Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 366–379 doi: 10.1111/2041-210X.12352

SPECIAL FEATURE – REVIEW

NEW OPPORTUNITIES AT THE INTERFACE BETWEEN ECOLOGY AND STATISTICS

Point process models for presence-only analysis

Ian W. Renner[1]*, Jane Elith[2], Adrian Baddeley[3], William Fithian[4], Trevor Hastie[4], Steven J. Phillips[5], Gordana Popovic[6] and David I. Warton[6]

[1]School of Mathematical and Physical Sciences, The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia; [2]School of BioSciences, The University of Melbourne, Parkville, Vic. 3010, Australia; [3]Department of Mathematics & Statistics, Curtin University, GPO Box U1987, Perth, WA 6845, Australia; [4]Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94303, USA; [5]2201 4th Street, Boulder, CO 80304, USA; and [6]School of Mathematics and Statistics and Evolution & Ecology Research Centre, The University of New South Wales, Sydney, NSW 2052, Australia

Summary

1. Presence-only data are widely used for species distribution modelling, and point process regression models are a flexible tool that has considerable potential for this problem, when data arise as point events.

2. In this paper, we review point process models, some of their advantages and some common methods of fitting them to presence-only data.

3. Advantages include (and are not limited to) clarification of what the response variable is that is modelled; a framework for choosing the number and location of quadrature points (commonly referred to as pseudo-absences or 'background points') objectively; clarity of model assumptions and tools for checking them; models to handle spatial dependence between points when it is present; and ways forward regarding difficult issues such as accounting for sampling bias.

4. Point process models are related to some common approaches to presence-only species distribution modelling, which means that a variety of different software tools can be used to fit these models, including MAXENT or generalised linear modelling software.

Key-words: Cox processes, Gibbs processes, MAXENT, pseudo-absences, species distribution modelling

Introduction

Species distribution modelling (SDM) provides a framework for determining the distribution of a species' habitat as a function of environmental variables and is a highly researched topic of interest to ecologists, biologists and climate change scientists. Often, the best available species data come in the form of a list of reported presence locations of a species without any corresponding information about where a species is absent. This type of data is known as 'presence-only' data (Pearce & Boyce 2006) and can be found in museums, atlases and herbaria. A researcher interested in exploring the relationship between a species and the environment is faced with the question of which methods to choose, and there have been calls for unification of SDM concepts (Elith & Leathwick 2009; Aarts, Fieberg & Matthiopoulos 2012). Here, using language common to the SDM literature, we aim to progress understanding of methods by focussing on emerging knowledge of the links between point process models (PPMs), regression and MAXENT.

Presence-only data typically arise as point events – a set of point locations where a species has been observed. In the statistical literature, a set of point events (in which the location and number of points is random) is known as a *point process*. Spatial statistics literature provides a suite of tools for modelling point processes (Cressie 1993; Diggle 2003), but only recently have point process models been proposed as a natural way for analysing species presence-only data in a regression framework (Warton & Shepherd 2010; Chakraborty *et al.* 2011). PPMs are closely connected to methods already in widespread use in ecology such as MAXENT (Aarts, Fieberg & Matthiopoulos 2012; Fithian & Hastie 2013; Renner & Warton 2013), some implementations of logistic regression (Baddeley *et al.* 2010; Warton & Shepherd 2010) and estimation of resource selection functions (Aarts, Fieberg & Matthiopoulos 2012; McDonald *et al.* 2013). PPMs enjoy particular benefits in interpretation and implementation. Consequently, we believe PPMs are a natural choice of analysis method for presence-only SDM, when the data arise as point events.

In this paper, we review PPMs for species distribution modellers, and different methods for fitting them. We show that the point process viewpoint resolves a number of important questions regarding presence-only data, including exactly what target quantity is being modelled, how to select background or pseudo-absence points (named quadrature points in the PPM literature), what assumptions are made and how they can be

*Correspondence author. E-mail: ian.renner@newcastle.edu.au checked. It also suggests a natural way to deal with biases. We provide a worked example to demonstrate how to fit PPMs with different methods.

Example – distribution of *Eucalyptus sparsifolia*

We will proceed by briefly describing an example data set to be used throughout the paper to illustrate key ideas. The data set comprises 230 presence-only locations of *Eucalyptus sparsifolia* within the Greater Blue Mountains World Heritage Area (GBMWHA) and a surrounding 100-km buffer zone (Fig. 1), a 86 227-km^2 area near Sydney, Australia (NSW Office of Environment and Heritage 2012). This species is known to be abundant and broadly distributed across this region (Hager & Benson 2010). Maps of environmental variables were available over the study region, and the goals of analysis were to map the distribution of *Eucalyptus sparsifolia* and identify its key environmental correlates.

The presence records were entirely from incidental sightings of the species collated by the responsible state department since 1972. These records were cleaned prior to analysis to remove data from systematically sampled transects, and records with high location errors, leaving only records of opportunistic sightings whose point location was known to a reasonable (1 km) degree of accuracy. As these observations are reported as locations, not counts in transects or grid cells, they are best described as point locations in continuous space, which motivates the use of point process models for analysis.

Environmental predictors selected as likely relevant to the distribution of the species and its records are minimum and maximum annual temperature, annual rainfall, number of fires since 1943 and a categorical soil variable. A description of the soil categories is presented in Section 1 of the Appendix S1. All variables were available at 100-m grid cell resolution (Renner *et al.* 2015).

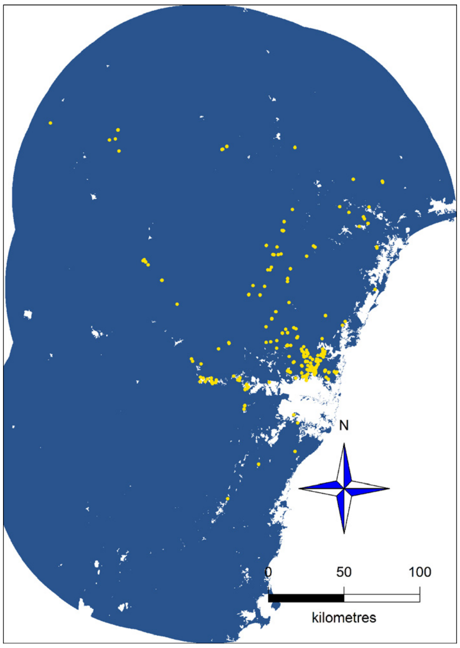

Fig. 1. Locations of 230 presence-only *Eucalyptus sparsifolia* observations within 100 km of the Greater Blue Mountains World Heritage Area.

Point process models

Presence-only data consist of a set of locations $\mathbf{s}_P = \{s_1, s_2, \ldots, s_m\}$ at which a species has been observed in some region $\mathcal{A}$. While methods for fitting PPMs are closely related to common regression models, particularly generalised linear models (GLMs, McCullagh & Nelder 1989), PPMs are posed differently. A regression model is typically used when the object of interest is a random variable y_i, for which we model the mean μ_i as a function of covariates $\mathbf{x}_i$. By contrast, the objects of primary interest in a PPM are the spatial locations of presence points $\mathbf{s}_P$ – that is the focus is on *where* the points were observed. We model the locations in $\mathbf{s}_P$ jointly with the number of presences m and characterise them via the *intensity* or limiting expected number of presence points per unit area $\lambda(s)$. The link to regression comes because we typically model $\lambda(s)$ as a function of covariates $\mathbf{x}(s)$ measured throughout the study region $\mathcal{A}$.

The first advantage of PPMs, before looking any further, is greater clarity about what exactly is being modelled (Aarts, Fieberg & Matthiopoulos 2012; Dorazio 2012). The target of interest, intensity, is not a probability and is instead a measure of abundance – the number of presence records per unit area. Thus, it need not have an upper bound of one (Aarts, Fieberg & Matthiopoulos 2012). Further, intensity as defined is a function of only two quantities – spatial patterning in the presence-only data, and the spatial measurement units. Changing the spatial units from kilometres to metres should change intensity proportionally (decreasing by a factor of 1000^2).

It should be emphasised that in most instances, the intensity $\lambda(s)$ does not reflect the expected abundance per unit area of a species; rather, it reflects the expected abundance of *species reportings*. It can typically only be used to make inferences about *relative* patterns in species abundance (Fithian & Hastie 2013).

THE POISSON CASE – NO SPATIAL DEPENDENCE

The simplest type of PPM of use in presence-only analysis is an *inhomogeneous* Poisson point process (hereafter referred to as a Poisson PPM), in which we assume (a) point events are independent of each other, which can be shown to imply that the total number of points in the study region is a Poisson random variable, and (b) that the intensity $\lambda(s)$ varies spatially (and so is indexed by location s). We will further assume it varies according to environmental conditions $\mathbf{x}(s)$.

Assumption (a), in which point locations are independent, is a restrictive assumption which often is not satisfied by presence-only data. Methods to check the independence assumption are described in Section 'Software for fitting point process models' and methods to fit models that account for dependence are described in Section 'Spatial dependence in point processes'.

Assumption (b) is often refined to a loglinearity assumption, where we model intensity as a loglinear function of environmental covariates:

$$\ln \lambda(s) = \mathbf{x}(s)'\beta, \qquad \text{(eqn 1)}$$

where $\beta = \{\beta_1, \ldots, \beta_p\}$ is a vector that contains the parameters corresponding to the p environmental covariates $\mathbf{x}(s)$. Loglinearity is a natural assumption because it ensures that intensity is a non-negative quantity, and it is the canonical link for Poisson data. While the form of this equation is loglinear, it can readily capture nonlinear relationships between intensity and the environment, for example, using quadratic and interaction terms, smoothed functions in generalised additive models (GAMs, Hastie & Tibshirani 1990) or via kernel regression (Guan 2008; Baddeley *et al.* 2012).

Loglinear models of the form of (1) are commonly fitted to count data, and one way to fit a PPM is in fact to break the data into grid cells and fit a Poisson loglinear model to the counts of presence points in each grid cell. However, such a model has the potential to lose information from the data during aggregation from point location to the grid cell level (Renner & Warton 2013). It may, however, be helpful to readers to think of a Poisson PPM as like a model for count data – the key distinction being that the data for analysis are the set of point locations of presences, rather than counts in grid cells.

REGRESSION MODELS OF PRESENCE–BACKGROUND DATA

Here we pause to briefly discuss connections with a common practice in the SDM literature, presence–background (PB) regression, also referred to as pseudo-absence regression (e.g. Chefaoui & Lobo 2008; Phillips *et al.* 2009; Barbet-Massin *et al.* 2012) and more recently as 'naïve regression' (Fithian & Hastie 2013). This models presence (y = 1) and background (treated as y = 0) with logistic regression methods usually used for presence–absence data. This approach can be understood as being 'naïve' essentially because of a mismatch between the model being fitted and the data that were collected – the presences (y = 1) are the raw data for which we wish to specify a model, and the background points (y = 0) are a fabrication. PB regression was motivated by the need to model distributions of species for which survey (PA) data were unavailable and, early on, by a lack of suitable alternatives. It has remained popular, perhaps because of the examples in which this approach seems to work reasonably well compared with other methods, and the fact that many ecologists are already familiar with regression methods. The approach is somewhat *ad hoc*. Some users apply arbitrary weights to the background samples, for pragmatic rather than statistically based reasons (Elith *et al.* 2006). The fitted quantity is interpreted as a relative likelihood of presence, with an unknown scaling linking it to the true probability of presence.

One of the most challenging steps in fitting a PB regression model is selection of the background points. A common choice has been to select a large number (thousands) of points across the landscape of interest (Elith *et al.* 2006). Other more complex schemes include identifying points more likely to represent a true absence, or at least avoiding presence points (Engler, Guisan & Rechsteiner 2004), or trying to specify an optimal number of background points (or presence–background ratio) for different methods (Barbet-Massin *et al.* 2012). These are usually based on an idea about data structure or on evidence that it performs better than an alternative in particular case studies or simulations, but without stronger statistical justification. These have created some confusion among users regarding which approach is the best to use. We think that efforts to clarify background sampling schemes under the naïve model are misdirected and that much is to be gained by changing to the point process viewpoint. As will be seen later, this viewpoint provides a solid statistical framework for understanding the role of background points and for deciding how many, placed where, are sufficient.

Warton & Shepherd (2010) and Fithian & Hastie (2013) discuss problems with using naïve logistic regression and its various extensions for presence-only data. One problem is scale dependence – the scale of PB regression predictions is meaningless since the predictions change as more background points are added to the sample. But Warton & Shepherd (2010) and Baddeley *et al.* (2010) showed that PB regression can be understood as an approximation to fitting a Poisson point process model and that the latter resolves the scale dependence issue, and many issues with background choice.

SPATIAL DEPENDENCE IN POINT PROCESSES

An underlying assumption of the Poisson PPMs (and most PB regression methods) is that data are conditionally independent given the covariates; that is, the similarities in intensity in nearby regions are fully explained by the environmental and sampling covariates in the model. This is, however, often not the case, and failing to account for spatial dependence can significantly alter conclusions (Dormann 2007). Common examples of spatial dependence are clustering through dispersal or social aggregation. Spatial dependence may also be induced by failing to measure environmental variables which act to make presence patterns in regions close together seem more similar than those further apart.

The two most common classes of point process models for species dependence are Gibbs and Cox processes.

Gibbs or 'interaction' processes relax the independence assumption by assuming interactions between sets of points. One useful example is area-interaction processes (Widom & Rowlinson 1970; Baddeley & van Lieshout 1995), which assume interactions among all points within a distance of $2r$. These interactions can be understood as introducing an additional term to the intensity function (conditional on the observed presence points):

$$\ln \lambda(s) = \mathbf{x}(s)'\beta + t_s(\mathbf{s}_P)\theta \qquad \text{(eqn 2)}$$

where θ is an interaction parameter (positive values implying clustering of points) and $t_s(\mathbf{y})$ is the area of a disc of radius r

centred at the location s that does not intersect with similar discs centred around each of the presence points $\mathbf{s}_P$.

Area-interaction models are potentially useful for SDMs because they are capable of modelling either clustering or inhibition. They also have some biological justification; for example, interaction radius r could be considered as a maximum dispersal distance, with intensity increasing at locations within a distance r of a known presence because of the chance of establishment from that presence point. Beyond area-interaction processes, there are a number of other types of Gibbs process, in particular processes that involve pairwise interaction between points (for a list, see Baddeley & Turner 2005, Section 9 of the Appendix S1).

An alternative way to deal with clustering and the effects of unmeasured covariates is by fitting a Cox process, the most common example of which is the spatial log-Gaussian Cox process (LGCP) (Møller, Syversveen & Waagepetersen 1998). This can be understood as a point process analogue of a generalised linear mixed model with a random intercept that is normally distributed.

The intensity $\lambda(s)$ in a LGCP is a function not just of environmental variables, but also of a stochastic Gaussian process $\xi(s)$:

$$\ln \lambda(s) = \mathbf{x}(s)'\beta + \xi(s). \qquad \text{(eqn 3)}$$

Here, $\xi(s)$ is a spatial Gaussian process with zero mean, and a covariance function that depends on the distance between observations, such that observations closer together in space are assumed to be more positively correlated than those further apart. The $\xi(s)$ can be understood as an unmeasured covariate which is associated with the distribution of the species. Conditional on this latent process, the point events are assumed to be inhomogeneous Poisson. In other words, it is assumed that any spatial dependence in the data is entirely captured by $\xi(s)$.

Fitting a point process model

This section provides an overview of the process of fitting a PPM, with more detailed information about software and example code provided in the Appendix S1. We use the example *Eucalyptus sparsifolia* data introduced previously in illustratory analyses.

FITTING POISSON POINT PROCESSES

The most common approach to fitting a Poisson PPM is to maximise the log-likelihood function (Cressie 1993), which can be written as:

$$l(\beta; \mathbf{s}_P) = \sum_{i=1}^{m} \ln \lambda(s_i) - \int_{\mathcal{A}} \lambda(s)ds. \qquad \text{(eqn 4)}$$

For a derivation of this log-likelihood, and how it differs from the homogeneous Poisson point process case, see Appendix S1 (Section 2). The integral in this expression can be interpreted as the expected number of presence points in the whole study region $\mathcal{A}$, and it is the approximation of this quantity that is the main challenge in model fitting.

Estimation of parameters in the inhomogeneous Poisson PPM is not straightforward, because the integral in (4) does not have a closed form and must be approximated in some way. A standard way to approximate this integral is through the use of numerical integration, otherwise known as 'quadrature' (Davis & Rabinowitz 1984). The general idea is to choose a set of 'quadrature points' at which the intensity function is evaluated, and these evaluations are then combined as a weighted sum to estimate the integral. Common examples of quadrature methods are Riemann sums and the trapezoidal rule. Irrespective of the quadrature method used, the likelihood can then be written as:

$$l(\beta; \mathbf{s}_P) \approx \sum_{i=1}^{m} \ln \lambda(s_i) - \sum_{j=1}^{m+n} w_j \lambda(s_j), \qquad \text{(eqn 5)}$$

$$= \sum_{j=1}^{m+n} w_j \big(y_j \ln \lambda(s_j) - \lambda(s_j)\big) \qquad \text{(eqn 6)}$$

where $\mathbf{w} = \{w_1, \ldots, w_{m+n}\}$ are quadrature weights, $\mathbf{s}_0 = \{s_{m+1}, \ldots, s_{m+n}\}$ are quadrature points and $y_j = \frac{1}{w_j}$ for presence points ($j = 1, \ldots, n$) and $y_j = 0$ otherwise. The quadrature weights w_j can be understood as applying a spatial scaling, so that the response being modelled (intensity, λ) has spatial units not observational units. For example, in analysing the *Eucalyptus sparsifolia* data, we modelled the expected number of presences *per square kilometre*. Broadly speaking, w_j represents the area of the neighbourhood around the point s_j, found after partitioning the study region $\mathcal{A}$ into neighbourhoods around each point (including presences and quadrature points).

Equation 6 is due to Berman & Turner (1992), and it reexpresses the likelihood as a Poisson likelihood with observation weights w_j. The significance of this result is that it makes Poisson PPMs relatively straightforward to fit – they can be fitted using any standard GLM software, such as the GLM function in R (R Development Core Team 2010), using the Poisson family and appropriate observation weightings. This can be done in just a few lines of code, as illustrated in the Appendix S1, although specialised packages are available that offer enhanced options – the SPATSTAT package on R (Baddeley & Turner 2005) has a suite of model-checking tools, and the PPMLASSO package adds a LASSO penalty for improved predictive performance.

The connection to Poisson GLMs in equation 6 has enabled equivalence results between Poisson PPMs and PB regression in large samples (Baddeley *et al.* 2010; Warton & Shepherd 2010) and more recently MAXENT (Aarts, Fieberg & Matthiopoulos 2012; Fithian & Hastie 2013; Renner & Warton 2013), offering some insight into these methods. For example, the scale dependence of PB regression can be understood as arising due to the omission of appropriate observation weights w_j. The equivalence with MAXENT implies that MAXENT software can be used to fit a Poisson PPM. We will discuss the capabilities of these alternate methods for fitting Poisson PPMs later.

HOW TO CHOOSE QUADRATURE POINTS

A first step in fitting a PPM is selection of quadrature points, to allow estimation of the PPM likelihood. This is an equivalent issue to that of background selection (Section 'Regression models of presence–background data'). However, our choice of the term 'quadrature points' in this Section reflects a desire to pose the question of their choice as a quadrature problem, which clarifies their role in analysis and provides a framework for their selection.

From the point process viewpoint, quadrature points are merely a device to estimate an integral. This is true across the various methods that can be used to fit a Poisson PPM (Section Software for fitting point process models, see also Fithian & Hastie 2013). Being such a device, the important question becomes: How many points, placed where, will be sufficient to accurately estimate the likelihood? This is primarily a question for which the number and location of *presence* records are irrelevant; hence, ideas of matching number of presence points and sampling far away from presence points are not relevant, except when computational efficiency is an issue, which then requires specifically designed schemes (see later).

Assuming that the extent of the study area is pre-determined by the analyst, two simple strategies to select the *location* of quadrature points within that extent are (i) to choose them on a regular mesh (e.g. at regularly spaced intervals in each cardinal direction) or (ii) to choose them randomly. Alternative strategies whereby the density of quadrature points increases with environmental variability may be helpful in scenarios where computation time is slow as these would lead to a smaller data set with negligible loss in accuracy. This is related to the ideas of importance sampling and adaptive quadrature (Davis & Rabinowitz 1984), which seem so far under-utilised in the SDM literature.

The *number* of quadrature points should be sufficient for an accurate estimate of the likelihood, which will lead to a stable model that is approximately invariant across repeat samples of the points. To determine an appropriate number of quadrature points, we advise that analysts check that sufficient accuracy is achieved by increasing the number of quadrature points until there is little appreciable change in model fit or in predictive performance (Phillips & Dudík 2008). For instance, for Fig. 2a, we fitted Poisson PPMs to *Eucalyptus sparsifolia* data repeatedly halving the spacing between quadrature points, selected across a regular mesh. This was done using the PPMLASSO package, which has a function specifically designed to perform this operation. The log-likelihood converged at a 1-km spacing, which required more than 86 000 quadrature points for our example (Fig. 2a), noticeably more than common software defaults (e.g. 10 000 in MAXENT).

Undersampling of quadrature points will lead to error in our coefficient estimates, although this error may still be small compared to the coefficients' overall standard errors. If quadrature points have been randomly sampled (rather than using a regular mesh), we can easily quantify this error by considering what might happen under repeated sampling of quadrature points. For example, in Fig. 2b, models were fitted with an increasing number of randomly chosen quadrature points, and results replicated 30 times to study how much results varied when using different sets of random quadrature points (code available in Section 5 of the Appendix S1). Figure 2b suggests that 100 000 or more randomly chosen quadrature points would be needed to reliably estimate the maximised log-likelihood – with significant variation from one set of quadrature points to another before this point. In particular, if using 10 000 random quadrature points, the maximised log-likelihood varied over a range of more than 50 for different samples of quadrature points.

Alternatively, the error introduced by quadrature can be estimated analytically quite easily. The integral was estimated as an average of estimated intensities at n random points (multiplied by $|\mathcal{A}|$), so its uncertainty can be estimated using the formula for the standard error of a sample mean, $\frac{|\mathcal{A}|\sigma}{\sqrt{n}}$. In our example data set, given an initial fit using 10 000 random quadrature points, the standard deviation of estimated intensities at the quadrature points was $s = 0{\cdot}0103$, yielding an estimated standard error of $\frac{86\,227 \times 0{\cdot}0103}{\sqrt{10\,000}} = 8{\cdot}89$. If we desire an

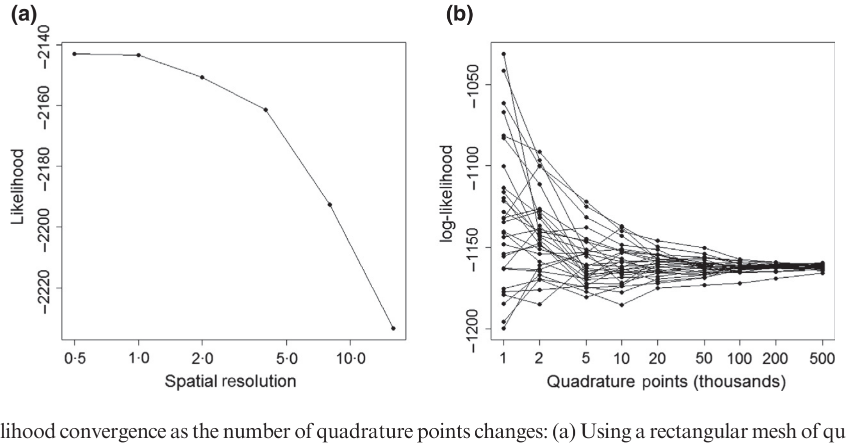

Fig. 2. Checking for likelihood convergence as the number of quadrature points changes: (a) Using a rectangular mesh of quadrature points at different spatial resolutions, as available in the PPMLASSO package; (b) using random sets of quadrature points and progressively increasing the sample size, as estimated using downweighted Poisson regression models. It appears that there is little benefit in analysing the data at a spatial resolution finer than 1 km (a), or with more than 100 000 quadrature points (b).

estimate of the log-likelihood to be within a standard error of two of its true value, we can estimate the required number of quadrature points as $|\mathcal{A}|^2 s^2/2^2 = 197\,745$. This corresponds well with the results of Fig. 2b.

The precise number of points needed for a sufficiently accurate estimate should vary with the roughness of the intensity surface (hence the difficulty of the integration problem). Thus, we may expect to need more quadrature points when environmental data are measured at a finer resolution (equivalently, smaller grid cell sizes) or broader spatial extent.

'Quadrature thinking' also leads to other emphases. First, if computational efficiency is of key importance, it could be prudent to sample fewer quadrature points (with higher corresponding quadrature weights) in areas where the species is unlikely to occur, which should have negligible impact on the intensity and the likelihood of the fitted model. However, this needs to be done with care because it relies on correct identification and inclusion in the model of the covariates causing species absence from certain parts of the landscape. An example of where this may be appropriate is in telemetry studies, where an individual's location is typically strongly associated with distance from the last observed location; hence, quadrature points far from that location make negligible likelihood contribution (Warton & Aarts 2013).

Secondly, the quadrature viewpoint tends to place more emphasis on specifying the model in a way that deals with biases, rather than fiddling with quadrature points. Hence with quadrature thinking, one is more likely to accommodate bias by explicitly specifying covariates in the model (e.g. Warton, Renner & Ramp 2013; Fithian *et al.* 2015) than through the selection of quadrature points (e.g. target-group background as in Phillips *et al.* 2009). For example, in our *Eucalyptus sparsifolia* model, we added two predictors related to site accessibility in order to model observer bias (distance from main roads and distance from urban areas). These two observer bias variables were included to try to account for spatial patterning in presence locations due to behaviour of observers rather than behaviour of the study species. By then making predictions at a common level of observer bias (e.g. distance equals zero), we can map *E. sparsifolia* distribution controlling for observer bias (Warton, Renner & Ramp 2013; Fithian *et al.* 2015).

Finally, most methods for fitting Poisson PPMs attach weights to the quadrature points (the w_j in equation 5) for a scale-invariant estimate of the log-likelihood that is comparable across sets of quadrature samples of different size. This can have advantages for model fitting and interpretation (Warton & Shepherd 2010).

CHECKING ASSUMPTIONS

A suite of diagnostic tools are available in the point process literature that can be used to ground-truth model assumptions. Just as with ordinary regression models, residual analysis (Baddeley *et al.* 2005) can be used to assess adequacy of the model for intensity, in particular by checking for a spatial trend in residuals. The assumption of independence among point locations can be checked using Ripley's *K*-function (Ripley 1977) and its generalisations (Baddeley, Møller & Waagepetersen 2000).

Consider, for example, a Poisson PPM fitted to the *E. sparsifolia* data. A check of the independence assumption suggests significant clustering of points at radii <10 km (Fig. 3a), so we fit an area-interaction model (see Section 3 of the Appendix S1 for details). The resulting model fit exhibits some pattern (Fig. 3b), but the magnitude of the residuals is not exceedingly large. Cumulative residuals for increasing longitude and latitude (x and y) do not significantly deviate from Monte Carlo simulation envelopes (Fig. 3c–d), so the model fit may be deemed sufficiently appropriate.

Other useful diagnostic features demonstrated in Section 3 of the Appendix S1 include influence, leverage and partial residual plots, all derived in direct analogy to how they are used in generalised linear models (as in Baddeley *et al.* 2013). All of these diagnostic plots are relatively easy to produce using the SPATSTAT package.

When checking the assumption of independence of presence points, an alternative approach is to treat models that incorporate spatial dependence as the default, to fit such models, and then study the level of dependence in the subsequent model fit, as in Fig. 4a. This approach makes particular sense as an approach if one is expecting spatial dependence *a priori*, and data of sufficient quality to see the spatial dependence signal. The fitting of such models is discussed below.

FITTING POINT PROCESSES WITH SPATIAL DEPENDENCE

Both Gibbs and Cox processes are more difficult to fit than Poisson PPMs, although for different reasons.

Gibbs processes are difficult to fit by maximum likelihood because their specification involves a proportionality constant that is difficult to estimate. One workaround is to use the Poisson process likelihood in place of the true likelihood (Besag 1977), that is, use (5) as a pseudo-likelihood. This is often done because the Gibbs likelihood has a complex form, but by using the Poisson pseudo-likelihood instead, estimates are readily available via GLM. This approach is implemented in SPATSTAT and PPMLASSO (see Sections 3 and 4 of the Appendix S1 for detailed code), with a number of types of interaction processes to choose from in SPATSTAT, as specified using the interaction argument. One issue to be aware of when using this approach is that traditional methods of likelihood-based inference (such as likelihood-based standard errors, likelihood ratio tests, AIC) may no longer apply, because parameters have not been estimated by maximum likelihood.

Cox process models are difficult to fit because they involve an unobserved Gaussian random process (the $\xi(s)$ in equation 3), and so maximum likelihood estimation would involve complex integrals. These models are hierarchical and are therefore naturally suited to a Bayesian hierarchical approach to estimation as in Illian, Møller & Waagepetersen (2009) and Chakraborty *et al.* (2011). Other estimation techniques include composite likelihood (Guan 2006) and weighted esti-

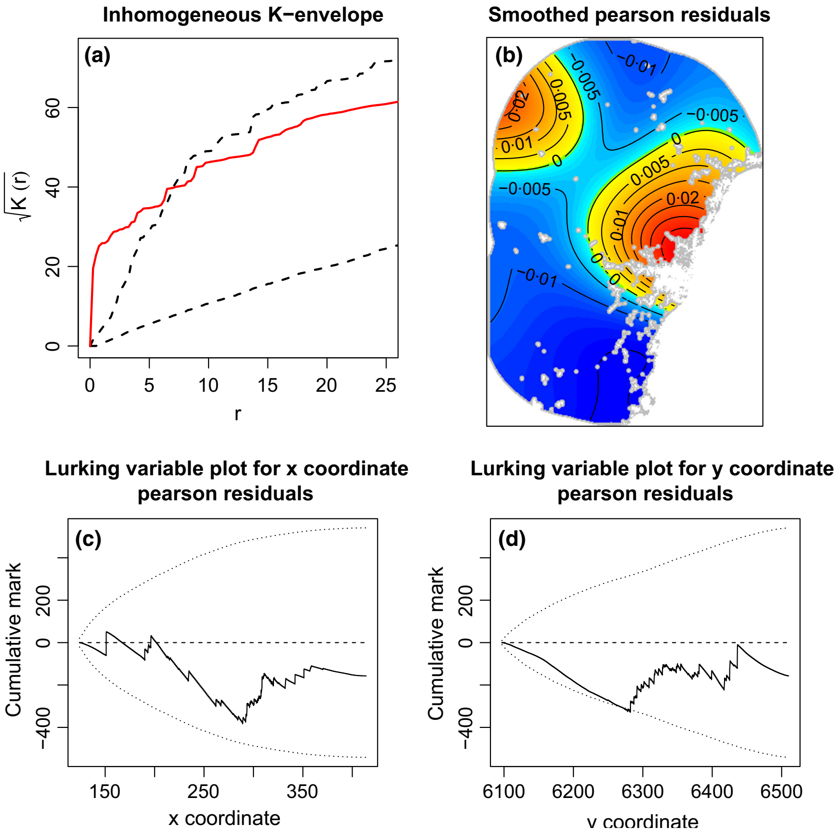

Fig. 3. Example diagnostic plots for a Poisson point process model using the SPATSTAT package – (a) an inhomogeneous *K*-function with 95% simulation envelope, (b) a map of smoothed Pearson residuals, (c) cumulative Pearson residuals for increasing longitude with 95% simulation envelope and (d) cumulative Pearson residuals for increasing latitude with 95% simulation envelope. The inhomogeneous *K*-function (a) suggests point clustering not accounted for by the model as the observed values (red) exceed the upper limit of the envelope (dashed) for radii below 10 km. The residual plot (b) exhibits some pattern, but the magnitude of the residuals is not exceedingly large. Cumulative residuals for increasing longitude and latitude do not significantly deviate from Monte Carlo simulation envelopes (c and d), so the model fit may be deemed sufficiently appropriate.

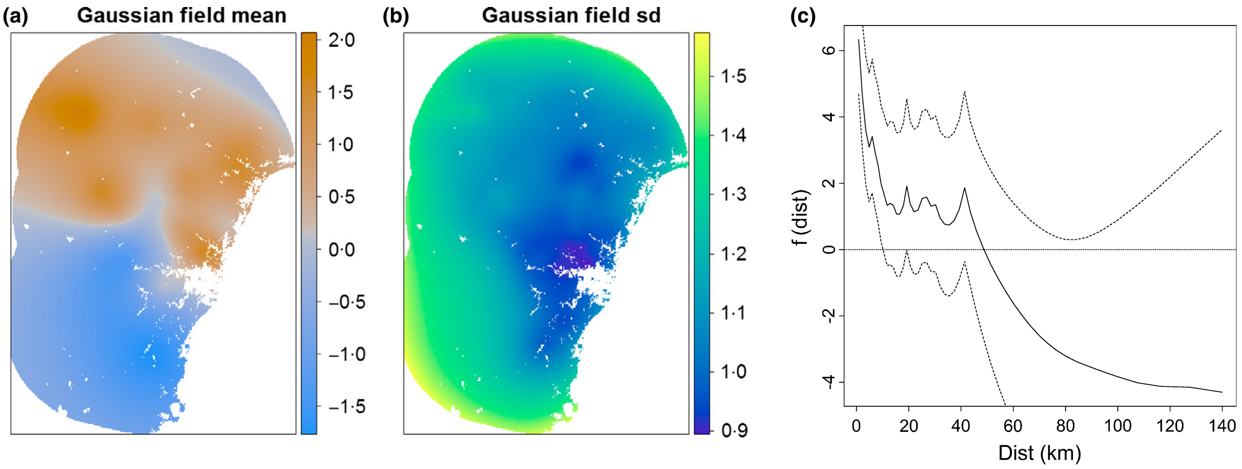

Fig. 4. Estimation of spatial dependence using a log-Gaussian Cox process model: (a) Mean and (b) standard deviation of the random Gaussian field, and (c) a posterior mean and 95% credible interval for the interaction coefficient, computed using the DIST function on R-INLA. The mean is significantly larger than the standard deviation in some regions, and the interaction coefficient in (c) has a prediction interval greater than zero at small distances, both of which suggest clustering in the data beyond that explained by covariates.

mating equations (Guan & Shen 2010), as implemented in SPATSTAT.

Two methods of Cox process estimation that are currently popular, and which both have been implemented under a Bayesian framework, are the integrated nested Laplace approximation (INLA, implemented in the R-INLA package, Rue, Martino & Chopin 2009) and a Markov chain Monte Carlo (MCMC) method (implemented in the LGCP package, Taylor *et al.* 2013). The INLA method seeks to calculate the integrals by a set of carefully chosen approximations. It is generally fast compared to MCMC methods, which can be quite time-consuming but have potentially greater accuracy. For a comparison of these, see Taylor & Diggle (2014).

Sections 7 and 8 of the Appendix S1 have example code for fitting a Cox process using both the R-INLA and LGCP packages. We found these packages much more difficult for the practitioner to use than Poisson PPMs, with specialist guidance often required. The main gain from this additional effort is the possibility of making valid inferences from the model, taking into account uncertainty, under the assumption of spatial dependence. This is otherwise harder to achieve without resorting to resampling, as below.

BLOCK RESAMPLING FOR INFERENCE IN THE PRESENCE OF SPATIAL DEPENDENCE

A complication that can arise when modelling spatially dependent point processes is that likelihood-based standard errors may be estimated to be too small. This problem arises when using the pseudo-likelihood approach to fit Gibbs processes or when failing to account for interpoint dependence correctly. This issue does not arise in fitting Cox processes, if correctly specified, a significant advantage of that approach.

Similarly, but irrespective of what model is fitted to data, standard cross-validation measures of out-of-sample prediction error can be over-optimistic in the presence of dependence, potentially leading us to erroneously prefer models that overfit to local structure in the data (Wenger & Olden 2012). Block resampling techniques can deal with short-range interpoint dependence nonparametrically.

Suppose that interpoint dependence is strong for nearby locations, but weak at radius r. Then if we tile our geographic domain into c rectangular blocks of size $r \times r$, the data falling in one block are approximately independent of the data in other blocks. If we believe this, then we can obtain accurate standard errors using a bootstrap algorithm that resamples whole blocks with replacement (Efron & Tibshirani 1993), as in Slavich *et al.* (2014). Likewise, we can obtain estimates of out-of-sample prediction error by cross-validation where blocks are assigned whole to each fold (Burman, Chow & Nolan 1994), as in Wenger & Olden (2012), Pearson *et al.* (2013) and Warton, Renner & Ramp (2013).

For example, the model for Fig. 5c involved a LASSO penalty which was estimated by 5-fold cross-validation using blocks of 32 km × 32 km. Section 4 of the Appendix S1 has example code for block cross-validation with the PPMLASSO package.

SOFTWARE FOR FITTING POINT PROCESS MODELS

The main software packages currently available for fitting point process models, and their key differences in properties, are summarised in Table 1. All are available in R. In the Appendix S1, we have developed short tutorials stepping the user through analysis of the *Eucalyptus sparsifolia* data using each of these packages, and we encourage new users to work through these resources when deciding on an approach to analysis of point process data.

The most established package for point process modelling is SPATSTAT, whose main advantages are the extensive suite of diagnostic tools, and its ability to simulate data from a given point process model. The PPMLASSO package was written to be SPATSTAT-compatible, so it inherits many useful diagnostic tools, while adding a couple of functions of particular interest for SDMs – the ability to regularise parameter estimates (i.e. shrink them towards zero to reduce variance, Hastie, Tibshirani & Friedman 2009) using the LASSO or elastic net, and functions to guide the user regarding quadrature point choice, as in Fig. 2a. Standard errors are not returned in PPMLASSO output, since these become very approximate when using a LASSO or other regularisation approach in parameter estimation.

Because point process models can be fitted using standard GLM software, one can entirely avoid using specialised point process software, but the onus is on the user to ensure that quadrature points have been chosen in sufficient numbers and locations and that the appropriate quadrature weights have been assigned. The main advantage of this approach is that the user has greater control over what type of model is fitted – for example, as well as GLM, one could use any of its extensions (GAM, elastic net, CART, …). Along these lines, Fithian & Hastie (2013) proposed a simple algorithm based on weighted logistic regression that can be used to estimate slope coefficients in a Poisson PPM, referred to as infinitely weighted logistic regression (IWLR). But IWLR only gives a solution proportional to a point process, because the intercept term and hence the scale of the log-likelihood is arbitrary. A modification of this idea, which we call ‘downweighted Poisson regression’ (DWPR), is proposed in Section 5 of the Appendix S1. Given a random set of pseudo-absences, this reduces the steps of assigning quadrature weights and fitting the model to just a few lines. When using DWPR, the intercept and hence the likelihood are estimable, so we can use this technique to look at questions like how many quadrature points to select (as done in Fig. 2b).

Given that MAXENT has recently been shown to be proportional to a Poisson PPM (Aarts, Fieberg & Matthiopoulos 2012; Fithian & Hastie 2013; Renner & Warton 2013), MAXENT software can also be thought of as a means of fitting a Poisson PPM. This does, however, require some departures from the default MAXENT software settings, as described in Section 6 of the Appendix S1. Using MAXENT to fit a Poisson PPM has the advantages that it is familiar to many users, has a lot of nice mapping features and is easy to use. Example features include the interactive ‘explain’ tool which visualises the link between the mapped prediction and the model at any selected point, and maps that show whether environments in new regions or times of interest are within the training range of the modelled data. MAXENT can be run from R using the package DISMO, which allows streamlined data preparation and modelling, and the potential to use block resampling along the lines of Wenger & Olden (2012). A short tutorial for fitting PPMs using DISMO is presented in Section 6 of the Appendix S1.

A key issue, however, with implementations of Poisson PPM using GLM and MAXENT software is the lack of assumption checking tools. One should not take ‘on faith’ the assumption that there is no spatial dependence in the data beyond that explained by environmental variables included in the model. A related issue is the lack of a capacity to fit models that account for point interactions using GLM or MAXENT.

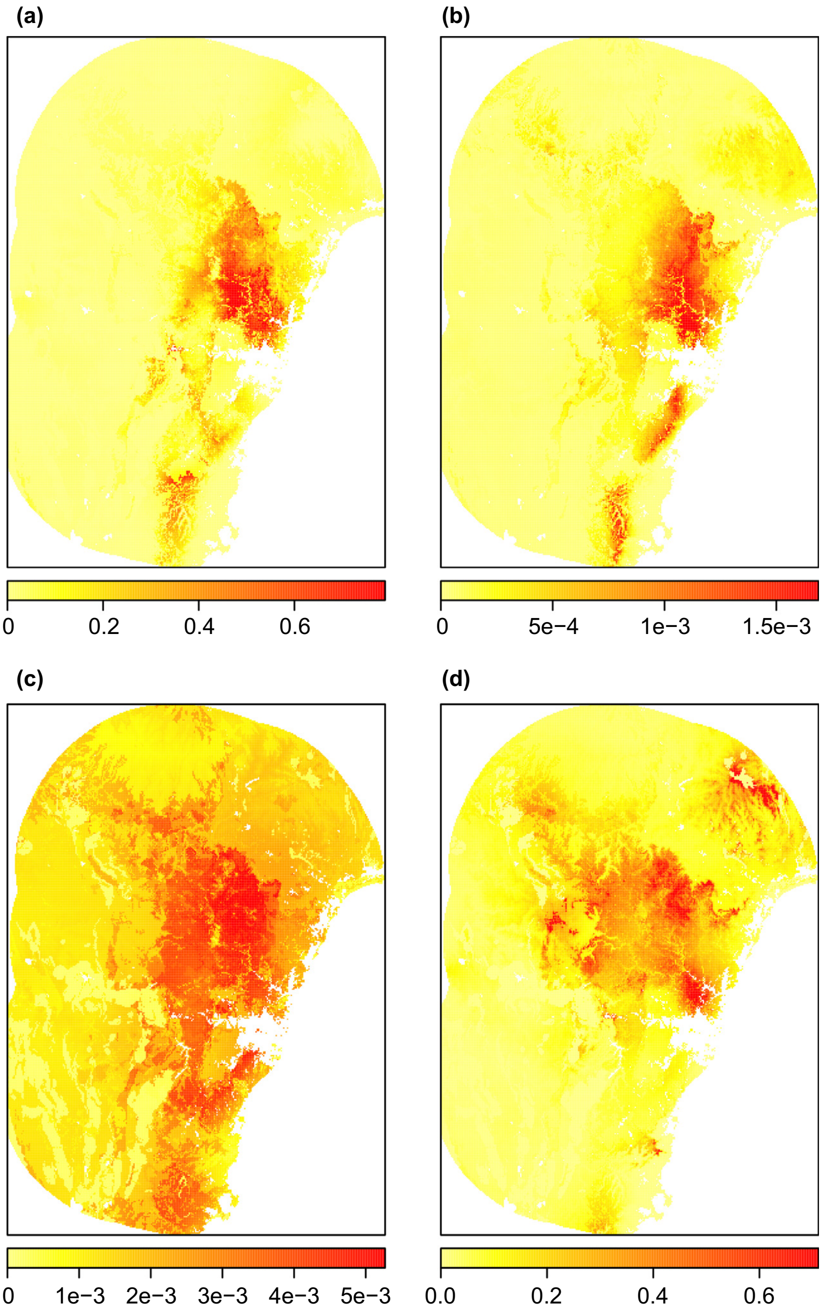

Fig. 5. Predicted intensity of *Eucalyptus sparsifolia* observations for a Poisson point process model fitted using (a) SPATSTAT or (b) MAXENT; (c) an area-interaction model fitted using PPMLASSO; and (d) a log-Gaussian Cox process fitted using R-INLA. Note that SPATSTAT and MAXENT results look quite similar, as they fit similar models, and that the PPMLASSO and R-INLA results, which account for spatial dependence, highlight additional areas of relatively high intensity further west.

Fitting Cox processes in a Bayesian framework, as in R-INLA and LGCP, has the advantage that spatial dependence can be estimated and accounted for in a flexible and statistically efficient way. However, the additional complexity of estimating the latent field results in much slower computation times as compared to competitors. Careful selection of the respective prior distribution for the Gaussian random field is important to avoid overfitting (Illian *et al.* 2013). But there is currently little guidance about prior selection – such guidelines are currently being developed for R-INLA.

Analysis of *Eucalyptus sparsifolia*

When fitting point process models to *Eucalyptus sparsifolia* data, results were broadly similar across methods of fitting Poisson PPMs (Fig. 5a–b), with some variation due to different decisions being made in default implementations (e.g. SPATSTAT does not apply a LASSO penalty, whereas MAXENT does). But as identified previously, there is evidence of spatial clustering between presence points in close proximity – at distances less than 10 *km*, the inhomogeneous *K*-function in Fig. 3a strays above its simulation envelope, and the point interaction coefficient from a fitted Cox process is significantly above zero (Fig. 4c). Further evidence of dependence can be seen in the mean of latent Gaussian field from the Cox process fit, which was sometimes large relative to its standard deviation (Fig. 4a–b). Maps produced by models which account for this spatial dependence (Fig. 5c and d) highlight areas of relatively higher intensity further west.

Table 1. Summary table of software properties

Property	SPATSTAT	PPMLASSO	IWLR	DWPR	MAXENT	R-INLA	LGCP
Regularisation	×	✓	✓	✓	✓[1]	×	×
Standard errors	✓[2]	×	✓[2]	✓[2]	×	✓	✓
Variable importance plots	×	×	×	×	✓	×	×
Diagnostic plots	✓	✓	×	×	×	×	×
Spatial dependence	✓	✓	×	×	×	✓	✓
Nonlinearity (e.g. smoothers)	✓	✓	✓	✓	✓	✓	✓
Scale invariant	✓	✓	×	✓	✓[3]	✓	✓

[1] LASSO only.
[2] For Poisson models only.
[3] Raw output only.

The analysis is largely consistent with pre-existing knowledge of the distribution of *Eucalyptus sparsifolia*, thought to prefer 'low nutrient soils, but some on medium and high nutrient soils, over a wide range of rainfall' (Hager & Benson 2010). The MAXENT response curve for rainfall (Fig. 6a) indicates suitability over a wide rainfall range, and several models (Poisson PPM and area-interaction models produced by PPMLASSO, and MAXENT) either dropped all rainfall terms or assessed them as relatively unimportant in describing the distribution of *E. sparsifolia* (results not shown). The species appears most strongly associated with low-nutrient high-quartz sedimentary soils and has low intensity in volcanic soils. See Section 9 of the Appendix S1 for a list of coefficient estimates.

Minimum annual temperature is strongly associated with *E. sparsofila* distribution, yet not mentioned by Hager & Benson (2010). The quadratic term is significantly negative in models produced by PPMLASSO and R-INLA, consistent with the response curve produced by MAXENT (Fig. 6a). This variable has implications for climate change projections, suggesting a substantial decrease in *E. sparsofila* intensity at the southern end of its range under warming scenarios (Fig. 6a).

A key distinction in the models produced by the different methods is in the number of significant variables (Section 9 of the Appendix S1). The Poisson PPM and area-interaction model fitted by PPMLASSO added 24 and 18 nonzero terms in the model, respectively, while the Cox process model produced by R-INLA added only seven, five of which were soil indicators. One possible explanation is that there may have been collinearity between the Gaussian random field and environmental predictors, dampening the environmental signal – such a 'spatial confounding' effect is seen elsewhere in spatial statistics, and adjustments can account for it when fitting spatial generalised linear mixed models (Hodges & Reich 2010; Hughes & Haran 2013). Extension of these ideas to address spatial confounding in Cox process models is a potential avenue for future research.

Extensions

To this point, the focus has been on spatial point processes, to describe a set of point locations in space. A number of potential variations on the method may be of interest to species distribution modellers.

A time stamp is often available with presences as well as their point location. It may be of interest to study the patterning of points jointly in space and time, thus fitting a spatio-temporal point process (Cressie & Wikle 2011). This seems especially relevant in telemetry (Hooten *et al.* 2013), where one would expect strong temporal dependence in spatial patterning (with individuals tending to be found near their last known location); thus, there is a strong case for modelling such point events jointly in space and time, in order to tease apart habitat preference from habitat availability. Another example where spatio-temporal modelling is of clear interest is in the study of invasive species not at equilibrium (Hooten & Wikle 2008).

Sometimes presences are observed along a network rather than in a region in space. For example, when modelling roadkill events (Ramp *et al.* 2005), presence points occur along a road network. Poisson PPMs can be fitted to point events arising along networks relatively easily; however, methods for studying and accounting for dependence in point events along networks are a little explored topic (Baddeley, Jammalamadaka & Nair 2014).

Simultaneously modelling data from multiple species has potential from a number of standpoints. Fithian *et al.* (2015) showed how estimating observer bias simultaneously across species can improve outcomes, making use of the idea that the sources of observer bias can often be reasonably assumed constant across species, given that this bias is a property of the observer more so than of the study species. Multispecies models could also be used to study species interaction, by specifying what is known as a marked point process model (Cressie 1993) which explicitly includes terms for species interaction.

An issue with presence-only data is data quality, and there are a number of potential extensions to take into account data of varying quality. For example, typically there is uncertainty in the spatial location of presence points, and locations are often assigned 'accuracy' scores to estimate this which can be accounted for in subsequent modelling (Hefley *et al.* 2014). Further, environmental data

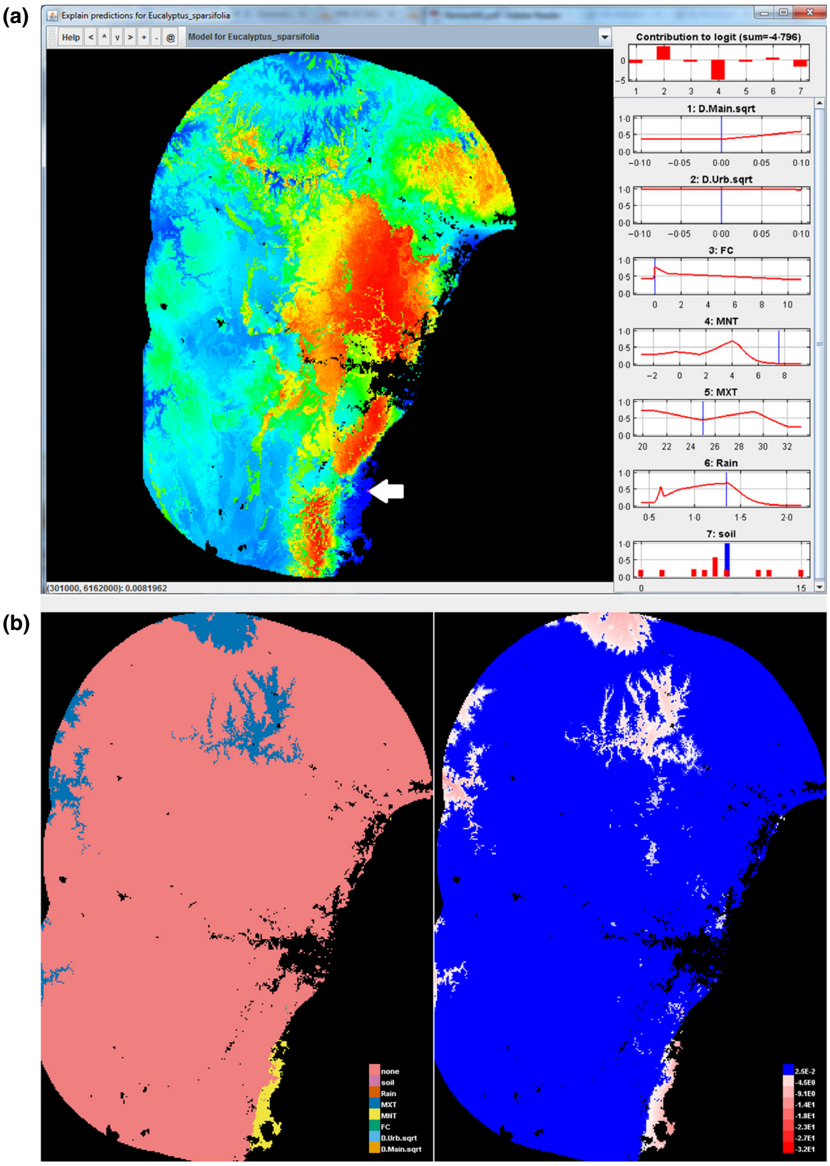

Fig. 6. Maps produced by MAXENT software to aid interpretation: (a) The explain tool – the location indicated by the arrow has low intensity, and the plots at the right suggest that the high minimum temperature is largely responsible. (b) Multivariate similarity surface plot (right) and most dissimilar variable plot (left) for climate change predictions – the redder colours in the left panel indicate areas that are most dissimilar to the environmental conditions used to build the model, while the right panel identifies which variables are most responsible for the dissimilarity. It seems minimum temperature may be limiting the distribution near the coast under the climate change model.

are also measured with uncertainty, maps of environmental data often being spatially interpolated from (often sparsely distributed) weather stations. This is a type of errors-in-variables problem (Carroll *et al.* 2012), and PPMs for such problems, while rare in the literature, should be a relatively straightforward extension given established methods for errors-in-variables approaches to GLM (Stoklosa *et al.* 2015). Presence-only data are furthermore subject to imperfect detectability, inducing biased estimates, but this bias can be reduced by building a hierarchical model including both presence-only data and independent presence–absence data (Dorazio 2014).

A key strength of point process models is that they operate at what is usually the most ecologically relevant of sampling levels – the level of the *individual*. This means they can (in principle) incorporate processes operating at the level of the individual, such as interactions between individual organisms, or covariates that vary across individuals. A limiting factor obviously is the quality of data at hand, but there is exciting potential in this framework.

Discussion

Point process modelling has been introduced as a natural framework for modelling presence-only data that is better understood as point events rather than as data from transects or grid cells.

A particular benefit of the PPM specification that we have highlighted is greater clarity around the issue of how to choose quadrature points, and the possibility of querying the data being analysed to verify that a given choice of quadrature points is appropriate (Fig. 2). We found in Section ‘How to choose quadrature points’ that for our example data, the number of quadrature points required for sufficient convergence of the log-likelihood (a change of <2) depended upon the selection method, but was closer to 100 000 than to the 10 000 usually advocated. Perhaps more quadrature points are needed when randomly chosen than when on a regular mesh, suggesting greater efficiency when using a regular mesh design, although at the cost of making it more difficult to quantify uncertainty in the approximation. This is related to classical results from survey sampling, where systematic sampling is often more efficient than random sampling under serial correlation, but for which it is more difficult to quantify uncertainty (Cochran 1946).

Our hope is that approaching the 'pseudo-absence problem' via numerical quadrature will shift attention of analysis away from quadrature point choice and towards where it belongs – developing and interpreting a plausible model for intensity as a function of environment and possibly observer bias variables.

PPMs as in this paper can be understood as applying regression methods to point event data, and as such, issues that arise in other areas of regression analysis apply equally well here. For example, there is increasing awareness of a dichotomy between prediction and explanation (Elith & Leathwick 2009) – many SDM researchers are interested primarily in the prediction problem and hence are mostly concerned with using a method which has good predictive performance. This leads the user down the road towards regularised methods (such as in PPMLASSO) or model averaging (Araújo & New 2007). Others are interested primarily in explanation – identifying key associations between environmental variables and a species. This leads the user to put greater focus on appropriately accounting for different sources of uncertainty, which requires careful consideration of the question of spatial dependence in the data, and potential problems like multicollinearity (Zuur, Ieno & Elphick 2010).

Care must be taken when interpreting a fitted PPM concerning what is actually being modelled, with particular reference to how the data were collected. For example, in order to interpret intensity as relative abundance of individuals per unit area, the intensity of presence records should be proportional to the intensity of individuals (i.e. abundance) of the species. Contrast this with how observers might collect data and how data are entered into online data bases. First, observers may tend to go to a site and only record one individual even though many are present, so individuals in abundant sites are under-represented. In that case, really one is modelling relative intensity of occupied sites or, equivalently, relative probability of presence (which opens a can of worms, since the extent of a site may be unknown or vary between observers). Secondly, in data aggregation services such as the Global Biodiversity Information Facility (GBIF, Belbin *et al.* 2013), records arrive from multiple providers. Duplicate records may represent the same individual, for instance, because the same record has been contributed through multiple channels or because different specimens of the same individual were lodged in different museums or herbaria. These duplicate records do not always have exactly the same coordinates because of variable data handling practices. These two examples illustrate typical problems in dealing with the realities of presence-only data, emphasising that it is important that appropriate data cleaning and model interrogation is considered, and that the interpretation of the final model not stray far from the data used to construct it.

While this paper has focussed on the merits from a model-fitting perspective of taking a point process approach, there are broader advantages afforded by a coherent modelling framework for presence-only data. From a technical perspective, PPMs can serve as a model for generation of simulated presence-only data, with and without point interactions (the SPATSTAT package makes this quite straightforward). From an analyst's perspective, PPMs offer a way forward regarding the assessment of goodness-of-fit to presence-only data, obviating the need to adapt tools like ROC curves to the presence-only context (Jiménez-Valverde 2012). Apart from a suite of diagnostic plots for goodness-of-fit, likelihood-based procedures can be used to quantify predictive success, for example Kullback-Leibler distance. From an ecologist's perspective, rather than aggregating to an arbitrary sampling unit, one can specify a model for location and behaviour of individual organisms, for example, incorporate covariate information particular to individuals into analyses, where available. In this respect, the point process framework offers an exciting platform that can be used to study ecological processes.

Data accessibility

Locations of *Eucalyptus sparsifolia*, including the 230 presence-only locations used in this paper, are available from http://www.bionet.nsw.gov.au.

Environmental data for Blue Mountains region: DRYAD entry doi:10.5061/dryad.985s5.

References

Aarts, G., Fieberg, J. & Matthiopoulos, J. (2012) Comparative interpretation of count, presence-absence and point methods for species distribution models. *Methods in Ecology and Evolution*, **3**, 177–187.

Araújo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, **22**, 42–47.

Baddeley, A.J. & van Lieshout, M.N.M. (1995) Area-interaction point processes. *Annals of the Institute of Statistical Mathematics*, **47**, 601–619.

Baddeley, A. & Turner, R. (2005) Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, **12**, 1–42.

Baddeley, A.J., Møller, J. & Waagepetersen, R. (2000) Non- and semiparametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, **54**, 329–350.

Baddeley, A.J., Turner, R., Møller, J. & Hazelton, M. (2005) Residual analysis for spatial point processes. *Journal of the Royal Statistical Society, Series B*, **67**, 617–666.

Baddeley, A., Berman, M., Fisher, N.I., Hardegen, A., Milne, R.K., Schuhmacher, D., Shah, R. & Turner, R. (2010) Spatial logistic regression and change-of-support in Poisson point processes. *Electronic Journal of Statistics*, **4**, 1151–1201.

Baddeley, A.J., Chang, Y.M., Song, Y. & Turner, R. (2012) Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and Its Interface*, **5**, 221–236.

Baddeley, A., Chang, Y.-M., Song, Y. & Turner, R. (2013) Residual diagnostics for covariate effects in spatial point process models. *Journal of Computational and Graphical Statistics*, **22**, 886–905.

Baddeley, A., Jammalamadaka, A. & Nair, G. (2014) Multitype point process analysis of spines on the dendrite network of a neuron. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**, 673–694.

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Belbin, L., Daly, J., Hirsch, T., Hobern, D. & La Salle, J. (2013) A specialists audit of aggregated occurrence records: an aggregators perspective. *ZooKeys*, **305**, 67–76.

Berman, M. & Turner, T.R. (1992) Approximating point process likelihoods with GLIM. *Journal of the Royal Statistics Society, Series C*, **41**, 31–38.

Besag, J. (1977) Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute*, **47**, 77–91.

Burman, P., Chow, E. & Nolan, D. (1994) A cross-validatory method for dependent data. *Biometrika*, **81**, 351–358.

Carroll, R.J., Ruppert, D., Stefanski, L.A. & Crainiceanu, C.M. (2012) *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press, Boca Raton, FL.

Chakraborty, A., Gelfand, A.E., Wilson, A.M., Latimer, A.M. & Silander, J.A. (2011) Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society, Series C*, **60**, 757–776.

Chefaoui, R.M. & Lobo, J.M. (2008) Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, **210**, 478–486.

Cochran, W.G. (1946) Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics*, **17**, 164–177.

Cressie, N.A.C. (1993) *Statistics for Spatial Data*. John Wiley & Sons, New York, NY.

Cressie, N. & Wikle, C.K. (2011) *Statistics for Spatio-temporal Data*. John Wiley & Sons, Hoboken.

Davis, P.J. & Rabinowitz, P. (1984) *Methods of Numerical Integration*. Academic Press, Orlando, FL.

Diggle, P. (2003) *Statistical Analysis of Spatial Point Patterns*. Oxford University Press, New York, NY.

Dorazio, R.M. (2012) Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, **68**, 1303–1312.

Dorazio, R.M. (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, **23**, 1472–1484.

Dormann, C.F. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16**, 129–138.

Efron, B. & Tibshirani, R. (1993) *An Introduction to the Bootstrap*. CRC press, Boca Raton, FL.

Elith, J. & Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

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

Fithian, W. & Hastie, T. (2013) Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, **7**, 1917–1939.

Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, **6**, 424–438.

Guan, Y. (2006) A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association*, **101**, 1502–1512.

Guan, Y. (2008) On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, **103**, 1238–1247.

Guan, Y. & Shen, Y. (2010) A weighted estimating equation approach for inhomogeneous spatial point processes. *Biometrika*, **97**, 867–880.

Hager, T. & Benson, D. (2010) The Eucalypts of the Greater Blue Mountains World Heritage Area: distribution, classification and habitats of the species of *Eucalyptus*, *Angophora* and *Corymbia* (family Myrtaceae) recorded in its eight conservation reserves. *Cunninghamia*, **10**, 425–444.

Hastie, T. & Tibshirani, R. (1990) *Generalized Additive Models*. Chapman & Hall, Boca Raton.

Hastie, H., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY.

Hefley, T.J., Baasch, D.M., Tyre, A.J. & Blankenship, E.E. (2014) Correction of location errors for presence-only species distribution models. *Methods in Ecology and Evolution*, **5**, 207–214.

Hodges, J.S. & Reich, B.J. (2010) Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, **64**, 325–334.

Hooten, M.B. & Wikle, C.K. (2008) A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, **15**, 59–70.

Hooten, M.B., Hanks, E.M., Johnson, D.S. & Alldredge, M.W. (2013) Reconciling resource utilization and resource selection functions. *Journal of Animal Ecology*, **82**, 1146–1154.

Hughes, J. & Haran, M. (2013) Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 139–159.

Illian, J.B., Møller, J. & Waagepetersen, R.P. (2009) Hierarchical spatial point process analysis for a plant community with high biodiversity. *Environmental and Ecological Statistics*, **16**, 389–405.

Illian, J.B., Martino, S., Sørbye, S.H., Gallego-Fernández, J.B., Zunzunegui, M., Esquivias, M.P. & Travis, J.M. (2013) Fitting complex ecological point process models with integrated nested Laplace approximation. *Methods in Ecology and Evolution*, **4**, 305–315.

Jiménez-Valverde, A. (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, **21**, 498–507.

Møller, J., Syversveen, A.R. & Waagepetersen, R.P. (1998) Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, **25**, 451–482.

McCullagh, P. & Nelder, J. (1989) *Generalized Linear Models*. Chapman and Hall, London.

McDonald, L., Manly, B., Huettmann, F. & Thogmartin, W. (2013) Location-only and use-availability data: analysis methods converge. *Journal of Animal Ecology*, **82**, 1120–1124.

NSW Office of Environment and Heritage (2012) Atlas of NSW Wildlife database. Data accessed 31/05/2012.

Pearce, J.L. & Boyce, M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**, 405–412.

Pearson, R.G., Phillips, S.J., Loranty, M.M., Beck, P.S., Damoulas, T., Knight, S.J. & Goetz, S.J. (2013) Shifts in Arctic vegetation and associated feedbacks under climate change. *Nature Climate Change*, **3**, 673–677.

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

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

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, AustriaISBN 3-900051-07-0

Ramp, D., Caldwell, J., Edwards, K.A., Warton, D. & Croft, D.B. (2005) Modelling of wildlife fatality hotspots along the Snowy Mountain Highway in New South Wales, Australia. *Biological Conservation*, **126**, 474–490.

Renner, I.W. & Warton, D.I. (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**, 274–281.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S., Popovic, G. & Warton, D. (2015) Data from: Point process models for presence-only analysis – a review. *Dryad Digital Repository*, doi: 10.5061/dryad.985s5.

Ripley, B.D. (1977) Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 172–212.

Rue, H., Martino, S. & Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 319–392.

Slavich, E., Warton, D.I., Ashcroft, M.B., Gollan, J.R. & Ramp, D. (2014) Topoclimate versus macroclimate: how does climate mapping methodology affect species distribution models and climate change projections? *Diversity and Distributions*, **20**, 952–963.

Stoklosa, J., Daly, C., Foster, S.D., Ashcroft, M.B. & Warton, D.I. (2015) A climate of uncertainty: accounting for error in climate variables for species distribution models. *Methods in Ecology and Evolution*, **6**, 412–423.

Taylor, B.M. & Diggle, P.J. (2014) INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, **84**, 2266–2284.

Taylor, B.M., Davies, T.M., Rowlingson, B.S. & Diggle, P.J. (2013) lgcp: an R package for inference with spatial and spatio-temporal log-Gaussian Cox processes. *Journal of Statistical Software*, **52**, 1–40.

Warton, D.I. & Aarts, G. (2013) Advancing our thinking in presence-only and used available analysis. *Journal of Animal Ecology*, **82**, 1125–1134.

Warton, D.I. & Shepherd, L.C. (2010) Poisson point process models solve the “pseudo-absence problem" for presence-only data in ecology. *Annals of Applied Statistics*, **4**, 1383–1402.

Warton, D.I., Renner, I.W. & Ramp, D. (2013) Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS One*, **8**, e79168.

Wenger, S.J. & Olden, J.D. (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, **3**, 260–267.

Widom, B. & Rowlinson, J.S. (1970) New model for the study of liquid–vapor phase transitions. *The Journal of Chemical Physics*, **52**, 1670–1984.

Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, **1**, 3–14.

Received 20 October 2014; accepted 12 January 2015
Handling Editor: Robert B. O'Hara

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Full description of environmental covariates, derivation of likelihood formulae, and detailed software tutorials.

Supplementary Material for "Point process models for presence-only analysis – a review"

March 1, 2015

1 Description of envrionmental covarites

Point process models fitted in the main text and in the Supplementary Material utilise five environmental covariates and two observer bias covariates. The environmental covariates include the number of fires since 1943, minimum and maximum annual temperature, annual rainfall, and a categorical covariate describing the type of soil. The values of the soil covariate are presented in Table 1.

Table 1: Description of soil categories used for model fitting.

Soil category	Description
2	siliceous (white) sandplains
5	low quartz sedimentary, ultramafic igneous and metamorphics, aeolian (red) sandplains, or limestone
6	felsic intrusives
7	high quartz sedimentary
8	mafic volcanics and intrusives
11	floodplain alluvium, lacustrine sediments, or estuarine sediments
12	residual alluvial sands or residual alluvial/colluvial sand and gravel
15	felsic volcanics

2 Point process model likelihood

2.1 Homogeneous Poisson point process

The simplest example of a PPM is a homogeneous Poisson point process, which is of little use in SDM but serves as a good conceptual starting point. A homogeneous Poisson point process is characterised by two conditions: first, all points are distributed independently and uniformly over $\mathcal{A}$, and second, the number of presences m is the observed realisation of a Poisson random variable M with mean $\lambda|\mathcal{A}|$ (Cressie, 1993; Diggle, 2003). As a consequence of these conditions, the intensity λ is constant throughout the region $\mathcal{A}$. More specifically, the first condition implies that given a point, the density of its location s is:

$$f(s) = \frac{1}{|\mathcal{A}|}, s \in \mathcal{A}.$$

Hence, conditional on $M = m$, the joint density of the m points $s_1, \dots, s_m$ is:

$$f(s_1, \dots, s_m | M = m) = \frac{1}{|\mathcal{A}|^m}. \tag{1}$$

The second condition describes the probability structure for the number of points m:

$$P(M = m) = \frac{e^{-\lambda|\mathcal{A}|}(\lambda|\mathcal{A}|)^m}{m!}, \; m = 0, 1, 2, \dots. \tag{2}$$

The form of the likelihood equation can be derived from (1) and (2):

$$\begin{aligned} L(\boldsymbol{\beta}; \mathbf{s}_P) &= m! f(s_1, \dots, s_m | M = m) P(M = m) \\ &= m! \frac{1}{|\mathcal{A}|^m} \frac{e^{-\lambda|\mathcal{A}|}(\lambda|\mathcal{A}|)^m}{m!} \\ &= \lambda^m e^{-\lambda|\mathcal{A}|}. \end{aligned} \tag{3}$$

The log-likelihood is found by taking the logarithm of (3):

$$l(\boldsymbol{\beta}; \mathbf{s}_P) = m \ln \lambda - \lambda|\mathcal{A}|.$$

Parameters of Poisson point process models are typically estimated by maximum likelihood (Cressie, 1993). In the case of a homogeneous PPM, the estimator of λ that maximises the likelihood is $|\mathcal{A}|/m$, which also happens to be the method-of-moments estimator. (The equivalence of maximum likelihood and method-of-moments estimators actually applies to all types of Poisson process.)

2.2 Inhomogeneous Poisson point process

The assumptions of an inhomogeneous Poisson point process are that (a) point events are independent of each other, which can be shown to imply that the total number of points in the study region is a Poisson random variable and (b) that the intensity $\lambda(s)$ varies spatially according to environmental conditions $\mathbf{x}(s)$ and is therefore indexed by location s.

The conditional independence of the points implies that given a point, the density of its location s is:

$$f(s) = \frac{\lambda(s)}{\int_{\mathcal{A}} \lambda(s)ds}, s \in \mathcal{A}.$$

Hence, conditional on the number of points $M = m$, the joint density of the m points $s_1, \ldots, s_m$ is:

$$f(s_1, \ldots, s_m | M = m) = \frac{\prod_{i=1}^{m} \lambda(s_i)}{(\int_{\mathcal{A}} \lambda(s)ds)^m}. \tag{4}$$

Because the number of points m is an observed realisation of a Poisson random variable M with mean $\int_{\mathcal{A}} \lambda(s)ds$, we describe the probability structure for the number of points m as follows:

$$P(M = m) = \frac{e^{-\int_{\mathcal{A}} \lambda(s)ds}(\int_{\mathcal{A}} \lambda(s)ds)^m}{m!}, \; m = 0, 1, 2, \ldots. \tag{5}$$

The form of the likelihood equation can be derived from (4) and (5):

$$\begin{aligned} L(\boldsymbol{\beta}; \mathbf{s}_P) &= m! f(s_1, \ldots, s_m | M = m) P(M = m) & (6) \\ &= m! \frac{\prod_{i=1}^{m} \lambda(s_i)}{(\int_{\mathcal{A}} \lambda(s)ds)^m} \frac{e^{-\int_{\mathcal{A}} \lambda(s)ds}(\int_{\mathcal{A}} \lambda(s)ds)^m}{m!} \\ &= e^{-\int_{\mathcal{A}} \lambda(s)ds} \prod_{i=1}^{m} \lambda(s_i). & (7) \end{aligned}$$

The $m!$ factor in (6) is included because we consider the m points of $\mathbf{s}_P$ to be unordered and thus there are $m!$ arrangements of the m points. An expression of the form of (6) is sometimes called a Janossy density (Daley & Vere-Jones, 1988). The log-likelihood is found by taking the logarithm of (7):

$$l(\boldsymbol{\beta}; \mathbf{s}_P) = \sum_{i=1}^{m} \ln \lambda(s_i) - \int_{\mathcal{A}} \lambda(s)ds.$$

3 spatstat

Perhaps the most well-known software for fitting PPMs is `spatstat` (Baddeley & Turner, 2005), an `R` package that offers a large suite of tools for analysing spatial point patterns including plotting, model fitting, diagnostic checks, and simulation. In this Section we provide code for each step required in fitting PPMs and checking results with spatstat.

3.1 Setting up the study window

```
>library(spatstat)
>load("Eucalyptus sparsifolia Atlas 2012.RData") #Contains X and Y
>load("Quad1000.RData") #Contains quad
>ux = sort(unique(quad$X))
>uy = sort(unique(quad$Y))
>nx = length(ux)
>ny = length(uy)
>col.ref = match(quad$X, ux)
>row.ref = match(quad$Y, uy)
>all.vec = rep(NA, max(row.ref)*max(col.ref))
>vec.ref = (col.ref - 1)*max(row.ref) + row.ref
>all.vec[vec.ref] = 1
>Sydney.mask = matrix(all.vec, max(row.ref), max(col.ref),
 dimnames = list(uy, ux))
>Sydney.win = as.owin(im(Sydney.mask, xcol = ux, yrow = uy))
```

3.2 Make point pattern and quadrature scheme

By default, quadrature points are randomly generated in spatstat, although user-entered quadrature points can be used instead as below, where they are contained in the data frame quad.

```
>ppp.dat = ppp(X, Y, window = Sydney.win, check = FALSE)
>quads = ppp(quad$X, quad$Y, window = Sydney.win)
>Q = quadscheme(data = ppp.dat, dummy = quads, method = "grid",
 ntile = c(nx, ny), npix = c(nx, ny))
```

3.3 Setting up covariate lists

We set up a list of covariates `int.list` for fitting the model `ft.int` and a corresponding list `pred.list` with distance-based covarates set to zero for prediction.

```
>X.des = cbind(poly(quad$FC, quad$MNT, quad$MXT, quad$Rain, degree = 2,
 raw = TRUE), poly(sqrt(quad$D.Main), sqrt(quad$D.Urb), degree = 2,
 raw = TRUE), quad$soil)
>int.list = list()
>for (i in 1:dim(X.des)[2])
>{
>all.vec = rep(NA, max(row.ref)*max(col.ref))
>vec.ref = (col.ref - 1)*max(row.ref) + row.ref
>all.vec[vec.ref] = X.des[,i]
>int.list[[i]] = im(matrix(all.vec, max(row.ref), max(col.ref),
 dimnames = list(uy, ux)), xcol = ux, yrow = uy)
>}
>names(int.list) = paste("V", 1:dim(X.des)[2], sep = "")
>pred.list = int.list
>set.0 = 15:19 #Variables to set to 0
>for (v in set.0)
>{
>pred.list[[v]]$v = 0*pred.list[[v]]$v
>}
```

3.4 Fit Poisson point process model

```
>int.form = as.formula(paste("~", paste(names(int.list), collapse = "+")))
>ft.int = ppm(Q, trend = as.formula(int.form), covariates = int.list)
```

The fitted coefficients of `ft.int` are presented in Table 3.

3.5 Inference for coefficients and models

We can make inferences about coefficients in a `ppm`, or compare several nested `ppm` objects, using `summary` and `anova` tools as they are used for standard regression objects. For example, the interaction coefficient between fire count and annual rainfall is small relative to its standard error, and an ANOVA comparing a model with and without this interaction term (`ft.int` and `ft.no.fc.Rain`) suggests it does not explain significant variation in intensity:

```
>anova(ft.no.fc.Rain, ft.int)
```

Note however that these inferential tools rely on the assumption of independence of presence points, and they are sensitive to violations of this assumption.

The `anova` function may also be used as a way to assess variable importance, by comparing the full model with models fitted without each of the covariates. Table 2 contains the differences in deviance between the full model and models fitted without each of the covariates. It suggests that the temperature variables and soil type are the most important environmental covariates in determining the distribution of *Eucalyptus sparsifolia*.

Table 2: Difference in deviance between the model with all covariates `ft.int` and models without each of the covariates, as produced by the `anova` function in `spatstat`.

Covariate	Difference in Deviance
Fire count since 1943	31.038
Minimum annual temperature	122.690
Maximum annual temperature	81.444
Annual rainfall	58.818
Soil type	85.490
Distance from the nearest main road	49.273
Distance from the nearest urban area	80.533

3.6 Fit area-interaction model

The `profilepl` function can be used to determine the appropriate interaction radius for fitting an area-interaction model.

```
>rs = data.frame(seq(2, 10, 0.2))
>names(rs) = "r"
>plike = profilepl(rs, AreaInter, Q, trend = int.form, covariates = int.list)
```

We now fit an area-interaction model with the suggested radius of 2km:

```
>ft.ai.2 = ppm(Q, trend = as.formula(int.form), covariates = int.list,
 interaction = AreaInter(2))
```

3.7 Model Diagnostics

There are a number of diagnostic tools available in `spatstat` to identify departures of model assumptions. The generic `envelope` function generates Monte Carlo simulation envelopes for some summary function of a fitted model. In the main text, we fitted a simulation envelope for the inhomogeneous K-function in Figure 3a as follows:

```
>envelope(ft.int, fun = Kinhom, nrank = 26, nsim = 1001)
```

Residual plots may be produced with a call to the `diagnose.ppm` function. Figure 3b of the main text is a plot of smoothed Pearson residuals for the model `ft.ai.2`, produced as follows:

```
>diagnose.ppm(ft.ai.2, which = "smooth", type = "Pearson", labcex = 1)
```

Figure 3b exhibits some pattern, with areas along the mid-coast and northwest showing the highest positive residuals, and the south showing the most negative residuals. Cumulative Pearson residuals for both increasing longitude (x) and latitude (y) do not show any significant departures from Monte Carlo simulation envelopes in Figure 3c-d of the main text, so the model fit may be assumed to be reasonable. The code to produce the plots in Figure 3c-d is as follows:

```
>diagnose.ppm(ft.ai.2, which = "x", type = "Pearson", compute.sd = TRUE)
>diagnose.ppm(ft.ai.2, which = "y", type = "Pearson", compute.sd = TRUE)
```

3.8 Maps of predicted intensity

```
>pred.ai.2 = predict(ft.ai.2, covariates = pred.list, ngrid = c(ny, nx))
>pred.int = predict(ft.int, covariates = pred.list, ngrid = c(ny, nx))
>plot(pred.int) #Fig. 1 (left)
>plot(pred.ai.2) #Fig. 1 (right)
```

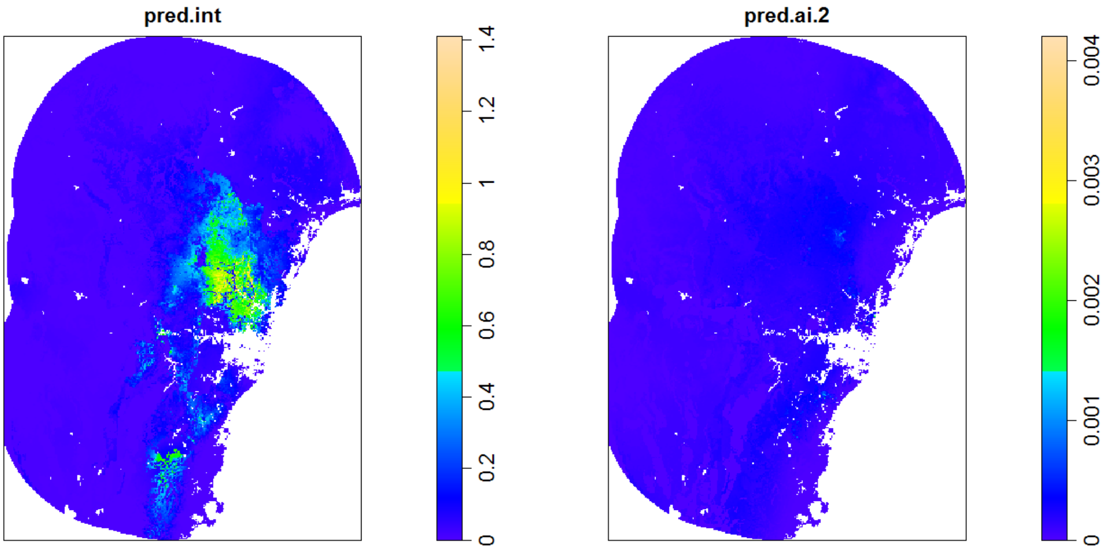

Figure 1: Maps of the spatial trend for the model `ft.int` (left) and `ft.ai.2` (right). The distance variables have been set to zero to account for observer bias.

3.9 Leverage and Influence

The `leverage` and `influence` functions provide insight into places and observed point locations in the study region that have potentially high impact on the fitted model:

```
>plot(leverage(ft.int)) #Fig. 2 (left)
>plot(influence(ft.int)) #Fig. 2 (right)
```

The interpretation is slightly different – places near the coast appear to have high leverage, suggesting that the fitted intensity would be most greatly increased if a species were to be observed there. Points with high influence suggest that if they were deleted, the change in maximum likelihood would be the greatest.

3.10 Validating the form of covariates

The `parres` function validates the parametric form of covariates by computing smoothed partial residuals, potentially suggesting a transformation if the fitted residuals deviate from the confidence bounds. Figure 3 shows the partial residual plot for maximum temperature `V3` in the fitted PPM `ft.int`, produced as follows:

```
>plot(parres(ft.int, "V3")) #Fig. 3
```

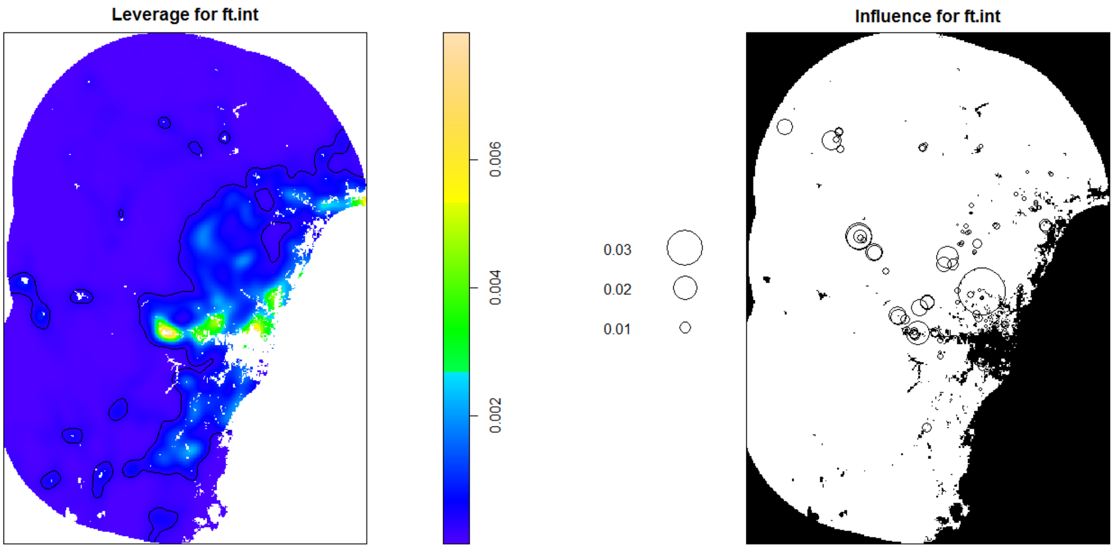

Figure 2: Maps of the leverage (left) and influence (right) of `ft.int`.

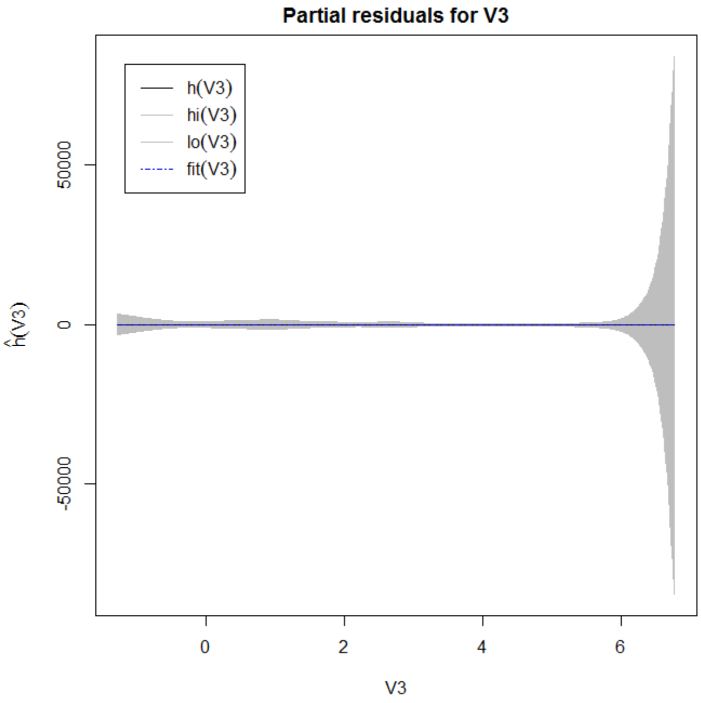

Figure 3: Smoothed partial residuals for the covariate "V3" (maximum temperature) for `ft.int`. The observed residuals fall well within the confidence bounds, suggesting that V3 does not need to be transformed.

Because the observed residuals follow the expected trend, there is no misspecification of maximum temperature. Similar plots for all 19 continuous covariates suggest that the model `ft.int` is properly specified.

4 ppmlasso

The spatstat package fits point process regression models by maximum likelihood, but there has been recent interest in regularised estimation of species distribution models, *e.g.* using the LASSO (Phillips *et al.*, 2006; Reineking & Schröder, 2006). The ppmlasso package on R (Renner & Warton, 2013) is a companion package to spatstat that permits coefficient penalisation via LASSO, adaptive LASSO (Zou, 2006), and elastic net (Zou & Hastie, 2005), while retaining much of the functionality originally available in the spatstat package. Various criteria are available for choosing how large the penalty should be, including AIC, BIC, and generalised cross validation (GCV). The same model is fitted as in spatstat – assuming intensity is a loglinear function of covariates – but ppmlasso fits PPMs via penalised likelihood, a method which is more numerically stable and robust to overfitting. The addition of a penalty term introduces bias but reduces variance (Hastie *et al.*, 2009) in an attempt to improve predictive performance.

4.1 Finding the appropriate spatial resolution for analysis

The findres function allows users to judge which spatial resolution to use for analysis based on convergence of the log-likelihood. Figure 2a of the main text was produced using the following code:

```
>library(ppmlasso) #Version 1.1
>load("Quad100m.RData") #Contains quad
>load("Eucalyptus sparsifolia Atlas 2012.RData") #Contains X and Y
>sp.xy = data.frame(X, Y)
>ppm.form = ~ poly(FC, MNT, MXT, Rain, degree = 2, raw = TRUE)
 + poly(sqrt(D.Main), sqrt(D.Urb), degree = 2, raw = TRUE)
 + soil
>scales = c(0.5, 1, 2, 4, 8, 16)
>findres(scales, formula = ppm.form, sp.xy = sp.xy, env.grid = quad)
```

This function suggests a spatial resolution of 1km is appropriate for analysis. For the categorical soil variable, a binary covariate needs to be made for each soil type. Version 1.1 of ppmlasso performs this step automatically.

4.2 Fitting a regularisation path of point process models

A regularisation path of Poisson point process models or area-interaction models may be fitted with a call to the `ppmlasso` function.

A Poisson PPM with a LASSO penalty that optimises non-linear GCV (Fu, 2005) can be fitted using the following line:

```
>quad.1k = sample.quad(quad, 1)
>ppm.fit = ppmlasso(ppm.form, sp.xy = sp.xy, env.grid = quad.1k,
 sp.scale = 1, criterion = "nlgcv")
```

The LASSO penalty forces uninformative covariates out of the model, so the process of model-fitting automatically does variable selection. This particular fit set three coefficients to zero, including the interaction term between fire count and annual rainfall, suggesting (as previously) that this interaction was not helpful for predicting the location of *Eucalyptus sparsifolia*, after other environmental variables.

4.3 Block cross-validation

Version 1.1 of `ppmlasso` includes block cross-validation as a method for choosing the LASSO penalty. The area-interaction model with radius 2km and LASSO penalty chosen by 5-fold block cross-validation in the main text was fitted as follows:

```
>final.fit = ppmlasso(ppm.form, sp.xy = sp.xy, env.grid = quad.1k,
 sp.scale = 1, criterion = "blockCV", n.blocks = 5, block.size = 32)
```

The fitted coefficients of this model `final.fit` are presented in Table 3.

5 Weighted GLMs

5.1 Infinitely Weighted Logistic Regression (IWLR)

Because the Poisson distribution approximates the binomial when probabilities are small, Fithian & Hastie (2013) suggested assigning large weights W to quadrature points to ensure predicted probabilities are small and approximate equivalence holds between logistic regression and a Poisson PPM, thereby validating use of logistic regression as an estimation procedure for a Poisson PPM. To fit an IWLR model in R requires only two lines of code. For $W = 10^6$:

```
> up.wt = (10^6)^(1 - Pres)
> iwlr = glm(Pres ~ X.des, family = binomial(), weights = up.wt)
```

Here, Pres is a binary vector indicating whether a location corresponds to a species location or a quadrature point, and X.des is the corresponding design matrix of covariates at species locations and quadrature points. Strictly speaking, spatstat and ppmlasso are also simple to fit, requiring a single line of code, but the difference here is that a generic function (glm) is used to do the work which will be already familiar to most users. However, with IWLR the onus is on the user to generate a sample of quadrature points (randomly, along a regular mesh or through some other scheme) in order to supply Pres and X.des to the glm call. While a glm was used above, Fithian & Hastie (2013) argue that any regression function designed for presence-absence data could be used in its place, provided that it permits weighted observations.

A disadvantage of IWLR is that it only gives a solution proportional to a point process because the intercept term is arbitrarily reduced by $\ln W$, and hence the scale of the log-likelihood is likewise arbitrary. Thus we can no longer use the method of choosing quadrature points previously advocated, via likelihood convergence, without suitable adjustments.

5.2 Downweighted Poisson Regression (DWPR)

We propose "downweighted Poisson regression" (DWPR) as a modification of IWLR that addresses these drawbacks. For DWPR, we set the weights equal to some small value (ϵ) at presence locations, but at quadrature points we set the weights equal to the area of the study region (in our case, 86,227km^2) divided by the number of quadrature points:

```
> p.wt = rep(1.e-6, length(Pres))
> p.wt[Pres == 0] = 86227/sum(Pres == 0)
> dwpr = glm(Pres/p.wt ~ X.des, family = poisson(), weights = p.wt)
```

This has similar advantages to IWLR, but additionally estimates the intercept term of the PPM correctly, and hence also fitted values and the log-likelihood, without the problem of scale dependence. We can then use this formulation to look at the question of how many randomly chosen quadrature points are needed to ensure likelihood convergence.

5.3 Assessing the variability in likelihood for different numbers of quadrature points

In Figure 2b of the main text, increasing numbers of randomly sampled quadrature points were used to fit PPMs using DWPR to judge the number of quadrature points for which the likelihood reasonably converged.

To generate subsets of quadrature points of increasing size:

```
>load("Quad100m.RData") #Contains bigquad
>n.quad = c(1000, 2000, 5000, 10000, 20000, 50000, 100000,
 200000, 500000)
>quad.inc = sample(1:dim(bigquad)[1], 1000)
>assign(paste("quad.", n.quad[1], sep = ""), bigquad[quad.inc[1:n.quad[1]],])
>for (i in 2:length(n.quad))
>{
>quad.inc = c(quad.inc, sample(setdiff(1:dim(bigquad)[1], quad.inc),
 (n.quad[i] - n.quad[i - 1])))
>assign(paste("quad.", n.quad[i], sep = ""), bigquad[quad.inc[1:n.quad[i]],])
>}
```

To compare the likelihood of PPMs fitted using downweighted Poisson regression:

```
>load("Eucalyptus sparsifolia Atlas 2012.RData") # Species Data
>sp.dat = data.frame(X, Y, D.Main, D.Urb, FC, MNT, MXT, Rain, soil)
>sp.dat$Pres = 1
>loglik = rep(NA, length(n.quad))
>for (i in 1:length(n.quad))
>{
>quad = get(paste("quad.", n.quad[i], sep = ""))
>quad$Pres = 0
>all.dat = data.frame(rbind(sp.dat, quad))
>X.des = as.matrix(cbind(poly(all.dat$FC, all.dat$MNT, all.dat$MXT,
 all.dat$Rain, degree = 2, raw = TRUE), poly(sqrt(all.dat$D.Main),
 sqrt(all.dat$D.Urb), degree = 2), all.dat$soil))
>p.wt = rep(1.e-8, dim(all.dat)[1])
>p.wt[all.dat$Pres == 0] = 86227/n.quad[i]
>z = all.dat$Pres/p.wt
>dwpr = glm(z ~ X.des, family = poisson(), weights = p.wt)
>mu = dwpr$fitted
>loglik[i] = sum(p.wt*(z*log(mu) - mu))
>}
>plot(n.quad, loglik, log = "x", type = "o")
```

6 MAXENT

When fitting a PPM using MAXENT, users need to adjust the default settings. Here we outline which settings best mimic the methods presented so far. Since Poisson PPMs are modelling intensity, the number of presence records is of interest. Hence if MAXENT is run with gridded environmental data, the default “Settings” option that reduces presence records to one per grid cell (“remove duplicate presence records”) needs to be unchecked. Alternatively if the samples with data (“SWD”) mode is used, the defaults are correct for PPMs: *i.e.* all points are retained. The raw output (not the default logistic output) can then be interpreted as a relative intensity, proportional to a Poisson PPM intensity. The maximum number of quadrature points (“background samples”) can be specified in the Settings window (though MAXENT will not put more than one point per grid cell), or any number of background points can be provided in SWD mode. The sufficiency of the number of points can be evaluated using the “gain”. The standard form of MAXENT’s log-likelihood involves a sum, analogous to the integral of equation 4 of the main text. A constant offset to the log-likelihood turns the sum into an average, which is independent of number of quadrature points; this value is reported as the “Regularised training gain” in MAXENT’s output (in the html file, the maxentResults.csv file and the maxent.log file). This gain could be compared between different numbers of quadrature points – changes of more than some small amount could be taken to indicate that the model is changing with different numbers of points. Alternatively, changes in the response curves, variable importance, or predictive performance across different numbers of quadrature points could be assessed.

The user also has the ability to adjust a number of settings, including the types of terms to include (linear, quadratic, product, hinge, and threshold), the convergence threshold for the model fit, the degree of regularisation via the LASSO penalty and other features to fine-tune the model. In order to use the “explain tool”, only linear, quadratic, and hinge terms may be included. However, there is no capacity for the model complexity (feature types and LASSO penalty) to be chosen in a data-driven manner as with `glmnet` or `ppmlasso`. Standard errors, information on residuals, capacity to explore clustering and other features of the previously described methods are also missing.

Further features of MAXENT not explored here include the ability to resample and subsample presence locations for bootstrapping and cross-validation for evaluation purposes. Whilst these cannot be implemented in spatial blocks (Wenger & Olden, 2012) from the interface, MAXENT can be run from R using the package `dismo`, so users could code their own specialised evaluation procedures. Comprehensive guides to MAXENT software are available with the program and elsewhere (Elith *et al.*, 2011; Merow *et al.*, 2013).

6.1 `dismo` package

Below we present some code for the `dismo` package for fitting a PPM.

6.1.1 Setting up to run Maxent from `R`

To run Maxent from R you need to install the package dismo and download maxent from Princeton University. Look at the help files for maxent within dismo, and here are some additional hints:

```
>library(dismo)

# find whether the maxent.jar file is in the correct place:
>jar <- paste(system.file(package="dismo"), "/java/maxent.jar", sep=’’)
>file.exists(jar)
```

If dismo does not think MaxEnt exists, manually put it (i.e. maxent.jar) in the right place by finding where R and dismo are installed. On many pcs it will be something like this:

```
C:\Program Files\R\R-3.0.1\library\dismo
```

On macs the R library is in the /Library/Frameworks directory - this is in the highest level Library folder - eg on an OSX MacBook pro it was here:

```
/Library/Frameworks/R.framework/Versions/3.1/Resources/library/dismo/java
```

For either PC or mac, within the dismo directory find the java directory and copy maxent.jar there. Then rerun this:

```
>file.exists(jar)  #the result should be TRUE
```

6.1.2 Preparing data

First set the working directory in whatever way you do that - we'll specify a name for the working directory:

```
>wd <- getwd()
```

Assuming you have gridded environmental data in a folder called "grids" in some format that can be read by the package raster. Also assuming that it is "masked" so any regions not being considered have "nodata" values.

Specify the names of the rasters – *e.g.* using variable names already introduced:

```
>vars <- c("FC", "MNT", "MXT", "Rain", "D.Main.sqrt", "D.Urb.sqrt", "soil")
```

Read these into a "stack", which is a collection of rasters with exactly the same extent, cell size etc:

```
>nsw.stack <- stack(file.path(wd, "grids/", vars))
```

To visualise the rasters:

```
>plot(nsw.stack)
```

Read in location records for the species in one of the format required for for maxent function - here we assume a .csv file with 2 columns of data specifying x (longitude) and y (latitude) for the sites (one row per site):

```
>locs <- read.csv(file.path(wd, "locs.csv"))
```

6.1.3 Fitting models and predicting

In dismo, Maxent models are fitted and predicted in 2 steps, as is common for many other modelling methods such as GLMs. The file format for inputs is described in dismo's help file for the function maxent:

```
?maxent
```

The arguments that the maxent function understands are identical to those in the set of flags that Maxent (the java program) understands – if you're not familiar with them, look at the help file for Maxent from the Maxent interface.

Since maxent-in-dismo implements fitting and predicting in 2 steps, only those arguments relevant to fitting will operate in the first step, and similarly for prediction.

Below is some example code with settings correct for point process models (note: data are not supplied; the code is indicative).

Fit a first maxent model:

```
>max1 <- maxent(x=nsw.stack, p=locs, path=paste(wd,’/max1’,sep=""),
 args=c("-P", "noautofeature", "nothreshold", "noproduct", -t soil,
 "maximumbackground=40000","noaddsamplestobackground","noremoveduplicates"))
```

The "path" as used above creates an output directory and will put the model outputs in there. If you don't specify the path, the maxent results will only be temporary files and you will lose them at the end of the session "args" above are, respectively (within the brackets): plot response curves, turn off autofeatures, turn off threshold features, turn off product features, toggle soil to categorical, take 40000 random background points, don't add samples to background, don't remove duplicate records from gridcells. Note the last 2 arguments are essential for the equivalence to a point process model. Any number of background points can be specified (though no more than one per grid cell will be sampled).

To see the results in a browser:

```
>max1
```

Plot showing importance of each variable:

```
>plot(max1)
```

To fit the same model but using SWD (SamplesWithData) format: first prepare some data:

```
>pres.env <- extract(nsw.stack, locs) #get the environmental data at
#presence points
>bg <- randomPoints(nsw.stack, 40000) #get 40,000 locations (or however
#many you want) in region, randomly selected from any cells with data
>bg.env <- extract(nsw.stack, bg)
>pbg.env <- rbind(pres.env, bg.env) # rowbind presence and bg
#environments
>pbg.which <- c(rep(1, nrow(locs)), rep(0, nrow(bg))) # a vector of
#1's and 0's to keep track of which rows are for presence (1) and
#which for bg (0)
```

Now for the SWD model:

```
>max1SWD <- maxent(x=pbg.env, p=pbg.which, path=paste(wd,'/max1SWD',sep=""),
 args=c("-P", "noautofeature", "nothreshold", "noproduct", -t soil,
 "maximumbackground=40000","noaddsamplestobackground"))
```

Note that this time you don't have to specify "`noremoveduplicates`" since the SWD format is ignorant of grid cells.

To predict: – *e.g.* here, to rasters:

If you are using observer bias covariates you'll need to make new rasters for the distance-to.. variables, and set these to a constant value. See help for the raster package for how to set all cells to one value. Assume we have a new stack with variables with the same name, but where the distance to.. variables are replaced with the constant-valued ones (call it `predict.stack`).

```
>max1.pred <- predict(max1, predict.stack, args="outputformat=raw",
 filename=paste(wd, '/max1/pred.asc', sep=""), format="ascii",
 progress="text")
```

Predictions will be in Maxent's raw format (correct for point process models, and different to the default logistic output), and here we are writing an .ascii file (other formats are possible).

7 INLA

7.1 Constructing the mesh and defining the SPDE

A key initial step in the R-INLA approach is to divide the study region into a series of triangles whose vertices form a "mesh" of points used in fitting. These triangular segments perform a similar role as grid cells, although there is no requirement that they be arranged in a regular rectangular grid.

We found that results were better when we defined a boundary for the study region before constructing the mesh. This can be done in R-INLA by creating a concave hull around the species locations and quadrature points (collected in a matrix data.xy) to serve as the boundary via the inla.nonconvex.hull function. The convex argument controls how close the hull is to its interior points and the resolution argument determines the number of points comprising the hull.

```
>hull = inla.nonconvex.hull(as.matrix(data.xy), convex = -0.01,
 resolution = 800)
```

We may now construct the mesh using hull as a boundary, with the cutoff, offset and max.edge arguments controlling the minimum distance between unique point locations, the extent of an inner and outer extension, and the maximum allowed edge length of triangles in the inner and outer extensions, respectively. The goal is to construct a set of triangles without small interior angles and long edges (Lindgren *et al.*, 2011).

For non-rectangular study regions, we make use of approximate solutions to stochastic partial differential equations (SPDEs, Lindgren *et al.*, 2011) for faster model-fitting. We use the constructed mesh to define the SPDE for the Gaussian field, setting initial prior parameters for the Gaussian field according to guidelines provided by R-INLA.

```
>mesh = inla.mesh.2d(boundary = hull, cutoff = 3, offset = c(20, 40),
 max.edge = c(6, 9))
>sigma0 = 0.05 ## field std.dev.
>range0 = 20
>kappa0 = sqrt(8)/range0
>tau0 = 1/(sqrt(4*pi)*kappa0*sigma0)
>spde = inla.spde2.matern(mesh, B.tau = cbind(log(tau0), -1, 1),
 B.kappa = cbind(log(kappa0), 0, -1), theta.prior.mean = c(0, 0),
 theta.prior.prec = c(0.1, 1))
```

The resulting mesh is illustrated in Figure 4, with the concave hull represented by the blue polygon and locations of *Eucalyptus sparsifolia* represented by red dots. We found that output was reasonably robust to prior specification, with fitted models using combinations of four values each for `sigma0` (0.01, 0.02, 0.05, 0.1) and `range0` (2, 10, 20, 40) producing largely similar results.

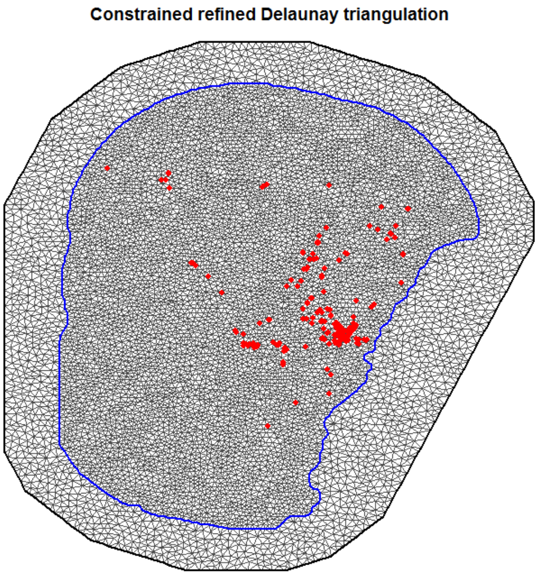

Figure 4: The mesh created in `R-INLA` for fitting the model with latent Gaussian field. The blue polygon represents the concave hull around the study region, and the red dots represent locations of *Eucalyptus sparsifolia*.

7.2 Projection matrices

Having constructed a mesh, we then create matrices that project the latent field from the mesh points to the species locations (`sp.xy`) and quadrature points (`quad.xy`):

```
>sp.mat = inla.spde.make.A(mesh, sp.xy)
>quad.mat = inla.spde.make.A(mesh, quad.xy)
```

7.3 Defining response

After defining the projection matrices sp.mat and quad.mat between the mesh and the species locations and quadrature points, respectively, we specify the covariates to be considered in the model by constructing data “stacks”. Note that the y covariate indicates presence or absence, the e covariate is the expected number of presences, with 0 at species locations and the square of the spatial resolution at quadrature points, and the dist covariate is the nearest distance to a presence point (here defined by pt.d for species presences and quad.d for quadrature points).

```
>attach(all.dat)
>stk.sp = inla.stack(data = list(y = 1, e = 0), A = list(sp.mat, 1),
 tag = ’sp’, effects = list(list(i = 1:mesh$n), data.frame(Intercept = 1,
 FC = FC[is.sp], MNT = MNT[is.sp], MXT = MXT[is.sp], Rain = Rain[is.sp],
 FC2 = FC[is.sp]^2, MNT2 = MNT[is.sp]^2, MXT2 = MXT[is.sp]^2,
 Rain2 = Rain[is.sp]^2, FC.MNT = FC[is.sp]*MNT[is.sp],
 FC.MXT = FC[is.sp]*MXT[is.sp], FC.Rain = FC[is.sp]*Rain[is.sp],
 MNT.MXT = MNT[is.sp]*MXT[is.sp], MNT.Rain = MNT[is.sp]*Rain[is.sp],
 MXT.Rain = MXT[is.sp]*Rain[is.sp], D.Main = sqrt(D.Main[is.sp]),
 D.Urb = sqrt(D.Urb[is.sp]), D.Main2 = sqrt(D.Main[is.sp])^2,
 D.Urb2 = sqrt(D.Urb[is.sp])^2,
 D.Main.D.Urb = sqrt(D.Main[is.sp])*sqrt(D.Urb[is.sp]),
 soil = soil[is.sp], dist = pt.d)))
>stk.quad = inla.stack(data = list(y = 0, e = spat.res^2),
 A = list(quad.mat, 1), tag = ’quad’, effects = list(list(i = 1:mesh$n),
 data.frame(Intercept = 1, FC = FC[is.quad], MNT = MNT[is.quad],
 MXT = MXT[is.quad], Rain = Rain[is.quad], FC2 = FC[is.quad]^2,
 MNT2 = MNT[is.quad]^2, MXT2 = MXT[is.quad]^2, Rain2 = Rain[is.quad]^2,
 FC.MNT = FC[is.quad]*MNT[is.quad], FC.MXT = FC[is.quad]*MXT[is.quad],
 FC.Rain = FC[is.quad]*Rain[is.quad], MNT.MXT = MNT[is.quad]*MXT[is.quad],
 MNT.Rain = MNT[is.quad]*Rain[is.quad], MXT.Rain = MXT[is.quad]*Rain[is.quad],
 D.Main = sqrt(D.Main[is.quad]), D.Urb = sqrt(D.Urb[is.quad]),
 D.Main2 = sqrt(D.Main[is.quad])^2, D.Urb2 = sqrt(D.Urb[is.quad])^2,
 D.Main.D.Urb = sqrt(D.Main[is.quad])*sqrt(D.Urb[is.quad]),
 soil = soil[is.quad], dist = quad.d)))
>stk.all = inla.stack(stk.sp.0, stk.q.0)
```

7.4 Fitting models

Having prepared the data for fitting, we fit a model containing environmental and distance-based terms, an interaction covariate (“constructed covariate”, Illian *et al.*, 2013) to capture point clustering or repulsion, and a random Gaussian field:

```
>ft.inla = inla(y ~ 0 + Intercept + FC + MNT + MXT + Rain + FC2 + MNT2 + MXT2
 + Rain2 + FC.MNT + FC.MXT + FC.Rain + MNT.MXT + MNT.Rain + MXT.Rain + D.Main
 + D.Urb + D.Main2 + D.Urb2 + D.Main.D.Urb + soil
 + f(inla.group(dist, n = 50, method = "quantile"), model = "rw1",
 scale.model = TRUE) + f(i, model = spde), family = 'poisson',
 data = inla.stack.data(stk.all),
 control.predictor = list(A = inla.stack.A(stk.all), compute = TRUE),
 E = inla.stack.data(stk.all)$e, control.compute = list(dic = TRUE),
 control.fixed = list(expand.factor.strategy = 'inla'))
```

The fitted coefficients of `ft.inla` are presented in Table 3.

We can project the mean and standard deviation of the latent field onto quadrature points using the `inla.mesh.projector` command as in Figure 4 of the main text:

```
>proj.quad = inla.mesh.projector(mesh, quad.xy)
>gf.mean = inla.mesh.project(proj.quad, ft.inla$summary.random$i$mean)
>gf.sd = inla.mesh.project(proj.quad, ft.inla$summary.random$i$sd)
```

We can produce a plot of the effect of the interaction covariate `dist` as follows:

```
>plot(ft.inla$summary.random[[2]][, 1:2], type = 'l',
 xlab = 'dist (km)', ylab = 'f(dist)'); abline(h = 0, lty = 3)
>for (i in c(4, 6)) lines(ft.inla$summary.random[[2]][,c(1, i)], lty = 2)
```

8 lgcp

First we need to define a boundary for the study region and build a point pattern object which includes the region boundary and presence points

```
>polygon <- data.frame(x=polyX,y=polyY)
>sd <- ppp(sp.dat$X,sp.dat$Y,poly=polygon)
```

where polyX and polyY are the co-ordinates of the boundary of the study region, and sp.dat are the data. Next we choose the cell width. To do this we can use the minimum.contrast function to find least squares estimates of model parameters. This gives us an idea of the extent of spatial variation via the 'scale' estimate. The chooseCellwidth function allows us to choose an appropriate cell width to optimize computational resource use. See (Taylor *et al.*, 2013) for further detail.

```
>minimum.contrast(sd, model = "exponential", method = "g",intens = density(sd),
 transform = log)
>chooseCellwidth(sd, cwinit=X)
>Cellwidth <- X
```

We will assume we have covariate information in a data frame quad which includes all relevant covariates as well as X and Y values in a grid inside the boundary of the study area. We then create the computational grid on which inference takes place. Note that the getpolyol is quite slow, and it is best to save the resulting object so that it can be used in multiple runs.

```
>covar=SpatialPixelsDataFrame(cbind(quad$X,quad$Y),quad)
>polyolay <- getpolyol(data = sd, pixelcovariates = covar,
 cellwidth = Cellwidth)
```

We also need to check the interpolation of covariates is accurate. Numeric variables are assigned interpolation by areal weighted mean while factor, character and other types of variable are assigned interpolation by majority vote. If any variables are incorrectly interpolated we can change the using the assigninterp function

```
>covar@data=guessinterp(covar@data)
>covar@data <- assigninterp(df = covar@data, vars "variable.name",
 value = "ArealWeightedSum")
```

We then need to create the Z matrix of covariates projected onto the computational grid. We do this using the getZmat function.

```
>Zmat <- getZmat(formula = X ˜ ...,data = sd, pixelcovariates = covar,
 cellwidth = Cellwidth, overl = polyolay)
```

We can choose from a number of flexible spatial covariance functions, see Appendix C of (Taylor *et al.*, 2013) for detail, for simplicity we choose the exponential.

```
>cf <- CovFunction(exponentialCovFct)
```

The last thing we need to specify are the priors for all out parameters. To specify priors we can use the `lgcpPrior`

```
>priors <-lgcpPrior( etaprior =
 PriorSpec(LogGaussianPrior(mean = log(c(5, 5)),variance = diag(c(5, 0.5)))),
 betaprior = PriorSpec(GaussianPrior(mean = rep(0, nbeta),
 variance = diag(rep(10^3, nbeta)))))
```

where `nbeta` is the number of predictors plus 1. The running of the MCMC code can be quite sensitive to prior choice. In particular if there are warnings during the running of the code, the prior for η may have to be tightened, refer to (Taylor *et al.*, 2013) for guidance. Now all that remains is to choose the length of the MCMC chain, the desired burn in and the number of iterations to thin by and run the chain. The length of time required by the MCMC depends on the length of the chain. We suggest you first choose small values to test the code is working.

```
>lbr=c(1000000,100000,1000) #length,burn-in and thin
>ft.lgcp=lgcpPredictSpatialPlusPars(formula = X ~ ...,sd = sd,Zmat = Zmat,
 model.priors = Priors, model.inits = NULL ,spatial.covmodel = cf,
 cellwidth = Cellwidth,poisson.offset = NULL,
 mcmc.control = mcmcpars(mala.length = lbr[1], burnin = lbr[2],retain = lbr[3],
 adaptivescheme = andrieuthomsh(inith = 1,
 alpha = 0.5, C = 1, targetacceptance = 0.574)))
```

The fitted coefficients of `ft.lgcp` are presented in Table 3.

9 Fitted Coefficients

A list of fitted coefficients appears in Table 3 for models fitted in `spatstat`, `ppmlasso`, `R-INLA`, and `lgcp`. The model fitted by `spatstat` did not account for spatial dependence, unlike the models fitted by `ppmlasso`, `R-INLA` and `lgcp`. We suspect that the latent

Gaussian fields of the Cox process models fitted by `R-INLA` and `lgcp` were correlated with the environmental predictors, and hence those models identified fewer significant covariates.

References

Baddeley, A. & Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, **12**, 1–42.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York.

Daley, D. J. & Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Springer, New York.

Diggle, P. (2003). *Statistical Analysis of Spatial Point Patterns*. Oxford University Press, New York.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E. & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Fithian, W. & Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, **7**, 1917–1939.

Fu, W. J. (2005). Nonlinear GCV and quasi-GCV for shrinkage models. *Journal of Statistical Planning and Inference*, **131**, 333–347.

Hastie, H., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Illian, J. B., Martino, S., Sørbye, S. H., Gallego-Fernández, J. B., Zunzunegui, M., Esquivias, M. P. & Travis, J. M. (2013). Fitting complex ecological point process models with integrated nested Laplace approximation. *Methods in Ecology and Evolution*, **4**, 305–315.

Lindgren, F., Rue, H. & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498.

Merow, C., Smith, M. J. & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Phillips, S. J., Anderson, R. P. & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Reineking, B. & Schröder, B. (2006). Constrain to perform: regularization of habitat models. *Ecological Modelling*, **193**, 675–690.

Renner, I. W. & Warton, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**, 274–281.

Taylor, B. M., Davies, T. M., Rowlingson, B. S. & Diggle, P. J. (2013). lgcp: An R Package for Inference with Spatial and Spatio-Temporal Log-Gaussian Cox Processes. *Journal of Statistical Software*, **52**.

Wenger, S. J. & Olden, J. D. (2012). Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, **3**, 260–267.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.

Table 3: Estimated coefficients for models fitted in `spatstat` (`ft.int`), the area-interaction model of radius 2km with LASSO penalty chosen via block cross validation in `ppmlasso` (`final.fit`), `R-INLA` (`ft.inla`) and `lgcp` (`ft.lgcp`). Coefficients in bold are significantly different from zero (available for `spatstat`, `R-INLA` and `lgcp`. Note that `spatstat` uses soil type 11 as a reference covariate, and hence it has no estimated coefficient.

Covariate	`spatstat`	`ppmlasso`	`R-INLA`	`lgcp`
Intercept	**-715.838**	-8.024	-18.851	-11.58
soil type 2	-15.663	-0.070	-21.263	-12.31
soil type 5	0.120	0.074	**34.238**	**17.31**
soil type 6	-13.550	-0.420	-19.116	-11.85
soil type 7	**1.888**	0.288	**34.795**	**18.38**
soil type 8	0.029	-0.038	**34.086**	15.66
soil type 11	NA	0	**34.314**	**17.90**
soil type 12	-0.690	0.073	**34.474**	-13.08
soil type 15	-14.178	-0.365	-17.369	-12.21
$\sqrt{\texttt{D.Main}}$	**-0.598**	0	-0.417	**-0.362**
$\sqrt{\texttt{D.Urb}}$	**-1.148**	-0.604	**-1.053**	0.017
`D.Main`	-0.137	-0.964	-0.131	-0.014
`D.Urb`	**0.093**	0.728	**0.101**	-0.003
$\sqrt{\texttt{D.Main*D.Urb}}$	0.133	0	0.130	**0.012**
`FC`	-3.391	0	0.010	-7.809
`MNT`	**-29.192**	0	-0.775	-1.845
`MXT`	**43.544**	0	-0.884	-0.064
`Rain`	**326.375**	0	-15.158	-7.581
$\texttt{FC}^2$	**-0.122**	-0.295	-0.000	0.042
$\texttt{MNT}^2$	**-0.400**	-0.217	-0.159	-0.244
$\texttt{MXT}^2$	**-0.676**	-0.228	0.000	-0.021
$\texttt{Rain}^2$	**-41.458**	0	2.703	-2.153
`FC*MNT`	**0.188**	-0.027	0.040	-0.114
`FC*MXT`	0.112	0.439	0.018	0.250
`FC*Rain`	0.146	-0.080	-0.607	1.180
`MNT*MXT`	**0.947**	0.848	0.064	0.161
`MNT*Rain`	**5.725**	-0.509	-0.280	-0.630
`MXT*Rain`	**-9.511**	0	0.439	0.517