**, 275–296.

Stockwell, D. R. B. and Peters, D. (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *Int. J. Geog. Inform. Sci.*, **13**, 143–158.

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K. and Possingham, H. (2003) Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecol. Applic.*, **13**, 1790–1801.

Van Lieshout, M. N. M. (2000) *Markov Point Processes and Their Applications*, 1st edn. London: Imperial College Press.

Veloz, S. D. (2009) Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *J. Biogeog.*, **36**, 2290–2299.

Ver Hoef, J. M., Cressie, N., Fisher, R. and Case, T. J. (2001) Uncertainty and spatial linear models for ecological data. In *Spatial Uncertainty for Ecology: Implications for Remote Sensing and GIS Applications* (eds C. T. Hunsaker, M. F. Goodchild, M. A. Friedl and T. J. Case), pp. 214–237. New York: Springer.

Ward, G., Hastie, T., Barry, S., Elith, J. and Leathwick, J. (2009) Presence-only data and the EM algorithm. *Biometrics*, **65**, 554–563.

Warton, D. I. and Shepherd, L. C. (2010) Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Ann. Appl. Statist.*, **4**, 1383–1402.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. and Guisan, A. (2008) Effects of sample size on the performance of species distribution models. *Divrsty Distribns*, **14**, 763–773.

Zhang, H. (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Am. Statist. Ass.*, **99**, 250–261.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary document for point pattern modeling for degraded presence-only data over large regions'.

Please note: Wiley–Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the author for correspondence for the article.