Advances in presence-only methods in ecology

Author:
Renner, Ian Walton

Publication Date:
2013

DOI:
https://doi.org/10.26190/unsworks/16297

Advances in Presence-Only Methods in Ecology

Ian Walton Renner
B. Sc. M. Stat

School of Mathematics and Statistics

Faculty of Science

The University of New South Wales, Australia

Submitted for the degree of Doctor of Philosophy, September 6, 2013

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: **Renner**

First name: **Ian** Other name/s: **Walton**

Abbreviation for degree as given in the University calendar: **PhD**

School: **School of Mathematics and Statistics** Faculty: **Faculty of Science**

Title: Advances in presence-only methods in ecology

Abstract 350 words maximum: (PLEASE TYPE)

Species distribution models are useful tools for relating the locations of species in a given region to environmental factors. This thesis will focus on the modelling of presence-only data, in which information is available about where species are reported present but not where species are reported absent. The aims of this thesis are to use theoretical tools from statistics to improve modern presence-only methods of analysis.

This thesis establishes that MAXENT, a popular method in ecology based on maximum entropy, is equivalent to Poisson point process modelling, a widely-used statistical method for analysing spatial point patterns only recently applied to species distribution modelling. This equivalence result significantly unifies the presence-only analysis literature and has important ramifications for MAXENT and point process models. Despite its good predictive performance, MAXENT has shortcomings in interpretation and implementation that can now be overcome. In particular, MAXENT users can inherit from point process models some well-developed tools for addressing model adequacy and the ability to model point interactions.

MAXENT's use of a LASSO penalty is known to improve predictive performance. However, the default penalty chosen by MAXENT software is *ad hoc*. Another focus of this thesis is implementing LASSO for point process models, which has rarely been done previously.

This thesis provides an asymptotic result for applying a LASSO penalty to point process models such that consistent estimates of model parameters and predictions can be achieved. A new consistent criterion for choosing the LASSO penalty (“MSI") is consequently developed as an alternative to the default MAXENT penalty which has better properties. MSI is found to be competitive with traditional methods of choosing the LASSO penalty and generally superior to the MAXENT penalty in a broad comparison using real and simulated species data.

This extension of point process models regularised with a LASSO penalty (“PPM-LASSO") therefore represents a significant advance of current species distribution modelling methods by combining the statistical foundations of point process models and the strong predictive performance of MAXENT via LASSO penalisation. I have developed the freely-available `ppmlasso` package for `R` so that PPM-LASSO models may now be fitted by users.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

………………………………………………………………	………………………………………..………………	……………………………..………
Signature	Witness	Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

Table of Contents

Originality Statement

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project’s design and conception or in style, presentation and linguistic expression is acknowledged.’

Ian Renner
September 6, 2013

Copyright Statement

‘I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstract International (this is applicable to doctoral theses only). I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.’

Ian Renner
September 6, 2013

Authenticity Statement

‘I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.’

Ian Renner
September 6, 2013

Abstract

Species distribution models are useful tools for relating the locations of species in a given region to environmental factors. This thesis will focus on the modelling of presence-only data, in which information is available about where species are reported present but not where species are reported absent. The aims of this thesis are to use theoretical tools from statistics to improve modern presence-only methods of analysis.

This thesis establishes that MAXENT, a popular method in ecology based on maximum entropy, is equivalent to Poisson point process modelling, a widely-used statistical method for analysing spatial point patterns only recently applied to species distribution modelling. This equivalence result significantly unifies the presence-only analysis literature and has important ramifications for MAXENT and point process models. Despite its good predictive performance, MAXENT has shortcomings in interpretation and implementation that can now be overcome. In particular, MAXENT users can inherit from point process models some well-developed tools for addressing model adequacy and the ability to model point interactions.

MAXENT's use of a LASSO penalty is known to improve predictive performance. However, the default penalty chosen by MAXENT software is *ad hoc*. Another focus of this thesis is implementing LASSO for point process models, which has rarely been done previously.

This thesis provides an asymptotic result for applying a LASSO penalty to point process models such that consistent estimates of model parameters and predictions can be achieved. A new consistent criterion for choosing the LASSO penalty ("MSI") is consequently developed as an alternative to the default MAXENT penalty which has better properties. MSI is found to be competitive with traditional methods of choosing the LASSO penalty and generally superior to the MAXENT penalty in a broad comparison using real and simulated species data.

This extension of point process models regularised with a LASSO penalty ("PPM-LASSO") therefore represents a significant advance of current species distribution modelling methods by combining the statistical foundations of point process models and the strong predictive performance of MAXENT via LASSO penalisation. I have developed the freely-available `ppmlasso` package for `R` so that PPM-LASSO models may now be fitted by users.

Acknowledgements

First and foremost I'd like to thank my supervisor David Warton for his tireless and peerless support. Not only has he expertly guided my research, but he has also taught me how to effectively communicate it through both writing and presentation. In doing so he has dedicated an enormous amount of time (perhaps too much!) in reading various drafts of presentation slides, journal articles, and this thesis. He has provided great encouragement when needed and has generally augmented my love of statistics.

This thesis has also been greatly improved through thoughtful suggestions by the UNSW Eco-Stats Research Group – Stephen Wright, Alexandra Brown, Yi Wang, Bénédicte Madon, Eve Slavich, Francis Hui, Jakub Stokłosa, and Sara Taskinen. Their contributions have helped me to present Chapters 5, 6, and 7 as well as various presentations more clearly and concisely.

I'd also like to thank Dan Ramp and Evan Webster for providing initial access to the species and environmental data, as well as for ironing out some difficulties along the way.

Finally, I would have not been able to complete this work without the unswerving support and encouragement of my family. This journey has been a lifetime in the making, and I would like to thank my parents and brother for their love and for encouraging me to chase my dreams. I'd like to particularly thank my beautiful star of a wife, Tracy. Thank you for always believing in me and boosting my spirits throughout this labour of love. You remind me every day why I married you.

I'd like to dedicate this thesis to my father, who filled me with a love of mathematics and statistics from a young age and always encouraged me to pursue my curiosity. I miss you very much and can only hope to be half as outstanding of a father as you as I embark on parenthood.

Part I —

Literature Review

Chapter 1

Introduction

Species distribution modelling (SDM), where the goal is to relate the distribution of a species' habitat to the environment (Figure 1.1), is a high-impact topic in ecology. ISI's Essential Science Indicators for July 2012 identifies SDM as one of the top five ranked research fronts in ecology and the environmental sciences. Indeed, leading articles published in 2006 (Phillips *et al.*, 2006; Elith *et al.*, 2006) have been cited over 2,000 times.

One reason for such high interest is that SDM aims to answer important biological and environmental questions. Species distribution models (SDMs) are useful in explaining what environmental factors influence the distribution of a particular species and hence can inform conservation efforts and studies of impacts of activities on habitats (Franklin, 2009). SDMs also facilitate prediction of species distributions or habitat suitability over an entire region (Elith & Leathwick, 2009), which can be useful for discovering unsampled areas that may be favourable for a species or predicting the advance of invasive species. Finally, SDM can be used for projection as it aims to address topical questions such as the potential effects of climate change on species distributions (Thullier *et al.*, 2008), although using SDM for this purpose has its limitations (Franklin, 2009).

Rapid progress in this field has been facilitated by recent significant technologi-

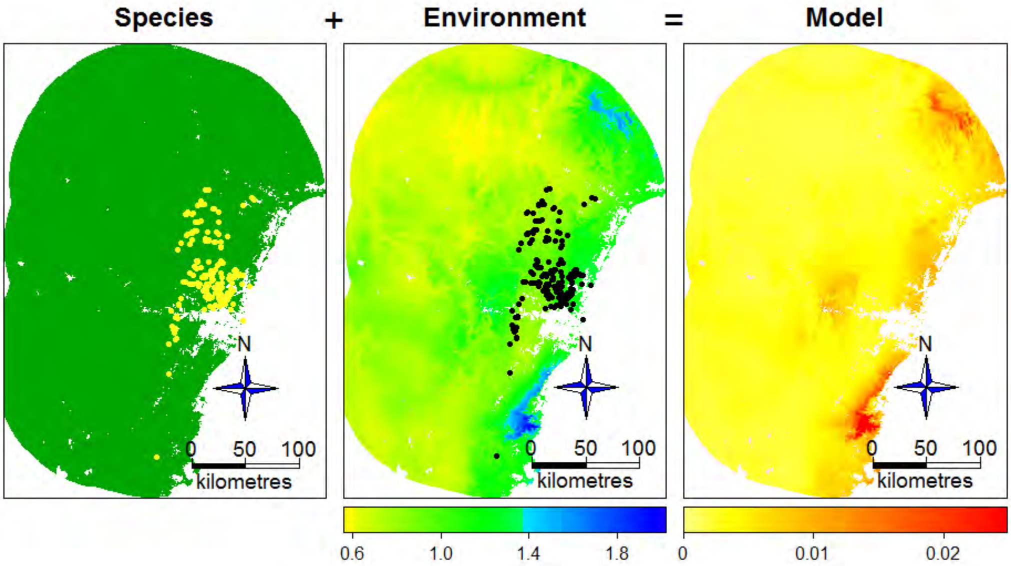

Figure 1.1: Species distribution modelling concept. SDMs relate the presence of a species (*Corymbia eximia*, left) to the environment (annual rainfall, middle) to predict the distribution of the species (right).

cal advances in remote sensing, GIS (O'Sullivan & Unwin, 2010), statistical software and computational power, enabling models to be built at increasingly fine resolutions and increasingly large spatial scales. Coinciding with this technological explosion has been the increasing availability of species data, in the form of digital records of species locations. The species data used for this thesis are 85,877 records of plant species locations in the Sydney-Newcastle region of New South Wales, Australia. These actually form a subset of over 1,400,000 records from the Office of Environment and Heritage (NSW Office of Environment and Heritage, 2012) across the same region. Such massive data sets are likewise available elsewhere in the world (Kadmon *et al.*, 2004; Elith *et al.*, 2006; Franklin, 2009), so it is now relatively easy to find data for a wide range of species and regions.

Ideally, a SDM could be constructed using systematically collected presence-absence data so that logistic regression (McCullagh & Nelder, 1989) and its extensions (Hastie & Tibshirani, 1990; Schapire, 2003) may be used. But often, the best available data are a list of locations where a species has been observed, with no 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. SDM methods for such presence-only data will be the focus of this thesis.

An example used throughout the thesis is the location of 302 observed presences of the eucalypt *Corymbia eximia* in the Blue Mountains region near Sydney (Figure 1.1). In its most basic form, a SDM relates the presence locations (left panel) to the environment (*e.g.* annual rainfall, middle panel) in order to make a map of the predicted species distribution throughout a study region (right panel). Although this appears to be a simple goal, there are a number of questions about how the SDM should be constructed. Which SDM method should be used, and what are the ramifications of this choice? Which variables should be included, and how should they be chosen?

The question of which SDM method should be chosen is not straightforward. The increased access to both environmental and species data as well as the widespread utility of SDM has led to the development of a large number of presence-only SDM methods, as reviewed in Section 2.1. These methods differ in response, the type of variables included, incorporation of points representing absences or the background of the study region, and methods of fitting. As discussed in Section 2.2, it is unclear for most methods how the predictions should be interpreted, how points representing absences or the background should be chosen, and how to address data challenges such as spatial autocorrelation and observer bias inherent in presence-only data. A particular focus of the thesis is point process models (PPMs, Warton & Shepherd, 2010), a method of analysing point pattern data capable of addressing these issues, examined in detail in Chapter 3.

The question of which variables should be included in an SDM (“variable selection”) is likewise complex. A number of strategies have been developed (Chapter 4) including implementation of a LASSO penalty (Tibshirani, 1996), which is a key focus of the thesis. Based on the choice of SDM method and how to perform variable selection, the model for a species can vary significantly in its complexity, form, and interpretation of its predicted distribution.

There is a clear need for a synthesis of SDM methods informed by their capacity to address the challenges described above. Some recent papers (Elith & Leathwick, 2009; Aarts *et al.*, 2012) have called for greater unification and synthesis of the SDM literature. To that end, Warton & Shepherd (2010) linked Poisson point process models with pseudo-absence regression, one of the most popular methods in practice (also see Baddeley *et al.* (2010)). Aarts *et al.* (2012) linked different classes of SDM through the likelihood of an inhomogeneous Poisson point process. This thesis provides further synthesis of the SDM literature.

The aim of this thesis is to improve modern presence-only methods of analysis by using theoretical tools from statistics. The principal outcome of this aim is an equivalence result between Poisson point process models and MAXENT, a popular presence-only SDM method. Leveraging off of this equivalence allows both Poisson point process models and MAXENT to be improved in practice. A particular extension is PPM-LASSO, an approach which takes cues from MAXENT on how to effectively apply a LASSO penalty, and extends them to a point process model framework. This extension addresses all of the difficulties that currently face presence-only methods and thereby yields advantages in interpretation, predictive performance and model fitting.

To that end, I prove the equivalence of Poisson point process models with MAXENT in Chapter 5. This work has been accepted for publication in *Biometrics* (Renner & Warton, 2013). In Chapter 6 I establish a novel and important asymptotic result for Poisson point process models that informs the choice of the method for determining the LASSO penalty. I then develop a novel criterion using this result. In Chapter 7, I compare the performance of both existing methods and this new criterion. Chapter 8 details software I developed for fitting point process models with a LASSO penalty that will be included in the forthcoming `ppmlasso` package in `R` (R Development Core Team, 2010).

Chapter 2

Current Presence-Only Species Distribution Modelling Methods

As mentioned in Chapter 1, there has been a recent proliferation of SDM methods. In this Chapter I will review the most common presence-only SDM methods in use today (Section 2.1) and discuss some of the current challenges in their interpretation and implementation (Section 2.2).

2.1 Review of Presence-Only SDM Methods

This thesis focusses on presence-only SDM methods. Let $\mathbf{y}_P = \{y_1, \dots, y_m\}$ be the vector of m presence-only locations for a particular species over some region $\mathcal{A}$ and $\mathbf{x}(y)' = \{1, x_1(y), \dots, x_p(y)\}'$ be the vector containing an intercept term and the values of p environmental variables corresponding to location $y \in \mathcal{A}$. The goal of SDM is to link the location of species presences $\mathbf{y}_P$ to the environment $\mathbf{x}(y)$, using one of a few methods.

2.1.1 Methods Which Only Use Presence-Only Locations

Some of the earliest methods used for modelling presence-only data were BIOCLIM (Busby, 1991), HABITAT (Walker & Cocks, 1991), and DOMAIN (Carpenter *et al.*, 1993). These "envelope" (Elith & Leathwick, 2007) or "profile" (Pearce & Boyce, 2006) methods exclusively use the environmental data $\mathbf{X}_P$ at presence-only locations $\mathbf{y}_P$ to determine the environmental regions (or "envelopes") amenable to the species, where $\mathbf{X}_P$ is an $m \times (p+1)$ matrix whose ith row is $\mathbf{x}(y_i)$ for $y_i \in \mathbf{y}_P$. BIOCLIM identifies the biotic range of presences for each environmental variable and proposes a bioclimate as a union of these ranges. To incorporate the effect of interactions among variables, HABITAT restricts the bioclimate to a convex hull of $\mathbf{X}_P$. Because both BIOCLIM and HABITAT exclude sites with environmental conditions very near but just outside the extremes measured at presence locations, Carpenter *et al.* (1993) proposed DOMAIN, which assigns a similarity score to sites $y \in \mathcal{A}$ based on the distance between their environmental conditions $\mathbf{x}(y)$ and those of known presence locations $\mathbf{X}_P$.

These methods do not output probabilities of species occurrence but rather degrees of classification within a climatic envelope. While some argue that they can be appropriate when data are scarce (Pearce & Boyce, 2006), profile methods have fallen in popularity due to weak predictive performance in comparison with newer methods (Elith *et al.*, 2006).

2.1.2 Methods Which Use Background Points

Most modern SDM methods contrast presence-only locations $\mathbf{y}_P$ with points chosen to represent the background of the study region ("pseudo-absences" or "background points") in order to model relative likelihood of species presence. Let $\mathbf{y}_0$ be a vector of $n - m$ background points and let $\mathbf{z} = \{z_1, \ldots, z_n\}$, where $z_i = I(i \in \{1, \ldots, m\})$ and $I(\cdot)$ is the indicator function. z_i therefore indicates whether location

i is a presence-location or a background point. Let $\mathbf{X} = (\mathbf{X}'_P, \mathbf{X}'_0)'$, where $\mathbf{X}_0$ is an $(n - m) \times (p + 1)$ matrix whose ith row is $\mathbf{x}(y_i)$ for $y_i \in \mathbf{y}_0$.

Ecological niche factor analysis (ENFA, Hirzel *et al.*, 2002) uses an approach similar to principal components analysis, first determining the "marginality factor" which transverses the centroid of environmental space for presence locations ($\overline{\mathbf{x}}_P$, say) and for background sites ($\overline{\mathbf{x}}_0$, say), and sequentially adding orthogonal "specialisation factors" to maximise the ratio of residual variance between presence sites and background sites (Engler *et al.*, 2004). Similar to the profile techniques above, ENFA does not output a probability but rather a habitat suitability index of sites based on their similarity to known presence sites. It is susceptible to optimistic prediction of species distributions (Pearce & Boyce, 2006).

Pseudo-absence regression approaches are fitted by regressing $\mathbf{z}$ against $\mathbf{X}$. Some common implementations of the pseudo-absence approach are:

- Logistic regression, a type of generalised linear model (GLM, McCullagh & Nelder, 1989) which uses a binomial response with logistic link.
- Generalised additive models (GAMs, Hastie & Tibshirani, 1990), which allow for non-linear effects of environmental variables through the implementation of nonparametric smoothers. The added flexibility of GAMs is both attractive and considered ecologically realistic (Elith & Leathwick, 2009), although they do not model interactions among environmental variables unless explicitly included in addition to smoothers.
- Multivariate adaptive regression splines (MARS, Elith & Leathwick, 2007), which fit piecewise linear splines to data using least squares and hence also can model complex relationships. MARS are faster to compute than GAMs (Elith & Leathwick, 2007; Franklin, 2009), and have been extended to modelling communities of species which has resulted in good predictive performance (Elith *et al.*, 2006).

Some other methods which use pseudo-absences have emerged from the machine learning and data-mining communities (Elith & Leathwick, 2009):

- Decision trees, which seek to divide the environmental space into distinct intervals over which species response is roughly homogeneous (Franklin, 2009).
- Boosted regression trees (BRTs, Elith *et al.*, 2008) and random forests (Breiman, 2001), which generate an "ensemble" of trees and average the results.
- Genetic algorithms for rule-set production (GARP, Stockwell, 1999), which stochastically generate presence and background locations that are contrasted to develop rules that evolve over the span of many iterations. Results are typically averaged over many such rule sets.

Despite the apparent difference in approach, many machine learning methods can be posed in terms of classical regression (Hastie *et al.*, 2009). For example, BRTs can be considered additive regression models with each term corresponding to a single tree (Elith *et al.*, 2008). Ensemble approaches such as BRTs, random forests, and GARP generally have advantages in predictive performance and stability over single implementations (Franklin, 2009).

All of these pseudo-absence regression approaches suffer from problems in model specification, interpretation and implementation (Warton & Shepherd, 2010) as a consequence of the reliance on pseudo-absences, detailed in Section 2.2.

A quite different approach to specifying a presence-only SDM is to use a Poisson PPM (Warton & Shepherd, 2010; Chakraborty *et al.*, 2011; Aarts *et al.*, 2012), which relates the number and location of presences $\mathbf{y}_P$ to $\mathbf{x}(y), y \in \mathcal{A}$. A key distinction of Poisson PPMs is that rather than merely modelling the probability of species occurrence, they directly model the intensity of species presence at sites, a formulation which leads to a number of attractive features. They address a number of key challenges in presence-only data analysis discussed in Section 2.2. While the model is posed in a different way to pseudo-absence approaches, it can be related to them through reexpression as a regression of $\mathbf{z}$ on $\mathbf{X}$ with weighted observations, as in Chapter 3. Poisson PPMs are a major focus of this thesis and are discussed in detail in Chapter 3.

A related and popular method is MAXENT (Phillips *et al.*, 2006), which is the second important method considered in this thesis (Chapter 5). Rather than using the presence-only locations $\mathbf{y}_P$ directly, MAXENT divides the study region into n grid cells and estimates the probability π_i that if there is a single presence, it is located in the ith grid cell. It calculates the probability by maximising entropy subject to $\sum_{i=1}^{n} \pi_i = 1$ and an additional constraint on the set of environmental variables. The idea to split the study region into square grid cells was first implemented by Agterberg (1974) in a logistic regression model of mineral deposits due to the suggestion of Tukey (1972), and eventually was applied in GIS (Bonham-Carter, 1994). Full details of the MAXENT procedure are provided in Chapter 5, where I establish the equivalence of MAXENT and Poisson regression, which enables a link to Poisson PPMs. MAXENT has been shown to have good predictive performance (Elith *et al.*, 2006), which may explain its popularity. However, it has a number of shortcomings in interpretation and implementation, discussed in Section 2.2 and Chapter 5.

2.2 Current Challenges for Presence-Only Analysis

There are a number of challenges in implementing and interpreting presence-only SDM methods.

2.2.1 Problems of Interpretation and Implementation

A key question for SDM methods that contrast presence locations with pseudo-absences or background points (Table 2.1) is how many of these generated absences should be chosen and where they should be. Some papers call for a fixed number (Chefaoui & Lobo, 2008; Hernandez *et al.*, 2008) or ratio (Hengl *et al.*, 2010; Lobo *et al.*, 2010) of pseudo-absences. Other have raised concern about the perceived dependence between pseudo-absences and known presence locations (Phillips *et al.*, 2009; Hengl *et al.*, 2010; Lobo *et al.*, 2010). The choice of pseudo-absences is also important because it has an impact on model predictions. The location of pseudo-absences can lead to categorically different predicted species distributions (Lobo *et al.*, 2010), and doubling the number of pseudo-absences will cause the scale of pseudo-probabilities to be roughly halved. Therefore, models that predict pseudo-probability can be said to be "scale-dependent".

Related to this question is the choice of spatial resolution used for analysis. For example, MAXENT splits up the study region into square grid cells, and hence the question becomes what the size of the grid cells should be. As $\sum_{i=1}^{n} \pi_i = 1$, the scale of these probabilities depends on the choice of grid cell size, and hence they are also scale-dependent. In addition to the difficulty in comparing models fitted with different numbers of pseudo-absences or grid cells, scale-dependent methods have other limitations – they can not be used to model species abundance as currently implemented, and as demonstrated in Chapter 5, they can not be used to determine the appropriate spatial resolution to be used for analysis.

Warton & Shepherd (2010) clarified the role of pseudo-absences for Poisson PPMs as quadrature points for approximating the integral in the likelihood function (Equation 3.7 in Chapter 3) and hence they should be chosen in such a way to provide a reasonable approximation, *e.g.* along a regular grid at increasingly fine spatial resolutions until the likelihood converges. This strategy permits the choice of the number and location of pseudo-absences as well as the spatial resolution used for analysis

Table 2.1: Properties and challenges of current SDM methods. Methods for which the scale of model predictions depends on choice of spatial resolution are scale dependent, and hence can not be used to model species abundance. Strictly speaking, many of these methods are algorithms rather than models, but it is commonplace in the SDM literature to refer to such methods as models.

SDM Method	**Pseudo-Absences**	**Model Prediction**	**Scale Dependence**
BIOCLIM	No	Envelope	N/A
HABITAT	No	Envelope	N/A
DOMAIN	No	Similarity Score	N/A
ENFA	Yes	Suitability Index	No
Pseudo-Absence Regression	Yes	Pseudo-Probability	Yes
GAM (Binomial)	Yes	Pseudo-Probability	Yes
MARS (Binomial)	Yes	Pseudo-Probability	Yes
Poisson PPM	Yes	Intensity	No
MAXENT	Yes	Probability	Yes
BRT (Binomial)	Yes	Pseudo-Probability	Yes
GARP	Yes	Pseudo-Probability	Yes

to be entirely determined by the data, which is not possible with scale-dependent methods.

Another notable challenge faced by current SDM methods is the interpretation of model predictions. Because there is no reliable absence data, most methods do not output probabilities of species presence or occupancy but rather relative likelihoods of habitat suitability (Pearce & Boyce, 2006; Elith & Leathwick, 2007; Franklin, 2009; Aarts *et al.*, 2012). Methods that do attempt to interpret output as probabilities still suffer a lack of clarity due to scale dependence. While a natural interpretation of probability would be the likelihood of species presence, both pseudo-probability and MAXENT's π_i are in essence also merely suitability indices.

2.2.2 Challenges Posed by the Data

Other challenges in presence-only analysis arise from the nature of the data. One such issue for presence-only data is observer bias. As presence-only data is by nature opportunistic, it reflects not only the distribution of the species in question but also the distribution of the observers (Phillips *et al.*, 2009). Subsequently, there is usually a greater concentration of observations in areas that are easier to access (*e.g.* areas that are in the vicinity of urban areas, roads and national parks). This bias can degrade the predictive ability of models that do not account for it, particularly when the environmental variables considered in the model differ between areas of easy access and remote areas (Kadmon *et al.*, 2004). Dorazio (2012) showed that coefficient estimates of a SDM are consistent as long as species detection probability is not correlated with the environmental variables used, although this can be difficult to establish.

Spatial autocorrelation, in which there is some dependence among the presence-only locations $\mathbf{y}_P$, can likewise have an impact on SDM, but it is often ignored (Franklin, 2009). Such dependence may arise in presence-only data because observations may come from the same observer sighting a species multiple times in a small area and knowledge about where a species is known to be found may influence the sampling effort (Chakraborty *et al.*, 2010). Spatial autocorrelation is an important issue because most SDM methods optimise a function of the joint density of presence-only locations and assume that this joint density can be expressed as the product of marginal densities at each point.

In the thesis, I address the potential for observer bias by introducing variables related to site accessibility, and I address spatial autocorrelation by considering models that account for point interactions, as in Chapter 3.

Chapter 3

Point Process Models

In Chapter 2, I reviewed a number of current SDM methods and outlined a number of obstacles in their interpretation and implementation. Poisson PPMs have numerous benefits that address these obstacles and therefore form a major focus of this thesis. In Sections 3.1 and 3.2 I describe Poisson PPMs and related models that account for interpoint interaction. In Section 3.3, I demonstrate diagnostic tools that may be used to check assumptions of PPMs.

3.1 Poisson Point Process Models

As the responses of interest in SDM are the number and location of species presences in some physical space, it is natural to view the distribution as a spatial point pattern (Cressie, 1993; Diggle, 2003). Models of spatial point patterns look to describe the entire spatial configuration of these points, such that it is possible to estimate μ, the intensity or limiting expected number of points per unit area throughout the region of interest $\mathcal{A}$:

$$\mu(y) = \lim_{|dy| \to 0} \left\{ \frac{E[M(dy)]}{|dy|} \right\}, \tag{3.1}$$

where $|dy|$ is the area of region dy, an arbitrarily small region in the neighbourhood of y which contains $M(dy)$ presences. These models can have a number of different forms given various assumptions about dependence among points and the environment itself (Cressie, 1993; Diggle, 2003; Baddeley & Turner, 2005; Chakraborty *et al.*, 2011). Some point processes do not have an intensity function, but such processes are not considered in this thesis.

The simplest representation of a spatial point pattern is a homogeneous Poisson point process, which assumes that (1) the number of presences m in the whole region $\mathcal{A}$ is the observed realisation of a Poisson random variable M and (2) that these presences $\mathbf{y}_P$ are uniformly and independently distributed over $\mathcal{A}$ (Cressie, 1993; Diggle, 2003). This means that the intensity function μ is constant throughout $\mathcal{A}$. Such processes are also said to be stationary and exhibit "complete spatial randomness".

Because the aim of SDM is to relate the location of species presences to the environment, it is natural to extend the framework of the homogeneous Poisson point process so that the intensity μ varies spatially according to the environment (and hence the process is non-stationary). Consequently, in an *inhomogeneous* Poisson point process, intensity is indexed by location as in (3.1) and presence locations $\mathbf{y}_P$ are assumed to be distributed independently conditional on the environment.

The conditions of both a homogeneous and an inhomogeneous Poisson point process can then be given as in Table 3.1. More specifically, the first condition for an inhomogeneous Poisson point process describes the probability structure for the number of points m:

$$P(M = m) = \frac{e^{-\mu_{\mathcal{A}}}(\mu_{\mathcal{A}})^m}{m!}, \ m = 0, 1, 2, \ldots, \tag{3.2}$$

Table 3.1: Assumptions of homogeneous and inhomogeneous Poisson point processes.

	Homogeneous	**Inhomogeneous**
Number of points M	Poisson($\mu\lvert\mathcal{A}\rvert$)	Poisson($\mu_{\mathcal{A}}$), where $\mu_{\mathcal{A}} = \int_{\mathcal{A}} \mu(y)dy$
Distribution of points	Independent Uniform on $\mathcal{A}$	Independent with density $\propto \mu(y)$

where $\mu_{\mathcal{A}}$ is the expected number of presence points in $\mathcal{A}$. The second condition for an inhomogeneous Poisson point process implies that given a point, the density of its location y is:

$$f(y) = \frac{\mu(y)}{\mu_{\mathcal{A}}}, y \in \mathcal{A}.$$

Hence, conditional on $M = m$, the joint density of the m points $y_1, \ldots, y_m$ is:

$$f(y_1, \ldots, y_m | M = m) = \frac{\prod_{i=1}^{m} \mu(y_i)}{(\mu_{\mathcal{A}})^m}. \tag{3.3}$$

Estimating the intensity $\mu(y)$ in a statistical model allows an analyst to describe the relationship with the environment. In a Poisson PPM, intensity is often modelled as a log-linear function of environmental covariates:

$$\ln \mu(y_i) = \mathbf{x}_i' \boldsymbol{\beta}, \tag{3.4}$$

where $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}$ is a vector that contains the parameters corresponding to the p environmental covariates $\mathbf{x}_i$. An advantage of Poisson PPMs is that because the intensity μ is modelled on a per-area basis, it is invariant to the choice of spatial resolution. This scale-invariance enables a method for choosing the appropriate spatial resolution for analysis illustrated in Chapter 5 that is unavailable to scale-dependent methods of Chapter 2.

The form of the likelihood equation can be derived from (3.2) and (3.3):

$$\begin{aligned} L(\boldsymbol{\beta}; \mathbf{y}_P) &= m! f(y_1, \ldots, y_m | M = m) P(M = m) && (3.5) \\ &= m! \frac{\prod_{i=1}^{m} \mu(y_i)}{(\mu_{\mathcal{A}})^m} \frac{e^{-\mu_{\mathcal{A}}} (\mu_{\mathcal{A}})^m}{m!} \\ &= e^{-\mu_{\mathcal{A}}} \prod_{i=1}^{m} \mu(y_i). && (3.6) \end{aligned}$$

The $m!$ factor in (3.5) is included because I consider the m points of $\mathbf{y}_P$ to be unordered and thus there are $m!$ arrangements of the m points. An expression of the form of (3.5) is sometimes called a Janossy density (Daley & Vere-Jones, 1988). The log-likelihood is found by taking the logarithm of (3.6):

$$l(\boldsymbol{\beta}; \mathbf{y}_P) = \sum_{i=1}^{m} \ln \mu(y_i) - \mu_{\mathcal{A}}. \tag{3.7}$$

Poisson PPMs are usually fitted by maximising this log-likelihood function (Cressie, 1993).

$\mu_{\mathcal{A}}$ is defined as an integral that is usually intractable and therefore must be approximated. For presence-absence data, (3.7) can be approximated by the likelihood of a logistic regression model applied to grid cells (Brillinger, 1978; Besag *et al.*, 1982). A less biased approach is to use binary regression with a complementary log-log link instead of typical logit link, with an offset equal to the logarithm of the grid cell area (Baddeley *et al.*, 2010). Numerical integration techniques (Davis & Rabinowitz, 1984) can likewise be applied in approximating $\mu_{\mathcal{A}}$:

$$\mu_{\mathcal{A}} \approx \sum_{i=1}^{n} w_i \mu(y_i), \tag{3.8}$$

where $\mathbf{w} = \{w_1, \ldots, w_n\}$ are quadrature weights and $\mathbf{y}_0 = \{y_{m+1}, \ldots, y_n\}$ are quadrature points. A natural way to choose quadrature points is to break the region $\mathcal{A}$ into a regular grid and insert a quadrature point at the centre of each cell. Each cell can then be assigned a quadrature weight which equals its area divided by the number of locations in $\{\mathbf{y}_P, \mathbf{y}_0\}$ contained in the cell.

Substituting (3.8) into (3.7) yields:

$$l(\boldsymbol{\beta}; \mathbf{y}_P) \approx l_{\text{ppm}}(\boldsymbol{\beta}; \mathbf{y}_P, \mathbf{y}_0, \mathbf{w}) = \sum_{i=1}^{m} \ln \mu(y_i) - \sum_{i=1}^{n} w_i \mu(y_i). \tag{3.9}$$

Berman & Turner (1992) showed that (3.9) can be written as a weighted Poisson likelihood:

$$l_{\text{ppm}}(\boldsymbol{\beta}; \mathbf{y}_P, \mathbf{y}_0, \mathbf{w}) = \sum_{i=1}^{n} w_i [z_{w,i} \ln\{\mu(y_i)\} - \mu(y_i)], \tag{3.10}$$

where $z_{w,i} = \frac{I(i \in 1, \ldots, m)}{w_i}$, and $I(\cdot)$ is the indicator function. This construction of the likelihood (3.10) allows PPMs to be posed as Poisson GLMs. Hence PPMs can be fitted using any standard GLM software (R Development Core Team, 2010), as with the `R` package `spatstat` (Baddeley & Turner, 2005).

An alternative representation of the point process likelihood is to use $I(i \in 1, \ldots, m)$ as the response and $\ln w_i$ as an offset term. This would produce a likelihood expression proportional to (3.9), but without the need for a non-integer response.

Assunção & Guttorp (1999) proposed an M-estimator as an alternative to maximum likelihood estimation for Poisson point processes that is robust to contamination (*e.g.* resulting from species misspecification), but the M-estimator is not considered here.

3.2 Processes with Point Interactions

Poisson point processes are defined in part by the assumption of independence of point locations. Hence these processes do not accommodate modelling the distribution of species for which point locations exhibit some form of dependence. Two common approaches to modelling processes with point interactions are Cox processes and Gibbs processes.

3.2.1 Cox Processes

Cox processes (also known as doubly stochastic Poisson processes) are a flexible class of spatial point process in which the intensity $\mu(y)$ is a realisation of some stochastic process $\xi(y)$. In the context of SDM, the assumption is that this stochastic process governs the spatial pattern of factors that influence the distribution of point locations (Diggle, 2003), such as environmental variables and sampling effort, and hence it accounts for spatial correlation in the response. Conditional on $\xi(y) = \mu(y)$, M is an inhomogeneous Poisson point process with intensity $\mu(y)$ (Cressie, 1993). Hence a Cox process can be modelled as a log-linear function of environmental variables plus ξ:

$$\ln \mu(y) = \mathbf{x}(y)'\boldsymbol{\beta} + \xi(y). \tag{3.11}$$

The goal is to maximise the likelihood, which does not have a closed form in general since it must be marginal with respect to the unobserved random ξ. Hence MCMC sampling is often used to estimate the posterior distribution (Møller *et al.*, 1998; Brix & Møller, 2001), but other methods, such as composite likelihood estimation (Guan, 2006) and integrated nested Laplace approximation (Beguin *et al.*, 2012), may also be used.

Some recent applications of Cox processes in SDM include log-Gaussian Cox processes (Møller *et al.*, 1998), their extension to inhomogeneous processes and multi type point pattern time series (Brix & Møller, 2001), and the development of estimating functions for inference of inhomogeneous cluster processes (Waagepetersen, 2007) and estimating functions for inhomogeneous cluster processes where covariate data is missing (Waagepetersen, 2008).

Recently, hierarchical Cox processes have been applied for presence-only SDM (Chakraborty *et al.*, 2011) at the grid cell level. Their hierarchical framework considered three surfaces – (1) the potential intensity surface which is of interest, (2) the availability surface which takes into account the impact of anthropogenic land transformation, and (3) the sampling effort surface. The stochastic process ξ in (3.11) was assumed to be a zero-mean Gaussian process.

3.2.2 Gibbs Processes

In this thesis I will model point interactions via finite Gibbs processes.

Gibbs processes are a broad class of spatial process that can be used to model a spatial pattern of m locations $\mathbf{y}_P = \{y_1, \ldots, y_m\}$. The probability that there are m locations in a Gibbs process is (Cressie, 1993):

$$P(M = m) = \begin{cases} e^{-\mu_{\mathcal{A}}}, & m = 0 \\ \frac{e^{-\mu_{\mathcal{A}}}}{m!} \int_{y \in \mathcal{A}} \mu(y_1, \ldots, y_m) dy_1 \ldots dy_m, & m \geq 1, \end{cases}$$

where $\mu(y_1, \ldots, y_m)$ is the joint intensity of points located in $\mathbf{y}_P$. The conditional density is proportional to $\mu(y_1, \ldots, y_m)$ (Cressie, 1993) and for $m \geq 1$ can thus be written:

$$f(y_1, \ldots, y_m | M = m) = \frac{\mu(y_1, \ldots, y_m)}{\int_{y \in \mathcal{A}} \mu(y_1, \ldots, y_m) dy_1 \ldots dy_m}.$$

Hence the joint (Janossy) density for $m \geq 1$ is:

$$\begin{aligned} f(y_1, \ldots, y_m, m) &= m! f(y_1, \ldots, y_m | M = m) P(M = m) \\ &= m! \frac{\mu(y_1, \ldots, y_m)}{\int_{y \in \mathcal{A}} \mu(y_1, \ldots, y_m) dy_1 \ldots dy_m} \frac{e^{-\mu_{\mathcal{A}}}}{m!} \int_{y \in \mathcal{A}} \mu(y_1, \ldots, y_m) dy_1 \ldots dy_m \\ &= e^{-\mu_{\mathcal{A}}} \mu(y_1, \ldots, y_m). \end{aligned} \tag{3.12}$$

Note that both homogeneous and inhomogeneous Poisson point processes are examples of Gibbs processes. In an inhomogeneous Poisson point process, $\mu(y_1, \ldots, y_m) = \prod_{i=1}^{m} \mu(y_i)$, and plugging this into (3.12) yields (3.6).

Point interactions can be introduced by writing the conditional density as:

$$f(y_1, \ldots, y_m | M = m) = \alpha \prod_{i=1}^{m} [\kappa(y_i)] \rho(\mathbf{y}_P),$$

where α is a normalisation constant, κ is an intensity parameter for the environmental variables and ρ is some function of interactions between points.

The simplest Gibbs interaction processes use interactions between distinct pairs of points and hence have the form:

$$f(y_1, \ldots, y_m | M = m) = \alpha \prod_{i=1}^{m} \kappa(y_i) \prod_{i<j} \rho(y_i, y_j).$$

The form of ρ determines how pairwise interactions affect the density. For example, in a Strauss process (Strauss, 1975),

$$\rho(y_i, y_j) = \begin{cases} 1, & \|y_i - y_j\| > r \\ \gamma, & \|y_i - y_j\| \leq r. \end{cases}$$

While theoretically a Strauss process can model both point repulsion ($\gamma < 1$) or clustering ($\gamma > 1$), it is not integrable for $\gamma > 1$ (Kelly & Ripley, 1976) and hence can only reliably model point inhibition.

Area-interaction processes (Widom & Rowlinson, 1970; Baddeley & van Lieshout, 1995), however, can accommodate both clustering and repulsion among points. Area-interaction processes use interactions among all points within a distance of $2r$ instead of pairwise interactions. They have a conditional density that can be written as follows:

$$f(y_1, \ldots, y_m | M = m) = \alpha \prod_{i=1}^{m} [\kappa(y_i)] \eta^{-U(\mathbf{y}_P)}, \tag{3.13}$$

where $\eta > 0$ and $U(\mathbf{y}_P)$ is the area of the region within $\mathcal{A}$ formed by the union of discs of radius r around each of the m points $y_1, \ldots, y_m \in \mathbf{y}_P$ (Figure 3.1). Note that for point locations $y \in \mathbf{y}_P$ within a distance of r of the boundary of the study region $\mathcal{A}$, the disc around y will extend outside of $\mathcal{A}$, and hence $U(y) < \pi r^2$. An equivalent "canonical scale-free form" of (3.13) used in `spatstat` that is easier to interpret is achieved by transforming κ and η:

$$f(y_1, \ldots, y_m | M = m) = \alpha \prod_{i=1}^{m} [\theta(y_i)] \nu^{-C(\mathbf{y}_P)},$$

where $\theta(y_i) = \kappa(y_i)\eta^{-\pi r^2}$, $\nu = \eta^{\pi r^2}$ and $C(\mathbf{y}_P) = U(\mathbf{y}_P) / \sum_{i=1}^{m} U(y_i) - m$.

Gibbs processes are typically analyzed using the conditional intensity (Papangelou, 1974; Baddeley & Turner, 2005). The conditional intensity $\mu(y, \mathbf{y}_P)$ at a location y given a configuration of locations $\mathbf{y}_P$ is:

$$\mu(y, \mathbf{y}_P) = \frac{f(\mathbf{y}_P \cup y | M = m)}{f(\mathbf{y}_P | M = m)}. \tag{3.14}$$

The conditional intensity essentially gives the conditional probability that a Gibbs Process Y has a point at y given the rest of the process coincides $\mathbf{y}_P$.

For an inhomogeneous Poisson point process,

$$\begin{aligned} \mu(y, \mathbf{y}_P) &= \frac{\prod_{i=1}^{m} \mu(y_i)\mu(y)/(\mu_{\mathcal{A}})^m}{\prod_{i=1}^{m} \mu(y_i)/(\mu_{\mathcal{A}})^m} \\ &= \mu(y). \end{aligned}$$

The fact that $\mu(y, \mathbf{y}_P)$ does not depend on $\mathbf{y}_P$ illustrates the independence of point locations for an inhomogeneous Poisson point process.

For an area-interaction process in canonical scale-free form,

$$\begin{aligned}
\mu(y, \mathbf{y}_P) &= \frac{\alpha \prod_{i=1}^{m} \theta(y_i)\theta(y)\nu^{-C(\mathbf{y}_P \cup y)}}{\alpha \prod_{i=1}^{m} \theta(y_i)\nu^{-C(\mathbf{y}_P)}} \\
&= \theta(y)\nu^{-\{C(\mathbf{y}_P \cup y) - C(\mathbf{y}_P)\}}. \qquad (3.15)
\end{aligned}$$

Closer inspection of the exponent in (3.15) reveals a nice interpretation:

$$\begin{aligned}
-\{C(\mathbf{y}_P \cup y) - C(\mathbf{y}_P)\} &= -\left[\frac{U(\mathbf{y}_P \cup y)}{\pi r^2} - (m+1) - \left\{\frac{U(\mathbf{y}_P)}{\pi r^2} - m\right\}\right] \\
&= -\left[\frac{U(\mathbf{y}_P \cup y) - U(\mathbf{y}_P)}{\pi r^2} - 1\right] \\
&= 1 - \frac{U(\mathbf{y}_P \cup y) - U(\mathbf{y}_P)}{\pi r^2} \\
&= t(y).
\end{aligned}$$

The quantity $t(y)$ is the proportion of the area of the disc of radius r centred around y that overlaps with the discs of radius r centred around the other points in $\mathbf{y}_P$ (Figure 3.1). This means that adding a point y to the pattern contributes a factor of $\theta(y)\nu^{t(y)}$ to the conditional intensity, and the conditional intensity can be simply represented as:

$$\mu(y, \mathbf{y}_P) = \theta(y)\nu^{t(y)}. \qquad (3.16)$$

The value of ν describes the behaviour of point interactions – processes with $\nu < 1$ exhibit inhibition between points, while processes with $\nu > 1$ exhibit clustering of points. Note that $\nu = 1$ reduces the area-interaction process to an inhomogeneous Poisson point process.

An area-interaction model fits the conditional intensity (3.16) at y as a log-linear function of environmental variables $\mathbf{x}(y)$ and point interaction $t(y)$ (Baddeley & Turner, 2005):

$$\ln \mu(y, \mathbf{y}_P) = \mathbf{x}(y)'\boldsymbol{\psi} + t(y) \ln \nu, \qquad (3.17)$$

where $\boldsymbol{\psi}$ is a vector of parameters corresponding to the explanatory variables in $\mathbf{x}(y)$. Note that (3.17) can be represented as log-linear in $(\mathbf{x}(y), t(y))$ with coefficients stored in $\boldsymbol{\beta} = (\boldsymbol{\psi}, \ln \nu)$.

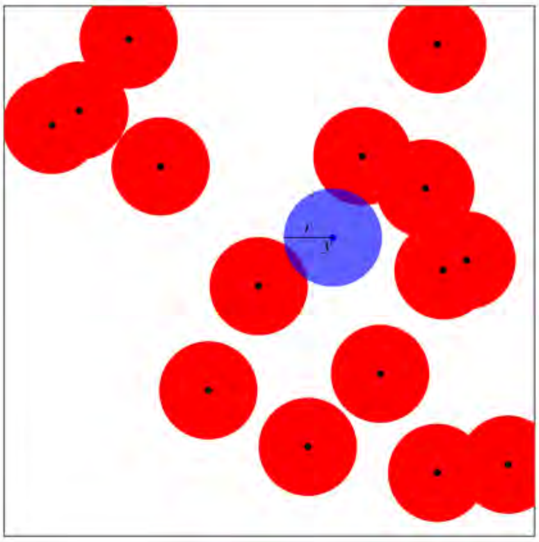

Figure 3.1: Calculating the point interaction for a given point y. First, discs of given radius r are drawn around all points $y_1, \ldots, y_m \in \mathbf{y}_P$, including y. $U(\mathbf{y}_P)$ is the area of the union of all discs. The point interaction at y is the proportion of the (blue) disc around y that intersects the (red) discs around the other points in $\mathbf{y}_P$.

Because α is not of closed-form (Baddeley & van Lieshout, 1995), it is difficult to fit area-interaction models by maximum likelihood. Besag (1977) introduced the pseudolikehood PL as an alternative to the likelihood function for point processes:

$$\ln PL(\boldsymbol{\beta}; \mathbf{y}_P) = \sum_{i=1}^{m} \ln \mu(y_i; \mathbf{y}_P) - \int_{y \in \mathcal{A}} \mu(y, \mathbf{y}_P) dy. \tag{3.18}$$

Note that this is identical to the likelihood of an inhomogeneous Poisson PPM (3.7) if the intensity $\mu(y)$ is replaced with the conditional intensity $\mu(y, \mathbf{y}_P)$. The derivative of (3.18) is an unbiased estimating function (Besag, 1977), and maximum pseudolikelihood estimates are consistent (Jensen & Møller, 1991) and asymptotically Normal (Jensen & Künsch, 1994), at least for pairwise interaction models.

For models with a log-linear conditional intensity such as area-interaction models, the integral in (3.18) is approximated using the Berman-Turner device as for Poisson PPMs (Baddeley & Turner, 2000), enabling area-interaction models to be fitted using standard GLM software (Baddeley & Turner, 2006). This can be done with the `spatstat` package in `R`. Given that area-interaction models have the same form and are fitted in the same way, they can also be fitted as Poisson PPMs with an extra covariate for point interactions. In Chapter 8, I describe the forthcoming `ppmlasso` package which can fit area-interaction models with a LASSO penalty.

Out of the suite of potential models that account for interpoint interaction, I have chosen to use area-interaction models as they can model processes with clustering or repulsion (Baddeley & van Lieshout, 1995) and they have interactions of order m which can capture the potential causes for interpoint dependence more realistically than pairwise interaction models (*e.g.* multiple presence locations observed by the same individual or varying sampling effort). There is no ecological reason why the potential interaction between observations should be restricted to pairs as for example in a Strauss process.

3.3 Goodness of Fit

A key benefit of applying PPMs to SDM is that the modelling framework facilitates the use of a number of goodness-of-fit techniques (Cressie, 1993; Baddeley *et al.*, 2005) to investigate whether the fitted model is appropriate. In particular, it is possible to check the assumption of independence among points (Table 3.1) and to diagnose spatial and environmental effects.

3.3.1 K-Function

One function commonly used for investigating departures from the assumption of independence among points is the K-function (Ripley, 1977). For a homogeneous Poisson point process with intensity μ, the K-function is defined as:

$$K_{\mathrm{H}}(r) = \frac{E[M_0(r)]}{\mu},$$

where $E[M_0(r)]$ is the expected number of further events located within a circle of radius r from a given event. Baddeley & Turner (2000) defined the K-function for inhomogeneous Poisson point processes as follows:

$$K(r) = \frac{1}{|\mathcal{A}|} E \sum_{y_i \in Y \cap \mathcal{A}} \sum_{y_j \in Y \setminus y_i} \frac{I(\|y_i - y_j\| \leq r)}{\mu(y_i)\mu(y_j)}, \tag{3.19}$$

where $\|y_i - y_j\|$ is the Euclidean distance between y_i and y_j. The presence of the intensity at both y_i and y_j in the denominator of (3.19) demonstrates that the K-function is related to the second-order intensity of the process (Ripley, 1976).

An unbiased estimator of the K-function (Baddeley & Turner, 2000) is given by:

$$\hat{K}(r) = \frac{1}{|\mathcal{A}|} \sum_{y_i \in Y \cap \mathcal{A}} \sum_{y_j \in Y \cap \mathcal{A} \setminus y_i} \frac{w_{i,j} I(\|y_i - y_j\| \leq r)}{\mu(y_i)\mu(y_j)}, \tag{3.20}$$

where $w_{i,j}$ is an edge correction factor (Ripley, 1977) equal to the reciprocal of the proportion of the circumference of a circle centred at y_i with radius y_j that is within the study area $\mathcal{A}$. This eliminates negative bias incurred at points that lie near the boundary of $\mathcal{A}$ for which there could be unobserved points within a distance of r but outside of $\mathcal{A}$.

As the intensity μ is unknown, it must be estimated using the data in order to determine $\hat{K}(r)$. For SDM, this is achieved by fitting a Poisson PPM to the data as a function of environmental covariates, and substituting the fitted intensity $\hat{\mu}$ in the demoninator of (3.20). This extra level of uncertainty from estimating μ can make it difficult to distinguish variation in the intensity surface due to the environment from variation due to dependence among species locations (Baddeley & Turner, 2000; Diggle, 2003).

Because the sampling distribution of $\hat{K}$ is intractable, goodness-of-fit tests using the K-function are performed in practice by comparing the observed value of $\hat{K}(r)$ to values $\hat{K}_{\text{sim}}(r)$ calculated from simulations of an inhomogeneous Poisson point process with true intensity surface equal to the observed intensity surface. It is possible to construct a C% simulation envelope (Diggle, 2003) by determining the $(1-C)/2$th and $1-(1-C)/2$th quantiles of the s simulations for varying values of the distance r. Figure 3.2 illustrates 95% simulation envelopes for the distribution of the

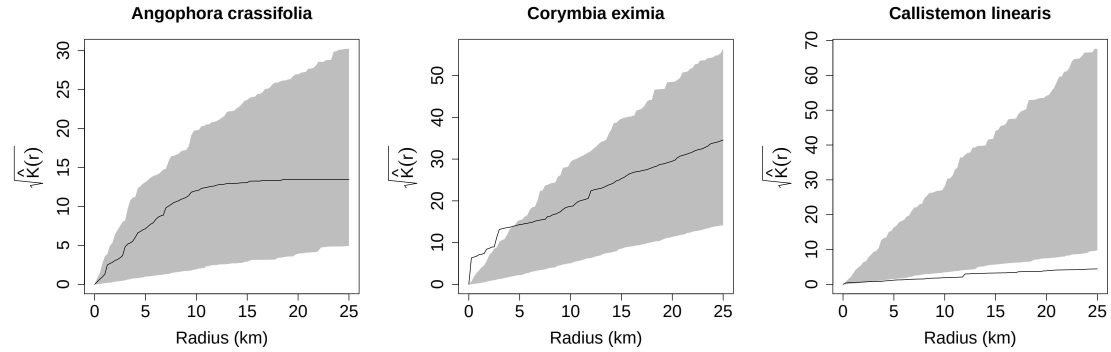

Figure 3.2: Assessing goodness-of-fit for a Poisson point process using 95% simulation envelopes of $\hat{K}(r)$ for three Sydney eucalypts. The observed function $\hat{K}(r)$ falls within the simulation envelope for *Angophora crassifolia*, suggesting that the assumption of independent events conditional on the environment is reasonable. However, the observed $\hat{K}(r)$ deviates above and below the envelope for *Corymbia eximia* and *Callistemon linearis*, respectively. This suggests additional clustering for *C. eximia* at radii $r < 5$ and inhibition of *C. linearis* for all $r \leq 25$.

---

locations of three eucalypts in the Blue Mountains Region near Sydney (NSW Office of Environment and Heritage, 2012). For *Angophora crassifolia*, the assumption of independent points conditional on the environment appears to be reasonable, as the observed $\hat{K}(r)$ falls within the simulation envelope at all radii r. However, $\hat{K}(r)$ deviates above the envelope for small radii r for *Corymbia eximia*, suggesting additional clustering than what is expected for an inhomogeneous Poisson point process. $\hat{K}(r)$ for *Callistemon linearis* falls below the envelope for all radii $r \leq 25$, suggesting repulsion among points or a “regular” process. Consequently, applying a model to *C. eximia* and *C. linearis* that accounts for dependence among points such as an area-interaction model may be appropriate.

Although the K-function provides a picture of the second-order properties of a spatial point process, these second-order properties do not completely characterise the process (Baddeley & Turner, 2000). Non-Poisson point processes with the same K-function as either a homogeneous or inhomogeneous point process can be constructed (Baddeley & Silverman, 1984; Baddeley & Turner, 2000). Consequently, K-functions should only be used to reject assumptions of independence among points, not fail to reject them.

3.3.2 Residuals

Baddeley *et al.* (2005) developed residuals and residual plots to assess goodness of fit by using the conditional intensity $\mu(y, \mathbf{y}_P)$ (3.14). For a model fitted with parameters $\widehat{\boldsymbol{\beta}}$ and conditional intensity $\hat{\mu}(y, \mathbf{y}_P)$, the residuals are defined over some region $B \in \mathcal{A}$ as:

$$R(B, \hat{h}, \widehat{\boldsymbol{\beta}}) = \sum_{y_i \in \mathbf{y}_P \cap B} \hat{h}(y_i, \mathbf{y}_P \backslash \{y_i\}) - \int_B \hat{h}(y, \mathbf{y}_P) \hat{\mu}(y, \mathbf{y}_P) dy,$$

where $\hat{h}(y, \mathbf{y}_P)$ is a non-negative function. The sum of residuals has expected value zero as a consequence of the Georgii-Nguyen-Zessin (GNZ) formula (Georgii, 1976; Xanh & Zessin, 1979):

$$E\left[\sum_{y_i \in \mathbf{y}_P} h(y_i, \mathbf{y}_P \backslash \{y_i\})\right] = E\left[\int_B h(y, \mathbf{y}_P) \mu(y, \mathbf{y}_P) dy\right]. \tag{3.21}$$

Different choices of the function $h(y, \mathbf{y}_P)$ lead to different forms of residual, as shown in Table 3.2.

Plots of these residuals can identify extreme values and departures in spatial trend from the fitted model. For example, Figure 3.3 shows smoothed Pearson residuals for a Poisson PPM fitted to *Corymbia eximia* presences. The presence of a pattern in which residuals are most positive toward the central coast part of the

---

Table 3.2: Different forms of residuals for checking goodness of fit.

Residual	$h(y, \mathbf{y}_P)$
Raw	1
Inverse μ	$1/\mu(y, \mathbf{y}_P)$
Pearson	$1/\sqrt{\mu(y, \mathbf{y}_P)}$
Pseudoscore	$\frac{\partial}{\partial \boldsymbol{\beta}} \ln \mu(y, \mathbf{y}_P)$

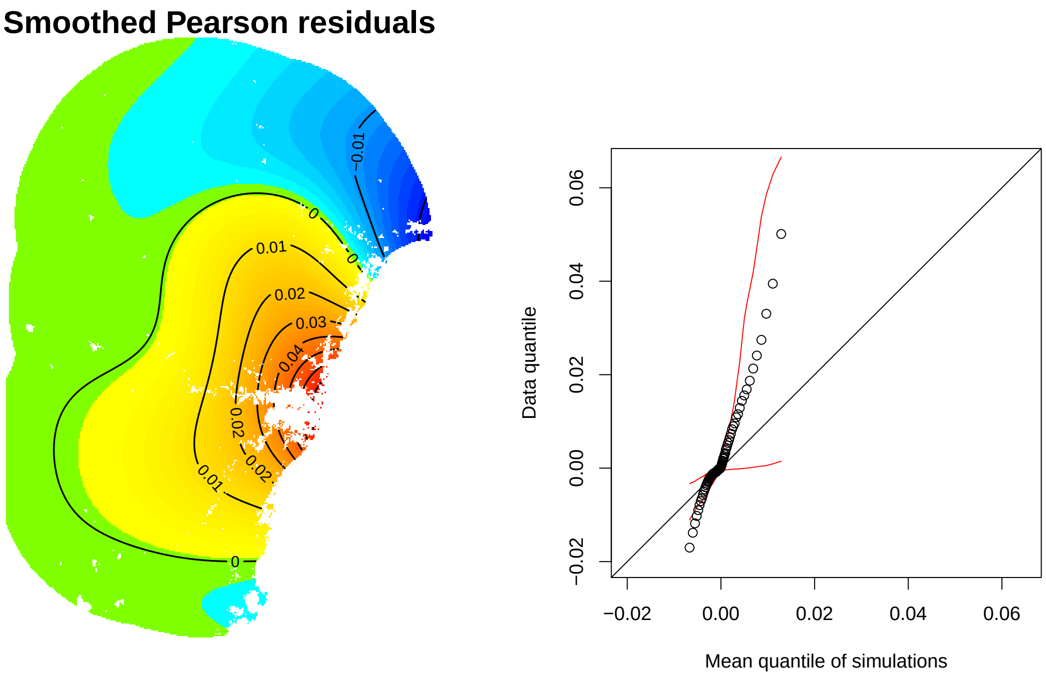

Figure 3.3: Spatial plot (left) and quantile plot (right) of smoothed Pearson residuals for a Poisson PPM. The existence of a trend in the spatial plot suggests that the fitted Poisson PPM is inappropriate. The significant departure from the straight line in the form of heavy tails for the data quantiles suggests that presence locations are clustered.

study region (near Sydney) while most negative along the northern coast suggests a spatial trend uncaptured by the fitted Poisson PPM.

An alternative to K-functions for identifying the presence of point interactions is to construct a quantile plot of the smoothed residuals versus their expected values under the fitted model obtained through simulation. Departures from a straight line indicate that point interactions are not properly modelled. For example the Poisson PPM fitted to the *C. eximia* presence data, Figure 3.3 deviates significantly from a straight line. The heavy tails for the data quantiles imply the existence of clustering among presence locations.

Given that the K-envelope (Figure 3.2) suggests clustering of points within a radius of 5 km for *C. eximia*, an area-interaction model with radius of 5 km is a

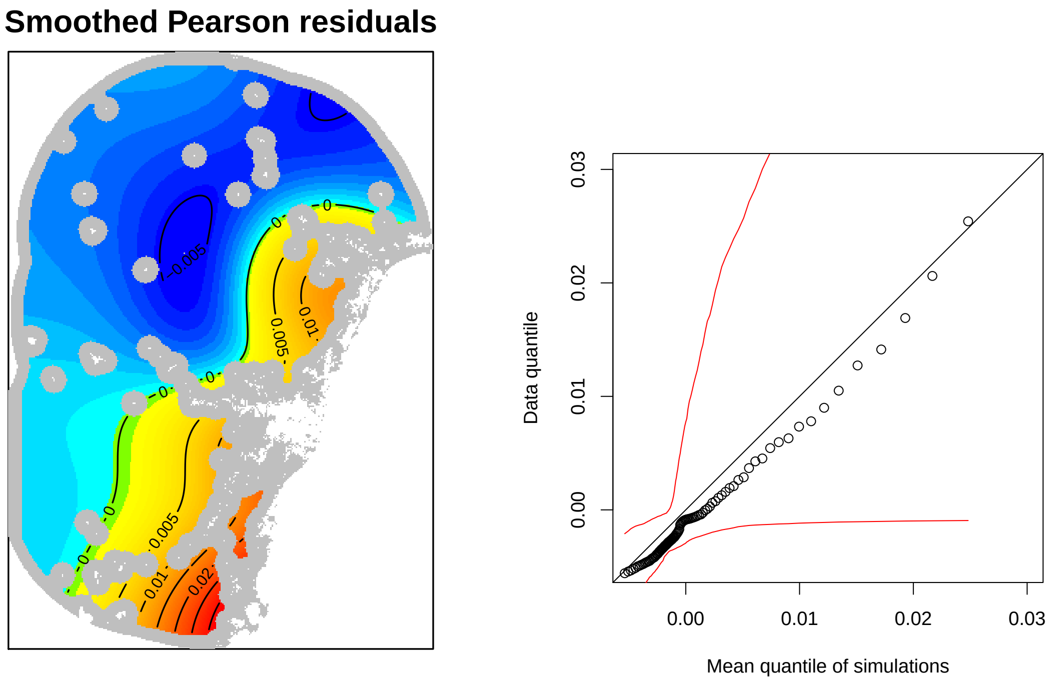

Figure 3.4: Spatial plot (left) and quantile plot (right) of the smoothed Pearson residuals for the area-interaction model with radius 5 km. There still exists a spatial trend, although it is different to that of Figure 3.3. The data quantiles fall along a straight line well within the 2.5th and 97.5th quantiles of simulations, suggesting that the model has adequately captured the structure of point clustering.

---

natural choice for an alternative model. Figure 3.4 shows the spatial and quantile plots of the smoothed Pearson residuals for this model. There is still a spatial trend evident, although the area underestimated by the model has shifted to the southern coast. However, the quantile plot suggests agreement between the residuals from the data and simulated residuals, and hence that the area-interaction model with radius 5 km adequately captures the structure of point clustering.

It is also possible to test for the dependence of a point process on a spatial covariate using both parametric (Berman, 1986; Lawson, 1988; Waller *et al.*, 1992) and non-parametric methods (Guan, 2008; Baddeley *et al.*, 2012), and hence identify potentially missing spatial variables in the model. In `spatstat`, this can be done by plotting them against cumulative residuals, although that is not explored here.

There are functions in `spatstat` for both fitting and performing these diagnostic checks on a large suite of Gibbs processes (Baddeley & Turner, 2005).

3.4 Summary

PPMs are a natural approach to presence-only modelling with strong theoretical foundations and readily available software for fitting and model checking. It is therefore surprising that while point pattern methods have been used in ecology for a long time (Cressie, 1993), PPMs have only recently been proposed for SDM (Warton & Shepherd, 2010; Chakraborty *et al.*, 2011).

Due to their advantages in interpretation, validation and ease of implementation, PPMs form a major component of the PPM-LASSO method advanced in this thesis, in which they are fitted with a LASSO penalty. In Chapter 5, I establish the equivalence of Poisson PPMs and MAXENT and further derive and illustrate the comparative advantages of point process models, while in Chapter 8, I describe the `ppmlasso` package for `R` that I have developed to fit PPMs with a LASSO penalty.

Chapter 4

LASSO and Its Extensions

4.1 Introduction

In surveys of predictive performance of SDM methods (*e.g.* Elith *et al.*, 2006), one way in which some high-performing methods differ from other methods is that they are applied with regularisation tools aimed at reducing model complexity. For example, MAXENT software by default uses a LASSO penalty, which shrinks parameter estimates $\widehat{\boldsymbol{\beta}}$ toward zero. In this Chapter I will review LASSO and related methods.

There are a number of benefits in shrinking parameter estimates toward zero:

- **Predictive Ability**: Unconstrained models are susceptible to overfitting, *i.e* they fit the data in the model well but may not predict the response for new data well. Shrinking parameter estimates introduces bias but decreases variance of $\widehat{\boldsymbol{\beta}}$ (Hastie *et al.*, 2009). If the decrease in variance is greater than the increase in bias, the mean squared error of $\widehat{\boldsymbol{\beta}}$ will be lower. Finding a balance in the "bias-variance tradeoff" is key to a model that fits the data reasonably well and has good predictive performance.
- **Numerical Stability**: The number of available variables p may exceed the number of observations m, in which case some sort of regression shrinkage is

required to find a unique solution.

- **Interpretation**: Some shrinkage procedures may perform variable selection by shrinking some parameter estimates to zero. Reducing the number of candidate variables helps to explain which biological factors are important in determining a species' distribution.

In Section 4.2, I describe a number of shrinkage procedures, including LASSO, which is a key component of this thesis. Sections 4.3 and 4.4 provides further details of how to fit models with a LASSO penalty and how to choose the LASSO penalty.

4.2 Regularisation Methods

Shrinkage procedures optimise some objective function subject to a constraint on the size of parameter estimates. In the case of GLMs, shrinkage procedures thereby maximise the constrained likelihood with penalty function $p(\boldsymbol{\beta})$:

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmax} l(\boldsymbol{\beta}) \quad \text{s.t.} \quad p(\boldsymbol{\beta}) \leq C, \tag{4.1}$$

where C is a constant. The form of $p(\boldsymbol{\beta})$ determines the shape of the constraint region (Figure 4.1), and is usually chosen to force the solution $\widehat{\boldsymbol{\beta}}$ to be within some neighbourhood of the origin. The choice of C controls the degree to which $\widehat{\boldsymbol{\beta}}$ is shrunk toward the origin.

An equivalent formulation of the optimisation problem given by (4.1) is found by applying the Lagrangian function with Lagrangian multiplier $\boldsymbol{\lambda}$ (Osborne *et al.*, 2000*b*):

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmax}[l(\boldsymbol{\beta}) - \boldsymbol{\lambda} p(\boldsymbol{\beta})].$$

Table 4.1 shows the most common shrinkage methods and their respective penalties $p(\boldsymbol{\beta})$. The LASSO ($\gamma = 1$) and ridge ($\gamma = 2$) penalties are special cases of what are known as bridge penalties (Frank & Friedman, 1993), which have the form

Table 4.1: Penalty terms for various shrinkage methods.

Shrinkage Method	**Penalty Term**
LASSO	$\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$
Ridge	$\lambda \sum_{j=1}^{p} \beta_j^2$
Fused LASSO	$\lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j=2}^{p} \lvert\beta_j - \beta_{j-1}\rvert$
Group LASSO	$\lambda \sum_{j=1}^{J} (\boldsymbol{\beta}_j' \mathbf{K} \boldsymbol{\beta}_j)^{1/2}$ for positive definite $\mathbf{K}$
Adaptive LASSO	$\lambda \sum_{j=1}^{p} w_j \lvert\beta_j\rvert$
Elastic Net	$\lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j=1}^{p} \beta_j^2$

$p(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^\gamma$. Bridge penalties with $\gamma \geq 1$ have convex constraint regions and hence are easier to fit (Hastie *et al.*, 2009). Note that a classical approach to variable selection (*e.g.* all-subsets selection) can be thought of as a bridge-type penalty method with $\gamma = 0$.

Ridge regression (Hoerl & Kennard, 1970) applies an $L2$ penalty on parameter coefficients, in effect shrinking them toward zero at a rate proportional to their magnitude. It is particularly effective in shrinking parameters of variables that are highly correlated with other variables (Hesterberg *et al.*, 2008). The form of the penalty is such that parameter estimates may be fitted even when the design matrix $\mathbf{X}$ is not full rank (Hastie *et al.*, 2009) but the shape of the constraint region precludes any of the coefficients from being shrunk exactly to zero (Hesterberg *et al.*, 2008; Hastie *et al.*, 2009).

LASSO (Tibshirani, 1996) is very popular, with the original article cited over 1,500 times in 2012 alone according to Google Scholar. One reason for the ubiquity of the LASSO is that it performs model fitting and variable selection simultaneously (Hastie *et al.*, 2009). For SDM, this is attractive because applying a LASSO penalty both fits a model to the environmental data and can eliminate environmental variables that are not informative in determining the species' distribution.

A number of extensions to the LASSO have arisen to improve it for particular types of data or to address some of its weaknesses. For example, the fused LASSO

(Tibshirani *et al.*, 2005) and group LASSO (Yuan & Lin, 2006) are useful to encourage neighbouring coefficients and groups of coefficients to be shrunk together, respectively.

One of the most popular extensions is adaptive LASSO (Zou, 2006), which adds an initial weight w_j to the penalty term for each coefficient β_j in order to shrink more important variables less, and hence reduce the bias incurred from shrinkage. Given an initial solution $\widehat{\boldsymbol{\beta}}_{\text{init}}$, the adaptive weight for the jth variable is $w_j = 1/|\hat{\beta}_{\text{init},j}|^\gamma$. Typically, these weights are chosen to be the reciprocal of either the unpenalised solution ($w_j = 1/|\hat{\beta}_{\text{GLM},j}|$) or the penalised solution that optimises one of the more common criteria like BIC (Schwarz, 1978) ($w_j = 1/|\hat{\beta}_{\text{BIC},j}|$). This strategy is consistent in variable selection (Zou, 2006), which is not necessarily the case with LASSO, as discussed in Section 4.5.

The elastic net (Zou & Hastie, 2005) affixes both a LASSO and a ridge penalty to the objective function, which produces sparse solutions that are often superior to LASSO for correlated variables and hence can distinguish unknown grouping within the variable structure.

Figure 4.1 shows the geometry of the constraint region for ridge regression, LASSO, adaptive LASSO and elastic net in the simple case of two parameters β_1 and β_2. As the unpenalised solution $\widehat{\boldsymbol{\beta}}_{\text{GLM}}$ falls outside each constraint region, each method shrinks parameter estimates toward zero. The LASSO, adaptive LASSO and elastic net regions have sharp corners, encouraging but not guaranteeing sparsity, while the ridge regression constraint region does not and hence will not eliminate any variables from the model. Note that the diamond shape of the adaptive LASSO is stretched in the direction of the parameter estimate of higher magnitude.

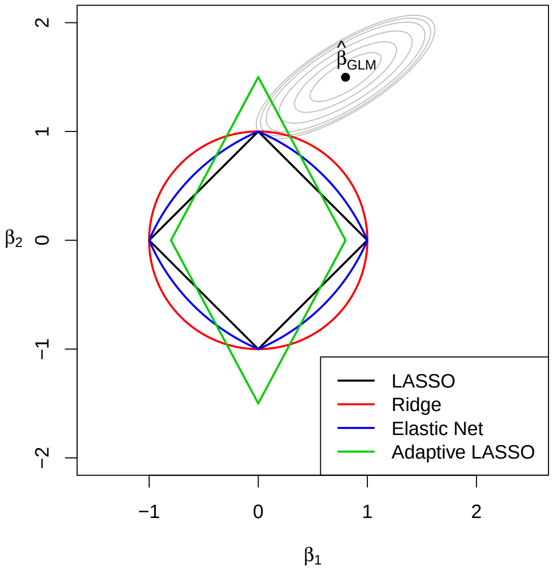

Figure 4.1: Constraint regions for different forms of penalty. The unpenalised solution occurs at $\widehat{\boldsymbol{\beta}}_{\text{GLM}} = (0.8, 1.5)$. For each type of penalty, the solution occurs at the point where contours of the likelihood surface first intersect the constraint region. For LASSO, $|\beta_1| + |\beta_2| = 1$. For ridge regression, $\beta_1^2 + \beta_2^2 = 1$. For the elastic net, $\alpha(|\beta_1| + |\beta_2|) + (1-\alpha)(\beta_1^2 + \beta_2^2) = 1$. For adaptive LASSO, $w_1|\beta_1| + w_2|\beta_2| = 1$, where $w_1 = 1/\hat{\beta}_{\text{GLM},1}$ and $w_2 = 1/\hat{\beta}_{\text{GLM},2}$.

4.3 Fitting Models with a LASSO Penalty

Because it has been used successfully in SDM and remains very popular, I will focus on the LASSO in this thesis. When applying a LASSO penalty to a GLM, the parameter estimates $\widehat{\boldsymbol{\beta}}$ are found by maximising the constrained likelihood:

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmax} l(\boldsymbol{\beta}) \quad \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq C,$$

where C is a constant, or equivalently,

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmax} l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j|. \tag{4.2}$$

The form of the constraint means the Karush-Kuhn-Tucker conditions (Osborne

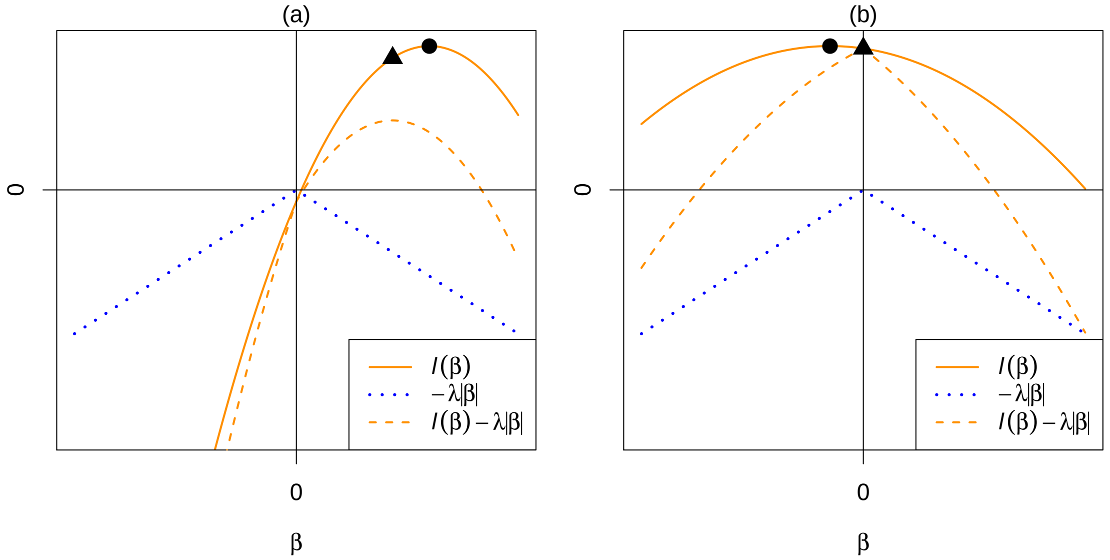

Figure 4.2: Geometry of the LASSO solution. For a GLM, the LASSO solution $\hat{\beta}$ (▲) is found by maximising the constrained likelihood (4.2). The constrained likelihood (dashed curve) is equal to the likelihood (solid curve) plus the penalty (dotted line). Geometrically, the LASSO solution $\hat{\beta}_j$ is found by moving down the likelihood curve from the unpenalised estimate $\hat{\beta}_{\text{GLM},j}$ (●) toward the origin until the derivative of the likelihood $s(\hat{\beta}_j)$ becomes equal to the absolute value of the penalty $|\lambda|$ (4.3). In (a), this means that at $\hat{\beta}$, the gain from reducing the penalty when moving toward zero no longer exceeds the loss in likelihood. If $|s(\beta_j)| < \lambda$ for all β_j between 0 and $\hat{\beta}_{\text{GLM},j}$ (●), $\hat{\beta}_j$ is set to 0 (4.4), as in case (b). Beyond 0, the penalty starts decreasing as well as $l(\beta)$, which is clearly sub-optimal.

---

et al., 2000*a*) ensure that:

$$s(\hat{\beta}_j) = \lambda\,\text{sign}(\hat{\beta}_j),\ \hat{\beta}_j \neq 0 \tag{4.3}$$

$$|s(\hat{\beta}_j)| \leq \lambda,\ \hat{\beta}_j = 0, \tag{4.4}$$

where $s(\beta_j) = \partial l(\boldsymbol{\beta})/\partial \beta_j$ is the jth score function. Figure 4.2 illustrates how the LASSO solution $\widehat{\boldsymbol{\beta}}$ is therefore derived.

A number of algorithms have been developed to fit models regularised with a LASSO penalty. Because of its intuition in relation to the geometry depicted in Figure 4.2, I will focus mostly on the coordinate descent algorithm of Osborne *et al.* (2000*b*) developed for linear models. In a linear model, the goal is to minimise residual sum of squares $\mathbf{r}'\mathbf{r}$, where $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. The Osborne algorithm is essentially as follows:

1. **Propose Update**: Let $\boldsymbol{\sigma}$ denote the active set, *i.e.* the indices of the nonzero coefficients. At the ith iteration, propose an update $\widehat{\boldsymbol{\beta}}^{i}_{\boldsymbol{\sigma}}$ to the current estimate $\widehat{\boldsymbol{\beta}}^{i-1}_{\boldsymbol{\sigma}}$ based on a local linearisation of $\mathbf{r}'\mathbf{r}$ about $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{i-1}_{\boldsymbol{\sigma}}$. Essentially, this moves all nonzero coefficients toward zero in a direction determined by their respective score functions, as in Figure 4.2.
2. **Delete Variables**: If $\text{sign}(\widehat{\boldsymbol{\beta}}^{i}_{\boldsymbol{\sigma}}) = \text{sign}(\widehat{\boldsymbol{\beta}}^{i-1}_{\boldsymbol{\sigma}})$, proceed to Step 3. Otherwise, calculate the direction of the update $\widehat{\boldsymbol{\beta}}^{\Delta}_{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\beta}}^{i}_{\boldsymbol{\sigma}} - \widehat{\boldsymbol{\beta}}^{i-1}_{\boldsymbol{\sigma}}$. Determine the first coefficient k to change sign, and calculate $\rho = |\hat{\beta}^{i-1}_{k} / \hat{\beta}^{\Delta}_{k}|$ such that $\hat{\beta}^{i-1}_{k} + \rho \hat{\beta}^{\Delta}_{k} = 0$. Set $\widehat{\boldsymbol{\beta}}^{i}_{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\beta}}^{i-1}_{\boldsymbol{\sigma}} + \rho \widehat{\boldsymbol{\beta}}^{\Delta}_{\boldsymbol{\sigma}}$. Delete k from the active set $\boldsymbol{\sigma}$. Return to Step 1.
3. **Add Variables**: Calculate the score functions $s(\widehat{\boldsymbol{\beta}}^{i}) = \mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{i})$. If $|s(\widehat{\boldsymbol{\beta}}^{i}_{k})| > \lambda$ for any $k \notin \boldsymbol{\sigma}$, add the most violated coefficient $\text{argmax}_{k \notin \boldsymbol{\sigma}} |s(\widehat{\boldsymbol{\beta}}^{i}_{k})|$ to the active set and return to Step 1.

It is important that variables are deleted and added one at a time to avoid infinite loops. Although the Osborne algorithm was originally posed for linear models, I have adapted it for fitting PPMs in the `ppmlasso` package in `R`, described in Chapter 8.

An alternative to the Osborne algorithm is a modification of the least angle regression (LARS) algorithm (Efron *et al.*, 2004), which can fit an entire regularisation path of LASSO-penalised models at the computational cost of a single least-squares fit. The LARS algorithm builds the solution from the intercept model by finding the variable that is most correlated with the current residuals and moving in its direction until another variable has equal correlation, at which point the algorithm moves in a direction equiangular between the two included variables, proceeding in this manner until the unpenalised solution. Park & Hastie (2007) extended the LARS algorithm to determine a regularisation path of LASSO-penalised GLM estimates, but these estimates are only approximations for values of λ that do not coincide with change points in the active set $\boldsymbol{\sigma}$.

4.4 Choosing the LASSO Penalty

A key question with fitting models penalised by LASSO is determining how large the penalty λ should be. The value of λ controls the complexity of the fitted model (4.2) and an appropriate choice can yield advantages in predictive performance. Tibshirani (1996) proposed generalised cross validation (Craven & Wahba, 1979) and Stein's unbiased risk estimator as criteria for choosing the LASSO penalty, while Fu (2005) proposed non-linear GCV as an extension of GCV for use in GLMs. A common approach is to choose the LASSO penalty by optimising AIC (Akaike, 1974) or BIC. Zou *et al.* (2007) established that the number of nonzero parameters is an unbiased estimate of the degrees of freedom of the LASSO, which has been used to calculate AIC and BIC. BIC is known to produce parameter estimates that are consistent in estimation, while AIC and GCV tend to overfit (Zhang *et al.*, 2010). However, AIC is asymptotically loss-efficient, while BIC is not (Zhang *et al.*, 2010). The question of which methods perform best for choosing the LASSO penalty in PPMs is explored in detail in Chapter 7.

Some desirable properties of any estimate $\widehat{\boldsymbol{\beta}}$ are that it be consistent in estimation, which implies that $\widehat{\boldsymbol{\beta}}$ converges to the vector of true parameter values $\boldsymbol{\beta}_*$ as sample size grows large, and consistent in variable selection, which implies that the subset of nonzero variables converges to the true subset of nonzero variables as the sample size grows large. The choice of LASSO penalty λ impacts whether consistency in estimation is achieved – Knight & Fu (2000) showed that LASSO estimates are $\sqrt{n}$-consistent as $n \rightarrow \infty$ under mild regularity conditions for linear models if λ has order $O(\sqrt{n})$. However, there is a further nontrivial condition involving correlation among variables that is necessary for the LASSO to be consistent in variable selection (Zhao & Yu, 2006; Zou, 2006; Yuan & Lin, 2007). This point will be elaborated in detail in Chapter 6. Indeed, the motivation for the development of adaptive LASSO was that it is consistent in both estimation and variable selection if the adaptive weights are generated from a $\sqrt{n}$-consistent estimate (Zou, 2006).

4.5 Summary

LASSO is a useful tool for regularisation that can improve models in the ways outlined in Section 4.1. As shown in this Chapter, the statistics literature has provided a number of algorithms for fitting models with a LASSO penalty as well as theory and methods concerning the choice of the LASSO penalty λ, which I use in this thesis to improve the way current SDM methods are implemented. In Chapter 5, I show that MAXENT can be further improved by using different choices of the LASSO penalty than what is currently used based on developments in the statistics literature.

Part II —

New Results for Presence-Only Analysis

Chapter 5

Equivalence of MAXENT and Poisson Point Process Models

5.1 Introduction

Recall from Chapter 2 that MAXENT (Phillips *et al.*, 2006), based on a maximum entropy approach, is a particularly popular method of presence-only analsysis, having been cited 606 times in 2012 according to Google Scholar. Its rise in popularity has been meteoric, having only been introduced to ecology seven years ago, although the concept of maximum entropy modelling has been around for a long time (Jaynes, 1957). One motivation for MAXENT is that it is said to make no additional assumptions to what is known from the data (Phillips *et al.*, 2006). A comprehensive study of current SDM methods found MAXENT to outperform nearly all other methods (Elith *et al.*, 2006), and this may explain its prevalence in the literature. The maximum entropy approach has been used elsewhere in ecology, *e.g.* to predict biodiversity using species traits (Shipley *et al.*, 2006), to predict species-area relationships at large spatial resolutions from small census plots (Harte *et al.*, 2009), and to infer the strength of interspecies interactions in tropical forests (Volkov *et al.*, 2009).

Nevertheless, MAXENT has a number of shortcomings, some of which are described in Section 2.2 and illustrated later in Sections 5.3 and 5.4. In particular, it is unclear what diagnostic tools may be used to assess whether the fitted model is reasonable. Moreover, MAXENT analyses data after first aggregating it into presence/absence grid cells (as in Figure 5.1), and it is currently unclear what spatial resolution should be used when constructing these grid cells. Further, as in Section 2.2 some key components of the output such as the intercept and fitted probabilities are dependent on this choice of spatial resolution and hence are scale-dependent. Given this dependence, analyses performed using MAXENT at different spatial resolutions may not be logically compatible, as in the case of logistic regression (Baddeley *et al.*, 2010). Moreover, while probabilities of species occurrence can be obtained for presence-only data, the fitted probabilities of MAXENT form a habitat suitability index of relative probabilities (Royle *et al.*, 2012). These drawbacks illustrate that MAXENT has not been described with strong statistical foundations. Elith *et al.* (2011) attempted to explain MAXENT in statistical terms by establishing that MAXENT minimises the relative entropy between the distribution estimated from the presence-only data and the distribution estimated from the background, but it is currently unclear how MAXENT relates to other SDM methods.

In this Chapter I show that MAXENT is mathematically equivalent to Poisson regression (McCullagh & Nelder, 1989) and related to a Poisson PPM (Chapter 3). Relationships between maximum likelihood and maximum entropy have been known for a long time – this relationship was explored for exponential families in the late 1950s (Kullback, 1959), while an equivalence for contingency tables was established in 1963 (Good, 1963), and maximum entropy was later linked to the maximum likelihood of a Gibbs distribution (Della Pietra *et al.*, 1997). Nonetheless, the direct link I make between MAXENT and Poisson PPMs is new.

Warton & Shepherd (2010) introduced Poisson PPMs as a way to address “problems of model specification, interpretation, and implementation” inherent in pseudo-

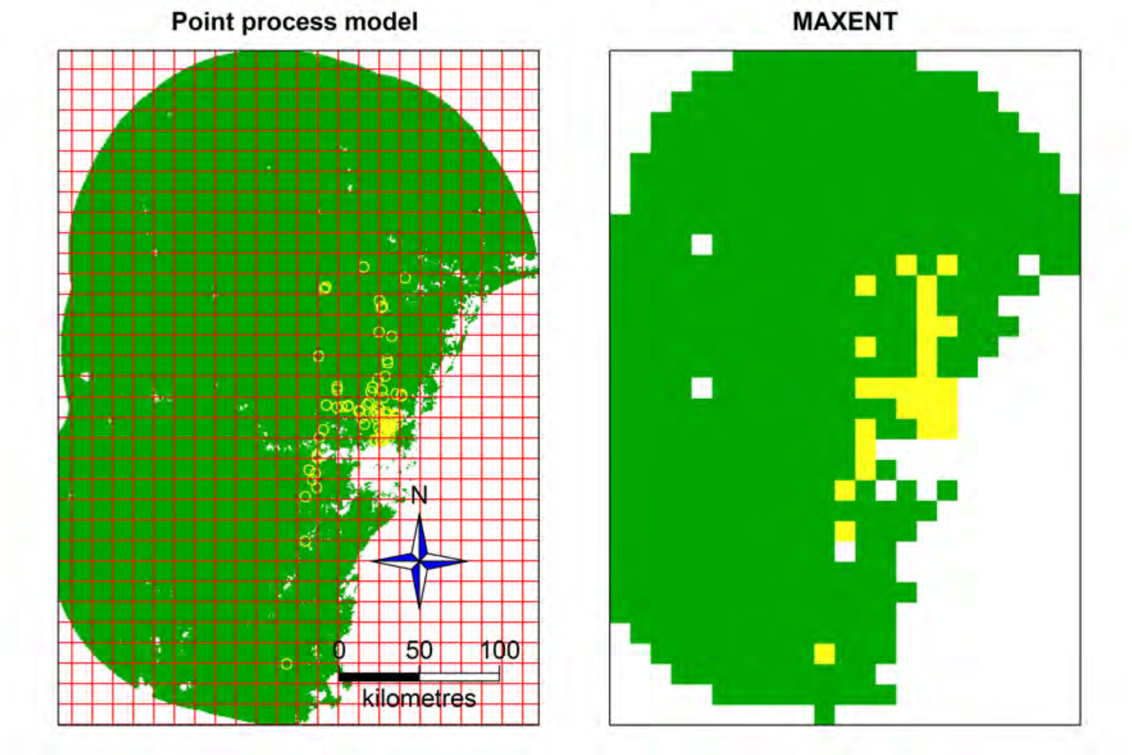

Figure 5.1: Comparison of point process and MAXENT framework for *Corymbia eximia*. A point process model (left) analyses presence points $\mathbf{y}_P = \{y_1, \ldots, y_m\}$; MAXENT (right) analyses presence/absence in grid cells $\{g_1, \ldots, g_n\}$, with $n = 258$ here. A key issue with MAXENT is determining how many grid cells n to use for analysis.

---

absence regression, another popular SDM method. This Chapter achieves a similar goal in relation to MAXENT – all of the problems described in Sections 5.3 and 5.4 can be addressed by reframing the problem using a Poisson PPM. Section 5.2 demonstrates the equivalence of Poisson PPMs and MAXENT. Section 5.3 demonstrates by example how this equivalence can improve on current practice in MAXENT modelling. Finally, Section 5.4 demonstrates that these proposed improvements can led to more accurate predictions of a species' actual distribution. The contents of this Chapter have been published in *Biometrics* (Renner & Warton, 2013).

5.2 Equivalence of MAXENT and Poisson point process models

Rather than using the presence-only locations $\mathbf{y}_P = \{y_1, \ldots, y_m\}$, the MAXENT procedure analyses data by splitting the study region $\mathcal{A}$ into n grid cells with centres at the locations in $\mathbf{g} = \{g_1, \ldots, g_n\}$. A binary response vector $\mathbf{z}^{(n)}(\mathbf{g}) = \{z^{(n)}(g_1), \ldots, z^{(n)}(g_n)\}$ is formed where $z^{(n)}(g_i) = 1/m^{(n)}$ if the ith grid cell contains at least one presence location and 0 otherwise, and $m^{(n)}$ is the count of grid cells that contain at least one presence location. Without loss of generality, I partition $\{g_1, \ldots, g_n\}$ as $\{\mathbf{g}_P, \mathbf{g}_0\}$, where $\mathbf{g}_P = \{g_1, \ldots, g_m^{(n)}\}$ are the $m^{(n)}$ presence cells. I index z and m with the superscript (n) to emphasise that these quantities depend on the spatial resolution (and hence the number of grid cells n) used in analysis. The goal in MAXENT is to model $\pi(g_i)$, the probability that if there is one presence then it is located in the ith grid cell as a function of p environmental variables $\mathbf{x}(g_i)$. $\boldsymbol{\pi}(\mathbf{g}) = \{\pi(g_1), \ldots, \pi(g_n)\}$ is estimated to maximise the entropy (Jaynes, 1957) $H\{\boldsymbol{\pi}(\mathbf{g})\} = -\sum_{i=1}^{n} \pi(g_i) \ln \pi(g_i)$, subject to two types of constraint:

$$\sum_{i=1}^{n} \pi(g_i)\, x_j(g_i) = \frac{1}{m^{(n)}} \sum_{i=1}^{m^{(n)}} x_j(g_i),\ \forall j, \tag{5.1}$$

$$\sum_{i=1}^{n} \pi(g_i) = 1. \tag{5.2}$$

(5.1) ensures that the predicted mean of each environmental variable equals its observed mean for the presence data while (5.2) ensures that the probabilities add to one.

In this Chapter, I assume the existence of a unique maximum of the objective function (entropy and likelihood for MAXENT and Poisson regression, respectively). I will show that the MAXENT procedure is equivalent to log-linear Poisson regression when applied to grid cell data $\mathbf{z}^{(n)}(\mathbf{g})$. That is, I model the mean of $z^{(n)}(g_i)$ as a log-linear model:

$$\ln \mu_i = \mathbf{x}(g_i)'\boldsymbol{\beta}. \tag{5.3}$$

The parameters $\boldsymbol{\beta}$ are estimated to maximise the likelihood function (McCullagh & Nelder, 1989):

$$l\{\boldsymbol{\beta};\mathbf{z}^{(n)}(\mathbf{g})\} = \sum_{i=1}^{n} z^{(n)}(g_i)\,\ln\mu(g_i) - \sum_{i=1}^{n}\mu(g_i) - \sum_{i=1}^{n}\ln\{z^{(n)}(g_i)!\}. \tag{5.4}$$

On face value, this analysis appears to be based on a nonsensical model for the data, as it implicitly assumes that a set of non-integer values comes from a Poisson distribution. However, I will show firstly that this is precisely what MAXENT does and later that this can be motivated as a PPM, which can be fitted for a non-integer response using the result of Berman & Turner (1992).

Theorem 5.1. *The MAXENT procedure and log-linear Poisson regression are equivalent. That is,*

1. They fit the same model:

$$\ln\pi(g_i) = \ln\mu(g_i) = \mathbf{x}(g_i)'\boldsymbol{\beta}.$$

2. They estimate parameters to maximise the same function up to a constant:

$$\Lambda\{\boldsymbol{\beta};\mathbf{z}^{(n)}(\mathbf{g})\} = l\{\boldsymbol{\beta};\mathbf{z}^{(n)}(\mathbf{g})\} + C,$$

where C is a constant and $\Lambda\{\boldsymbol{\beta};\mathbf{z}^{(n)}(\mathbf{g})\}$ is the Lagrangian function to maximise entropy $H\{\boldsymbol{\pi}(\mathbf{g})\}$ subject to the constraints stated in equations (5.1-5.2) above. Hence the maximum entropy estimate $\widehat{\boldsymbol{\beta}}_{MAXENT}$ equals the maximum likelihood estimate from Poisson regression $\widehat{\boldsymbol{\beta}}_{GLM}$.

Proof. 1. I maximise the entropy $H\{\boldsymbol{\pi}(\mathbf{g})\} = -\sum_{i=1}^{n}\pi(g_i)\,\ln\pi(g_i)$ subject to the given constraints by rephrasing the task as a minimisation problem and constructing the Lagrangian function with multipliers $\boldsymbol{\gamma} = \{\gamma_0, \gamma_1, \ldots, \gamma_p\}$:

$$\Lambda_\lambda\{\boldsymbol{\pi}(\mathbf{g}),\boldsymbol{\gamma};\mathbf{z}^{(n)}(\mathbf{g})\} =$$
$$\sum_{i=1}^{n}\pi(g_i)\,\ln\pi(g_i) + \sum_{j=1}^{p}\gamma_j\left\{\sum_{i=1}^{n}\pi(g_i)\,x_j(g_i) - \overline{x}_j(\mathbf{g})\right\} + \gamma_0\left\{\sum_{i=1}^{n}\pi(g_i) - 1\right\}. \tag{5.5}$$

Differentiating $\Lambda_\lambda\{\boldsymbol{\pi}(\mathbf{g}),\boldsymbol{\gamma};\mathbf{z}^{(n)}(\mathbf{g})\}$ with respect to $\pi(g_i)$ yields:

$$\frac{\partial\Lambda_\lambda\{\boldsymbol{\pi}(\mathbf{g}),\boldsymbol{\gamma};\mathbf{z}^{(n)}(\mathbf{g})\}}{\partial\pi(g_i)} = \frac{\pi(g_i)}{\pi(g_i)} + \ln\pi(g_i) + \sum_{j=1}^{p}\gamma_j x_j(g_i) + \gamma_0.$$

Setting the right hand side to zero and solving for $\pi(g_i)$ leads to the following expression for $\ln \hat{\pi}(g_i)$:

$$\ln \hat{\pi}(g_i) \quad = \quad -1 - \sum_{j=1}^{p} \gamma_j x_j(g_i) - \gamma_0. \tag{5.6}$$

This solution can easily be shown to be the global minimiser of $\Lambda_\lambda\{\boldsymbol{\pi}(\mathbf{g}), \boldsymbol{\gamma}; \mathbf{z}^{(n)}(\mathbf{g})\}$.

This model has the same form as the log-linear model used in Poisson regression (5.3), where $\gamma_j = -\beta_j - I(j = 0)$ and $I(\cdot)$ is the indicator function.

2. Plugging (5.6) back into (5.5) yields:

$$\Lambda_\lambda\{\widehat{\boldsymbol{\pi}}(\mathbf{g}), \boldsymbol{\gamma}; \mathbf{z}^{(n)}(\mathbf{g})\} \quad = \quad -\sum_{i=1}^{n} \hat{\pi}(g_i) - \sum_{j=1}^{p} \gamma_j \overline{x}_j(\mathbf{g}) - \gamma_0.$$

Because $\overline{x}_j(\mathbf{g}) = \sum_{i=1}^{n} z^{(n)}(g_i)\, x_j(g_i)$ and $\gamma_0 = \sum_{i=1}^{n} z^{(n)}(g_i)\, \gamma_0$,

$$\begin{aligned}
\Lambda_\lambda\{\widehat{\boldsymbol{\pi}}(\mathbf{g}), \boldsymbol{\gamma}; \mathbf{z}^{(n)}(\mathbf{g})\} \quad &= \quad -\sum_{i=1}^{n} \hat{\pi}(g_i) + \sum_{i=1}^{n} z^{(n)}(g_i) \left\{ -\sum_{j=1}^{p} \gamma_j x_j(g_i) - \gamma_0 \right\}, \\
&= \quad -\sum_{i=1}^{n} \hat{\pi}(g_i) + \sum_{i=1}^{n} z^{(n)}(g_i)\, \ln \hat{\pi}(g_i) + \sum_{i=1}^{n} z^{(n)}(g_i), \qquad (5.7) \\
&= \quad \Lambda\{\boldsymbol{\beta}; \mathbf{z}^{(n)}(\mathbf{g})\}, \text{say.}
\end{aligned}$$

Since $\gamma_j = -\beta_j - I(j = 0)$, I have reparameterised $\Lambda_\lambda\{\widehat{\boldsymbol{\pi}}, \boldsymbol{\gamma}; \mathbf{z}^{(n)}(\mathbf{g})\}$ as a function of $\boldsymbol{\beta}$ as $\Lambda\{\boldsymbol{\beta}; \mathbf{z}^{(n)}(\mathbf{g})\}$. Note that the expression for $\Lambda\{\boldsymbol{\beta}; \mathbf{z}^{(n)}(\mathbf{g})\}$ in (5.7) differs from the Poisson log-likelihood (5.4) only by a term C that is constant with respect to $\boldsymbol{\beta}$.

Convex duality (Boyd & Vandenberghe, 2004) suggests that the dual function (5.7) provides a lower bound for (5.5). Because entropy is strictly convex, Slater's condition ensures that the solution found by maximising (5.7) is indeed equivalent to the solution of the minimisation problem (5.5). Consequently,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{GLM}} \quad &= \quad \arg\max_{\boldsymbol{\beta}} [l\{\boldsymbol{\beta}; \mathbf{z}^{(n)}(\mathbf{g})\}], \\
&= \quad \arg\max_{\boldsymbol{\beta}} [\Lambda\{\boldsymbol{\beta}; \mathbf{z}^{(n)}(\mathbf{g})\} + C], \\
&= \quad \widehat{\boldsymbol{\beta}}_{\text{MAXENT}},
\end{aligned}$$

where C is a constant with respect to $\boldsymbol{\beta}$. □

Part 1 of Theorem 5.1 (that MAXENT fits a log-linear model) is well-known (e.g. Dutta, 1966), but Part 2 (the link to Poisson regression) is new. This link to Poisson

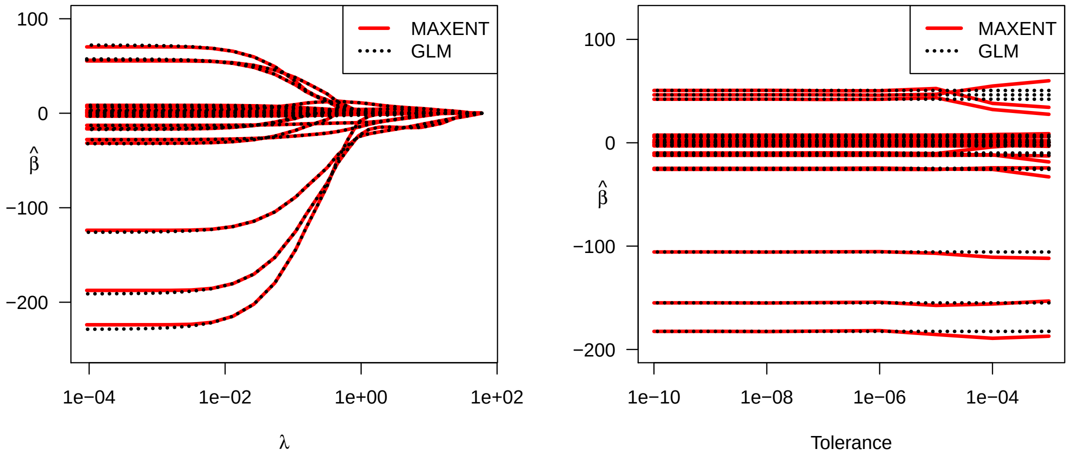

Figure 5.2: Numerical equivalence of MAXENT and GLM. Parameter coefficients as calculated by MAXENT's software (Version 3.3.3e) and using GLM for various values of the LASSO penalty λ (left) and for different tolerance levels (right). Results are numerically equivalent (hence overlapping) for most λ and tolerance levels. Coefficients calculated by the GLM code converged at a higher tolerance, and could be computed much faster (Table 5.2).

likelihood was enabled by specifying the MAXENT model in a slightly different way to what is conventional in the maximum entropy literature. It is typical to exclude the intercept term from the model and introduce a normalisation constant in its place after optimisation to ensure that the sum of $\boldsymbol{\pi}$ is one. Instead, I included an intercept term and the constraint given by (5.2) in the optimisation problem, which was key to the derivation. Hence I have shown that some maximum entropy problems, including MAXENT, can be solved using *standard generalised linear modelling software* via Poisson regression. I demonstrate this result numerically in Figure 5.2. Further, this enables a link with Poisson PPMs below.

Recall from Chapter 3 that a Poisson PPM analyses m presence-only locations $\mathbf{y}_P = \{y_1, \ldots, y_m\}$ as a point process in which the locations of the m points are assumed to be independent. Unlike MAXENT, which only models probability $\pi(g_i)$

per grid cell, a Poisson PPM directly models the intensity $\mu(y)$ *per unit area* (Cressie, 1993) for any location $y \in \mathcal{A}$. As in Chapter 3, intensity is modelled as a log-linear function of p explanatory variables: $\ln\{\mu(y)\} = \mathbf{x}(y)'\boldsymbol{\beta}$. Analysis on a per area basis rather than a per grid cell basis is a key distinction between a PPM and MAXENT.

Recall the log-likelihood of a Poisson PPM (Equation 3.7) can be approximated as a weighted Poisson likelihood (Equation 3.10). As in Chapter 3, quadrature points can be chosen by dividing the region $\mathcal{A}$ into a regular grid and inserting a quadrature point at the centre of each cell, meaning that $\mathbf{y}_0 = \mathbf{g}_0$.

I find a relation in Theorem 5.2 below between MAXENT and the above formulation for Poisson PPMs by analysing data at grid cell locations $\{\mathbf{g}_P, \mathbf{g}_0\}$ instead of $\{\mathbf{y}_P, \mathbf{y}_0\}$. That is, I use in analysis the same quadrature points $\mathbf{y}_0 = \mathbf{g}_0$, but use the locations of the $m^{(n)}$ presence grid cells $\mathbf{g}_P$ in place of the m actual presence locations in $\mathbf{y}_P$. This results in some loss of information, discussed in Section 5.3.

Theorem 5.2. *Consider a Poisson PPM fitted to grid cell data $\mathbf{z}^{(n)}(\mathbf{g})$, with parameter estimates fitted by maximum likelihood stored in $\widehat{\boldsymbol{\beta}}_{PPM}$. Then:*

$$\widehat{\boldsymbol{\beta}}_{MAXENT} = \widehat{\boldsymbol{\beta}}_{PPM} + J_C,$$

where $J_C = \{\ln C, 0, \dots, 0\}$ is a vector of length $p+1$, and $C = |\mathcal{A}|/(m^{(n)}n)$.

In other words, the MAXENT and PPM solutions for grid cell data are proportional, and estimates of slope parameters are identical.

Proof. $\widehat{\boldsymbol{\beta}}_{\text{MAXENT}}$ solves $\frac{\partial\Lambda\{\boldsymbol{\beta};\mathbf{z}^{(n)}(\mathbf{g})\}}{\partial\beta_j} = 0$:

$$\begin{aligned}\frac{\partial\Lambda\{\boldsymbol{\beta};\mathbf{z}^{(n)}(\mathbf{g})\}}{\partial\beta_j} &= \sum_{i=1}^{n}\left\{\frac{I(i \in \{1,\dots,m^{(n)}\})}{m^{(n)}\pi(g_i)}x_j(g_i)\,\pi(g_i) - x_j(g_i)\,\pi(g_i)\right\},\\ 0 &= \frac{1}{m^{(n)}}\sum_{i=1}^{n}x_j(g_i)\left\{I(i \in \{1,\dots,m^{(n)}\}) - m^{(n)}\hat{\pi}(g_i)\right\}. \qquad (5.8)\end{aligned}$$

$\widehat{\boldsymbol{\beta}}_{\text{PPM}}$ solves $\frac{\partial l_{\text{PPM}}(\boldsymbol{\beta};\mathbf{y}_P,\mathbf{y}_0,\mathbf{w})}{\partial\beta_j} = 0$:

$$\frac{\partial l_{\text{PPM}}(\boldsymbol{\beta};\mathbf{y}_P,\mathbf{y}_0,\mathbf{w})}{\partial\beta_j} = \sum_{i=1}^{n}w_i\left\{\frac{z_{w,i}\mu(y_i)\,x_j(y_i)}{\mu(y_i)} - \mu(y_i)\,x_j(y_i)\right\}.$$

For grid cell data, $w_i = |\mathcal{A}|/n$, so $\widehat{\boldsymbol{\beta}}_{\text{PPM}}$ solves:

$$0 = \sum_{i=1}^{n} x_j(g_i) \left\{ I(i \in \{1, \ldots, m^{(n)}\}) - \frac{|\mathcal{A}|}{n} \hat{\mu}(g_i) \right\}. \tag{5.9}$$

(5.8) and (5.9) are related by the identity $\hat{\pi}(g_i) = \frac{|\mathcal{A}|}{m^{(n)}n} \hat{\mu}(g_i)$.

Taking logarithms of both sides yields:

$$\mathbf{x}(g_i)' \widehat{\boldsymbol{\beta}}_{\text{MAXENT}} = \mathbf{x}(g_i)' \widehat{\boldsymbol{\beta}}_{\text{PPM}} + \ln |\mathcal{A}| - \ln n - \ln m^{(n)}.$$

Hence $\widehat{\boldsymbol{\beta}}_{\text{MAXENT}} = \widehat{\boldsymbol{\beta}}_{\text{PPM}} + J_C$, where $C = \frac{|\mathcal{A}|}{m^{(n)}n}$. □

Fithian & Hastie (in review) have independently shown that MAXENT and a Poisson PPM determine the same solution, in a preprint available on *ArXiv*.

Corollary 5.3. *For a given presence-only dataset* $\mathbf{y}_P$*, consider a set of vectors of grid cell data constructed at increasingly fine spatial resolutions (*e.g *by recursively partitioning* $\{\mathbf{z}^{(n)}(\mathbf{g}); n = 1, 2, 2^2, 2^3, \ldots\}$*). As* $n \to \infty$*, the MAXENT solution for* $\mathbf{z}^{(n)}(\mathbf{g})$ *becomes proportional to the Poisson PPM solution for* $\mathbf{y}_P$*. That is:*

$$\widehat{\boldsymbol{\beta}}_{MAXENT} - J_C \to \widehat{\boldsymbol{\beta}},$$

where J_C *is defined in Theorem 5.2.*

The proof follows by noting that as $n \to \infty$, the number and location of presence points in $\mathbf{g}_P$ approaches those in $\mathbf{y}_P$ and the quadrature approximation (Equation 3.10) approaches the exact solution (Equation 3.7).

This result is similar to Theorem 3.2 of Warton & Shepherd (2010) who showed that when fitting a Poisson PPM with constant quadrature weights C, ignoring these weights changes the solution by the factor C. MAXENT can be represented as a Poisson PPM ignoring quadrature weights, so a similar result applies. These quadrature weights are the mechanism that ensures that analysis is performed on an area basis instead of a grid cell basis (Warton & Shepherd, 2010). Hence while Poisson PPM and MAXENT solutions are qualitatively identical, analysing data on a grid cell basis instead of an area basis induces scale dependence in MAXENT: as $n \to \infty$, $\pi(g_i) \to 0$. Hence the maps in Figure 5.3 look the same, but only for the Poisson PPMs is the scale unchanged by changing spatial resolution.

5.3 Model Application

I will now demonstrate the application of a PPM to the presence-only locations of *Corymbia eximia* introduced in Chapter 1, illustrating many features currently unavailable to MAXENT. Software for the below analyses including example data is available in the `R` package `ppmlasso`, described in Chapter 8. The analysis will consist of four steps: (1) determine the appropriate spatial resolution for analysis, (2) assess whether a Poisson PPM is appropriate, (3) estimate the LASSO parameter for regularisation, and (4) compare results with a MAXENT model. A LASSO penalty is included because MAXENT applies one by default. I use four environmental variables as in Warton & Shepherd (2010) – minimum and maximum temperature, number of fires since 1943, and annual rainfall. Likelihood of observing a presence point depends not just on the spatial distribution of the species, but also on the spatial distribution of observers, which is strongly affected by site accessibility. Hence I include two variables to measure site accessibility – distance from main roads and distance from urban areas. Intensity of *C. eximia* was modelled as a quadratic function of the six available variables, including interactions between the four environmental variables and between the two accessibility variables (but assuming additivity between environmental and accessibility variables). So long as all six of these variables are independent of variables associated with species detection probability, parameter estimates from a Poisson PPM will be consistently estimated (Dorazio, 2012).

Prior to applying the LASSO to PPMs, variables were standardised to have mean zero and variance one as in Tibshirani (1996), such that the LASSO penalty was ap-

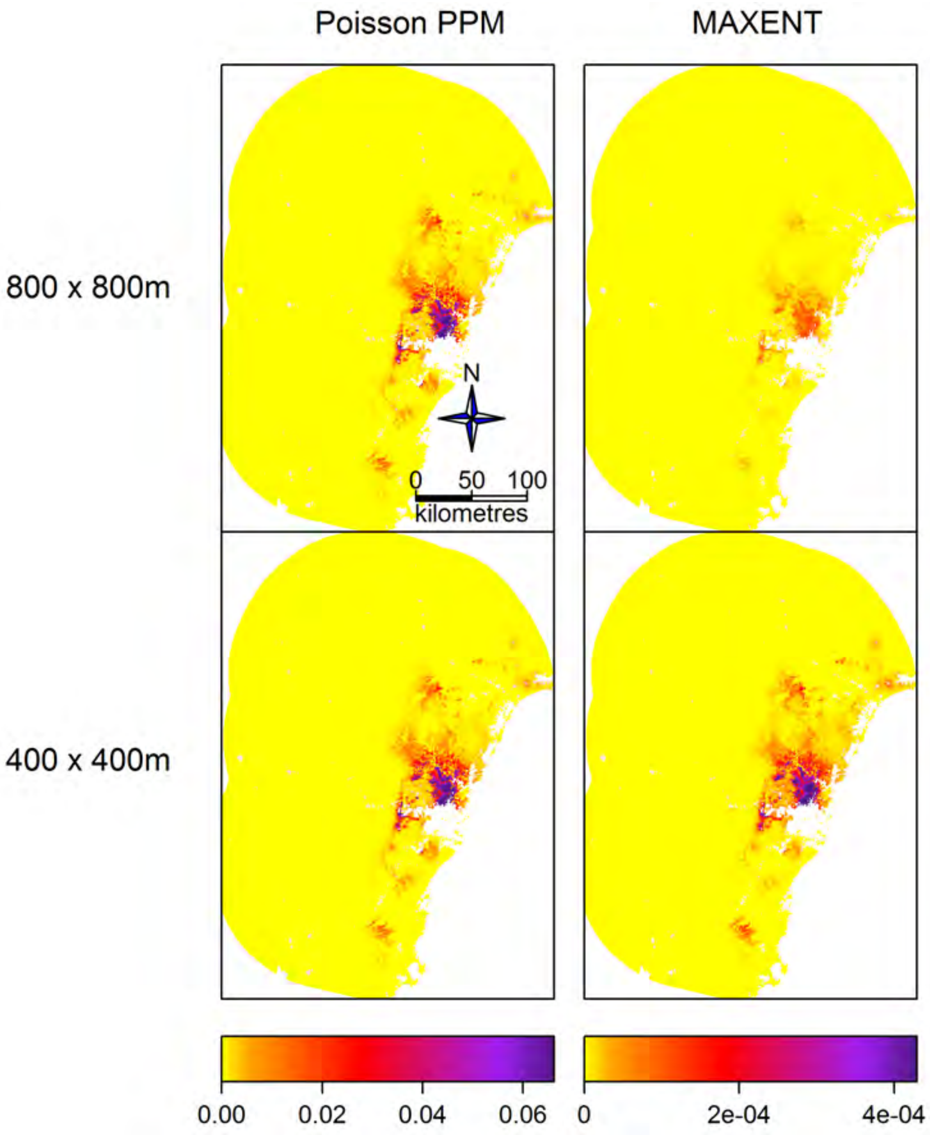

Figure 5.3: Predicted intensity maps. Predicted intensities for *Corymbia eximia* using presence-only data generated by two different methods at the 800m resolution (top) and the 400m resolution (bottom). The maps have the same pattern at each resolution, but the predicted values are scale-dependent for MAXENT while remaining constant for the Poisson PPM.

plied to standardised coefficients. In MAXENT, variables instead were standardised to have minimum zero and maximum one.

5.3.1 Choosing the appropriate spatial resolution

NSW Office of Environment and Heritage (2010) provides environmental data over the study region at the 100m resolution. However, performing an analysis at such a fine resolution is computationally expensive and may not be necessary. Using a Poisson PPM specification facilitates the use of a numerical integration framework for choosing an appropriate spatial resolution for a particular species. Because the absence grid cells $\mathbf{g}_0$ are used as quadrature points, the question of what spatial resolution needs to be used can be rephrased as a question of how many quadrature points are needed to obtain a sufficiently accurate estimate of the log-likelihood (Equation 3.7).

Following Warton & Shepherd (2010), quadrature points are added at increasingly fine resolutions until the log-likelihood has converged. For *Corymbia eximia*, the likelihood appears to converge at a spatial resolution of 800m (Figure 5.4a), suggesting that model output will not appreciably change at finer spatial resolutions. However, the entropy of analogous MAXENT models does not converge due to the scale dependence of $\boldsymbol{\pi}(\mathbf{g})$ and hence MAXENT is not informative about which spatial resolution to use for analysis. The scale dependence of MAXENT can be adjusted for (using "gain", defined as $\ln n-$ entropy) in part, but not completely, since the loss of information incurred by absorbing the m presence locations into a smaller number $m^{(n)}$ of presence grid cells varies with the choice of spatial resolution. Hence the gain will not converge until $m^{(n)}$ converges.

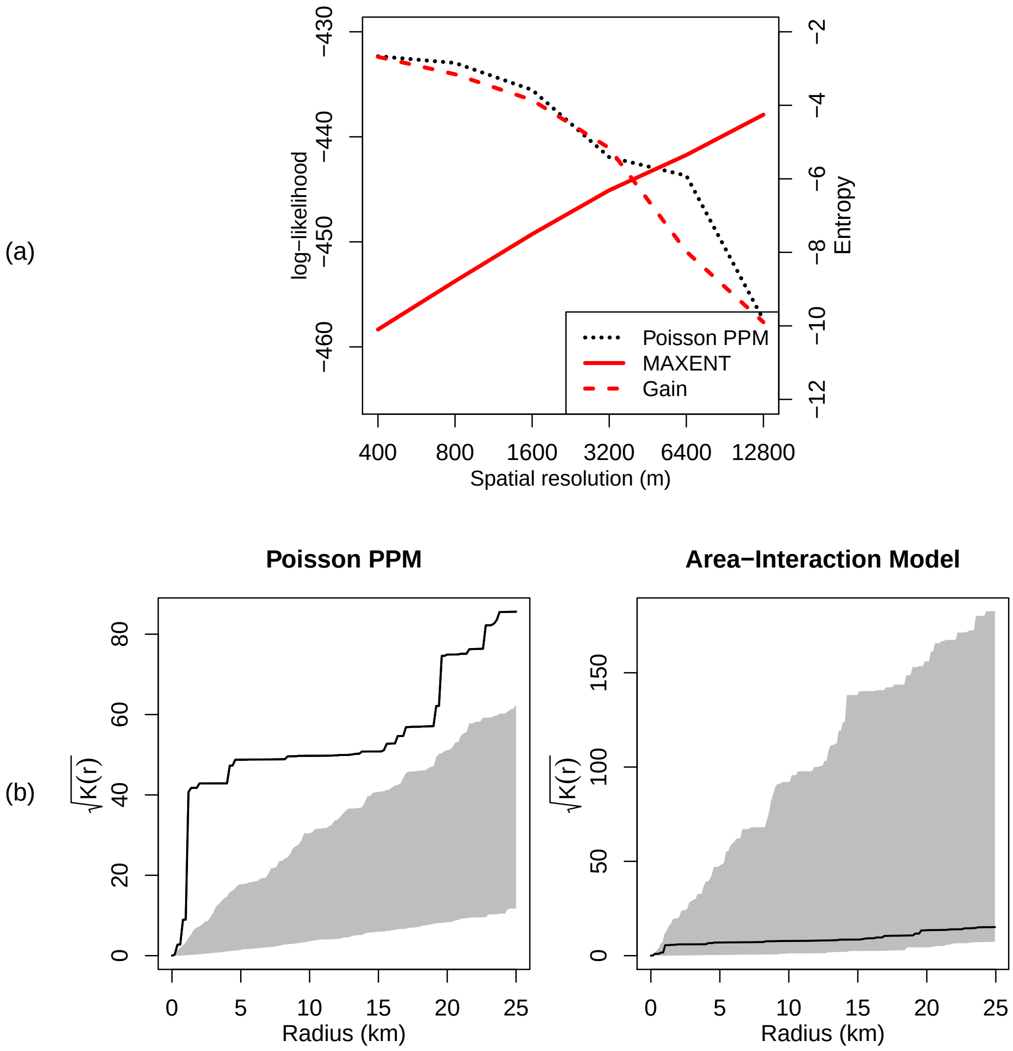

Figure 5.4: Model checking for the *Corymbia eximia* analysis: (a) Spatial resolution can be chosen for a PPM from a plot of maximised log-likelihood at differing spatial resolutions. Convergence is achieved at the 800m resolution for the Poisson PPM, suggesting this is the optimal spatial resolution at which to perform analysis. There is no convergence for the entropy used by MAXENT. We can attempt to address this by analysing “gain” (defined as $\ln n$ - entropy), but gain (rescaled) does not converge until the number of presence cells $m^{(n)}$ converges. (b) Inhomogeneous K-function, with simulation envelope, for a Poisson PPM (left) and an area-interaction model with radius 5 km (right). The deviation from the envelope for the Poisson PPM suggests additional clustering unaccounted for in the model.

5.3.2 Is a Poisson PPM appropriate?

An underlying assumption of a Poisson PPM (and by equivalence, MAXENT) is that the point locations are independent, conditional on model covariates (Table 3.1). This may not be appropriate for *Corymbia eximia*. While MAXENT offers no method for checking this assumption, the diagnostic tools of Chapter 3 (Cressie, 1993; Baddeley *et al.*, 2005) may be applied to assess adequacy of a Poisson PPM. In Figure 5.4b, it can be seen that for *C. eximia*, a Poisson PPM may not be suitable for the data, as the observed estimate of the inhomogeneous K-function $\hat{K}(r)$ falls well outside a 95% envelope formed by simulating 1000 realisations from a Poisson PPM with intensity function as estimated from the *C. eximia* data. As in Chapter 3, the deviation above the envelope suggests that the presence locations of *C. eximia* are more clustered than would be expected for a true Poisson PPM. Instead, Figure 5.4b demonstrates that an area-interaction model with radius 5 km is more appropriate.

5.3.3 Choosing the LASSO parameter

MAXENT is often fitted using a LASSO penalty (Chapter 4) to control for overfitting. However, rather than using the data-driven approaches of Chapter 4 to choose the value of the LASSO parameter λ, MAXENT software used an *ad hoc* value of $\frac{9}{70}$ for *Corymbia eximia*, which was chosen without any consideration for predictive performance of the model at hand but rather based entirely on the number of presence cells (90), as per (Phillips & Dudík, 2008). As an alternative to the default MAXENT penalty, I used a simple line search algorithm to find the value that minimised non-linear GCV (Fu, 2005), which returned a value of 4.907.

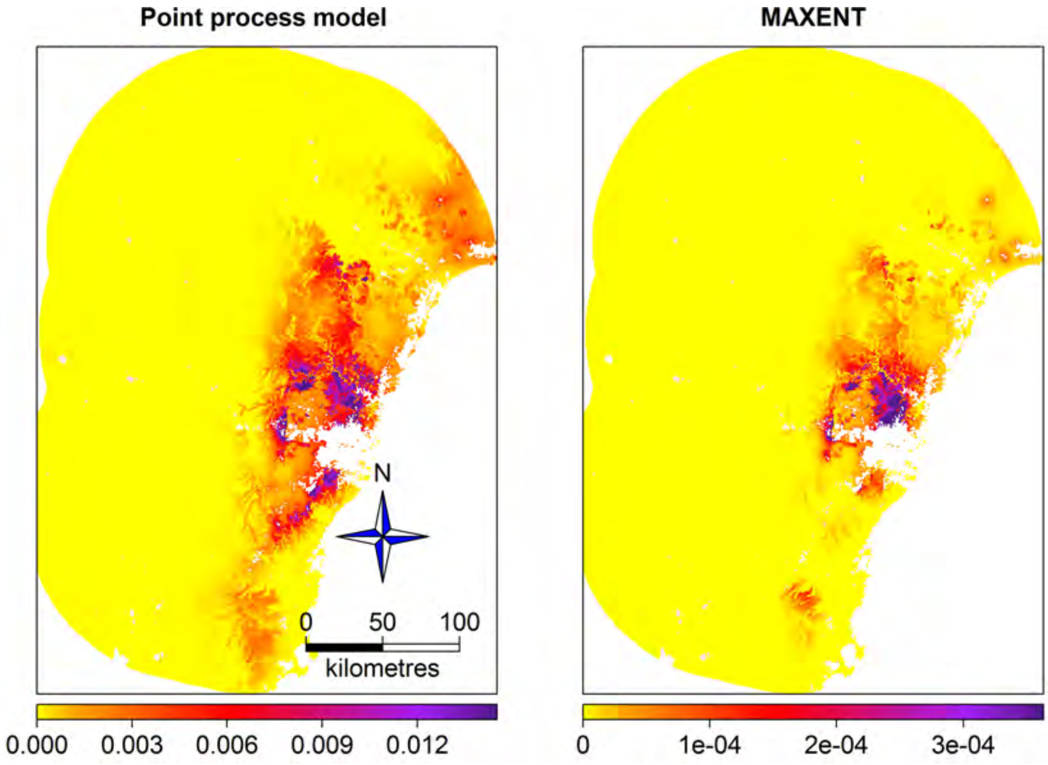

Figure 5.5: Predicted species distribution maps for an area-interaction model (left) and MAXENT (right).

5.3.4 Results

The coefficients for both the PPM and the MAXENT model (Table 5.1) are qualitatively different due largely to the different LASSO parameters. Of the 19 model coefficients, only 11 are non-zero in the PPM, as opposed to 17 for MAXENT. Moreover, the harsher LASSO penalty of the PPM ensures that each of the estimated coefficients are smaller than the corresponding coefficients of the MAXENT model. Otherwise, the models are broadly similar and hence the maps produced by both models identify the same geographic hot spots for *Corymbia eximia* (Figure 5.5).

5.3.5 Summary

Analysing the *Corymbia eximia* data has illustrated advantages of the Poisson PPM approach in choosing the spatial resolution, assessing model adequacy, and choosing

Table 5.1: Model coefficients for a PPM (area-interaction with radius 5 km) and MAXENT model. The variables included are number of fires since 1943 (FC), minimum and maximum annual temperature (MNT and MXT), annual rainfall (Rain), distance from main roads (D.Main) and distance from urban areas (D.Urb). Both models were built at the 800m spatial resolution. The LASSO parameters for each model are 4.907 and 9/70, respectively.

Coefficient	PPM	MAXENT model
Intercept	-10.545	56.457
FC	0.275	2.898
FC^2	-0.321	-6.789
MNT	1.268	8.386
FC∗MNT	0.200	8.399
MNT^2	-1.158	-24.593
MXT	0	34.394
FC∗MXT	0.322	5.897
MNT∗MXT	-0.077	15.140
MXT^2	-1.281	-73.619
Rain	0	0
FC∗Rain	-0.035	-14.104
MNT∗Rain	0.539	27.071
MXT∗Rain	0	-114.774
$Rain^2$	-0.824	-105.588
D.Main	-0.469	-3.511
$D.Main^2$	-0.118	-2.538
D.Urb	-0.547	-5.310
D.Main∗D.Urb	0	0.495
$D.Urb^2$	0	0
Point Interaction	0.176	NA

Table 5.2: Current problems with MAXENT and their proposed solutions available through reexpression as a Poisson PPM.

MAXENT problem	**Poisson PPM solution**
Predicted probabilities are scale-dependent	Predicted intensities are scale-invariant
How to determine spatial resolution?	Increase until log-likelihood converges
How to assess model adequacy?	Goodness-of-fit procedures (Chapter 3)
How to choose LASSO parameter?	Data-driven methods (Chapter 7)
Available in MAXENT software only	Standard GLM software (Chapter 8)
130 seconds to fit models in Figure 5.5	12 seconds to fit models in Figure 5.5

the LASSO parameter. These are summarised in Table 5.2. Another potential advantage is in assessing model uncertainty – a point process framework can be used to put standard errors on model coefficients and predictions, although when using the LASSO in estimation (Fan & Li, 2005) there are some difficulties (Kyung *et al.*, 2010). A final advantage worthy of mention is in computation time: Figure 5.5 took 12 seconds to produce for the PPM, but 130 seconds using MAXENT software (Table 5.2).

5.4 Improvements in Predictive Performance

I will now compare the predictive performance of the point process approach described in Section 5.3 to MAXENT in order to assess whether the performance of the model has been improved by the proposed refinements (in particular, modelling point interactions and data-driven estimation of the LASSO penalty parameter). The approach I took was to model *Corymbia eximia* presence-only data and predict to new areas, assessing predictive performance using a separate presence-absence dataset from 8678 systematically collected transects (NSW Office of Environment and Heritage, 2010), as in Elith *et al.* (2006). This presence-absence dataset may

Table 5.3: Predictive performance (measured as average area under the ROC curve for 20 different 5-fold spatial cross-validation schemes) of different presence-only models for *C. eximia* when predicting to a separate presence-absence dataset. Note that the point process approach proposed in Section 5.3 has the highest predictive performance.

Model	**LASSO Penalty Criteria**	**AUC**	**Standard Error**
Poisson PPM	No penalty	0.7555	0.0070
MAXENT	*ad hoc* MAXENT	0.8508	0.0060
Poisson PPM	Non-Linear GCV	0.8813	0.0051
Area-interaction	Non-Linear GCV	0.9066	0.0036

be considered a "gold standard", where observers have gone to each of the 8678 sites and specifically noted presences of *C. eximia*. I applied a spatial 5-fold cross-validation in which sites were assigned to 30 square 64 × 64 km spatial blocks that were randomly assigned to test and training samples. I employed this procedure to minimise the influence of spatial autocorrelation, which was not considered by MAXENT.

I evaluated the performance of MAXENT and various models from the point process approach by comparing predicted intensities at the systematically collected transects against observed presence/absence, using area under a ROC curve (Hastie *et al.*, 2009). Table 5.3 reveals that choosing the LASSO parameter to minimise the non-linear GCV performed better than using MAXENT's default method for *C. eximia* for both PPMs. Hence, while MAXENT achieves high predictive performance relative to other SDM methods (Elith *et al.*, 2006), there is the potential to improve it further by using the data to inform the choice of the LASSO parameter.

5.5 Discussion

Some recent papers (Elith & Leathwick, 2009; Aarts *et al.*, 2012) have called for greater unification and synthesis of the literature on SDM. To that end, I have demonstrated equivalence of MAXENT and a Poisson PPM. Warton & Shepherd (2010) showed the equivalence of Poisson PPMs and pseudo-absence regression, which aside from MAXENT is the most commonly used approach to presence-only modelling at the moment. Hence this work represents a significant unification of the literature, using Poisson PPMs to link the two most widely used presence-only methods, MAXENT and pseudo-absence regression. This work has significant practical ramifications, given that MAXENT (Table 5.2) and pseudo-absence regression (Warton & Shepherd, 2010) have shortcomings stemming largely from the framework used for modelling, which can be resolved by using a Poisson PPM instead. Others have made further connections between PPMs and alternative approaches to analysis – Aarts *et al.* (2012) and Baddeley *et al.* (2012) made a connection to the estimation of "resource selection functions" via presence-absence analysis, and Dorazio (2012) to case-augmented binary regression. PPMs are a natural framework for analysing presence-only data and it is interesting that a variety of different methods of analysis can all be connected to them in some way, and in many instances, improved through this connection.

A key distinction between PPMs and MAXENT is that in the former $\mu(y)$ is modelled on a per area basis whereas for the latter, $\pi(g_i)$ is modelled per grid cell – the per area analysis is thus invariant under choice of spatial resolution while the per grid cell analysis is not (because increasing spatial resolution increases the number of grid cells). This is related to the distinction between probability and frequency models (Aarts *et al.*, 2012). It is this distinction that enables the likelihood convergence for a Poisson PPM (Figure 5.4a) and hence a data-driven choice of spatial resolution. However, MAXENT is proportional to a Poisson PPM (Theorem 5.2), which suggests that it can achieve the same qualitative answer but with the disadvantage of scale dependence of the predicted probabilities and an arbitrary choice of spatial resolution.

One important disadvantage of MAXENT is that in its current form, it does not estimate the intercept consistently (Elith *et al.*, 2011). The intercept term diverges to $-\infty$ as spatial resolution increases. Theorem 5.2 gives the form of the term causing this divergence. Such discrepancies between models fitted to grid cell data at different spatial resolutions have been described extensively for logistic regression in Baddeley *et al.* (2010). This means that MAXENT as currently posed cannot predict species intensity for any subset of the study region $\mathcal{A}$ in the way that PPMs can.

The ability to use data to estimate spatial resolution (Figure 5.4a) is of interest for a couple of reasons. First, the resolution of the process is largely a function of biological factors and measurement error, and estimating this resolution is informative about the spatial scale at which such processes are operating. Second, the resolution of the process is of interest for computational reasons, because data are becoming available at increasingly fine resolutions - I originally had access to 8,620,092 points at the 100m resolution, but even finer resolutions are now available - and analysis at such fine resolutions can be very computationally intensive. Colleagues analysing this type of data in biology departments have constructed their own parallel computing arrays to analyse this type of data for multiple species at fine resolutions. Hence it is of considerable practical interest to know whether such a fine resolution is required, and in this case, it clearly was not required as I only needed 134,716 quadrature points and was able to analyse data in seconds on a desktop computer (Table 5.2), with negligible loss of information.

An alternative approach to using all grid cells in MAXENT analysis is to randomly select empty grid cells as “background points” for analysis, as in Bonham-Carter (1994). This obviates any computational need to coarsen resolution for analysis. The default approach that has been advocated (Phillips & Dudík, 2008) and implemented in MAXENT software is to use 10,000 random background points, which for my data was clearly insufficient (Figure 5.4), equivalent to using a resolution of nearly 3 km. I advise that as a matter of routine, presence-only analysts should use their data to identify a spatial resolution appropriate for analysis, or equivalently, to identify the number of "background points" to use in analysis.

In Section 5.4 I demonstrated that PPMs achieve a higher predictive performance for *Corymbia eximia* by choosing the LASSO penalty parameter to minimise non-linear GCV. However, this may not be true of all species. In Chapter 7, I investigate the question of how predictive performance and variable selection varies with different methods of choosing the LASSO parameter across multiple species using two real data sets and simulation.

Chapter 6

LASSO Asymptotics for Poisson Point Process Models

6.1 Introduction

In Chapter 5, I established the equivalence of MAXENT and Poisson PPMs. Therefore, any advantage in predictive performance inherent in MAXENT is not due to a fundamental difference with GLM methods. One area in which the two approaches differ is in the incorporation of a LASSO penalty. Applying a LASSO penalty shrinks parameter estimates toward zero, thereby reducing their variance. A judicious choice of the LASSO penalty can therefore improve predictive ability. Given that applying a LASSO penalty controls the complexity of the model and impacts its predictive performance, it is clearly of interest to explore how to choose the penalty.

The LASSO penalty applied by MAXENT is a function of the number of presences. However, the precise function is *ad hoc*, so an alternative may have superior properties.

Knight & Fu (2000) examined the behaviour of LASSO-type estimators for linear models of sample size n and established that applying a LASSO penalty of order $O(\sqrt{n})$ results in parameter estimates that are $\sqrt{n}$-consistent. In this Chapter I extend this result to the setting of Poisson PPMs, which to my knowledge has not been investigated previously. This is not a trivial exercise because while in a typical study of asymptotic behaviour there is a fixed sample size n that is allowed to grow large, there is no fixed n in a Poisson PPM. Instead, the natural measure is the number of presences m, but it is a random quantity. I resolve this issue by exploring the asymptotic behaviour of the LASSO estimator as the *expected* number of presences $\mu_{\mathcal{A}}$ approaches infinity. In Section 6.3 I show that for a set of m presence-only locations $\mathbf{y}_P$, applying a LASSO penalty of order $O(\sqrt{m})$ will yield parameter estimates that are $\sqrt{m}$-consistent for a Poisson PPM. It is a slight abuse of notation to discuss $\sqrt{m}$-consistency, because as above m is a random quantity and it is $\mu_{\mathcal{A}}$ which is sent to ∞ via the intercept. However, $m/\mu_{\mathcal{A}} \to_P 1$ so asymptotic results for m and $\mu_{\mathcal{A}}$ are interchangeable.

The $\sqrt{m}$-consistent result of this Chapter provides a means of choosing λ in such a way that $\widehat{\boldsymbol{\beta}}$ has desirable properties. The motivation for the focus on $\sqrt{m}$-consistency is that this ensures that the tradeoff between bias and variance in parameter estimates is managed in such a way that both vanish at the same rate asymptotically.

I begin with an asymptotic result for Poisson GLMs in Section 6.2, which serves as an intermediate step to the asymptotic result for Poisson PPMs established in Section 6.3. The result of Section 6.3 motivates a new criterion for choosing the LASSO penalty in Section 6.4 that is $\sqrt{m}$-consistent in estimation. A discussion follows in Section 6.5.

6.2 Asymptotic Behaviour of Poisson GLM LASSO

In this Section I will provide asymptotic results for a Poisson GLM as the sample size n goes to infinity. This is a straightforward extension of the results of Knight & Fu (2000), who established a result for linear models. This GLM result will provide a framework to be used for examining the asymptotic behaviour of Poisson PPMs in

Section 6.3. This result is very similar to the asymptotic result of adaptive LASSO fitted to GLMs given in Zou (2006).

In this Section I assume that $\mathbf{z} = \{z_1, \ldots, z_n\}$ is a set of count data which can be modelled as a function of p explanatory variables $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. Consider a Poisson GLM, which has mean function:

$$E(z_i) = \mu_i = e^{\mathbf{x}_i \boldsymbol{\beta}}.$$

Hence the mean μ_i can be modelled as a loglinear function of environmental covariates $\mathbf{x}_i$:

$$\ln \mu_i = \mathbf{x}_i \boldsymbol{\beta}.$$

This model can be fitted with a LASSO penalty by maximising the constrained likelihood:

$$l(\boldsymbol{\beta}; \mathbf{z}) = -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} z_i \ln \mu_i - \lambda_n \sum_{j=1}^{p} |\beta_j|. \tag{6.1}$$

I denote the LASSO penalty parameter by λ_n in this Section to emphasise that it will be chosen as a function of n. Denote by $\widehat{\boldsymbol{\beta}}_n$ the vector of parameter values that maximises (6.1) and $\boldsymbol{\beta}_*$ the vector of true parameter values.

To investigate the consistency of the LASSO estimate $\widehat{\boldsymbol{\beta}}_n$, I will assume that $n \to \infty$ in such a way that:

$$\mathbf{C}_n = \frac{1}{n} \sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i' \to \mathbf{C}, \tag{6.2}$$

where $\mathbf{C}$ is a nonnegative definite matrix,

$$\begin{aligned} \frac{1}{n} \max_{1 \le i \le n} \mu_i \mathbf{x}_i' \mathbf{x}_i &\to 0, \quad \text{and} && (6.3) \\ \frac{1}{n} l'''(\boldsymbol{\beta}) &\to_P \mathbf{K} && (6.4) \end{aligned}$$

for $\boldsymbol{\beta}$ in the neighbourhood of $\boldsymbol{\beta}_*$, where the elements of $\mathbf{K}$ are finite.

Conditions (6.2) and (6.3) are similar to those in Knight & Fu (2000) – the only difference here is the inclusion of μ_i. This follows from the natural extension to GLMs, as μ_i emerges from differentiating the log-likelihood (6.1). Condition (6.4), however, is similar to that of Zou (2006) and is necessary for the third derivative term in a Taylor expansion of the likelihood to vanish asymptotically. Note that $\mathbf{C}_n$ is the Fisher information of $\boldsymbol{\beta}_*$.

The following theorem shows that $\sqrt{n}$-consistency is achievable for a Poisson GLM with the right choice of λ_n when conditions (6.2), (6.3), and (6.4) hold.

Theorem 6.1. *Assume that regularity conditions (6.2), (6.3), and (6.4) hold. If* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ *and* $\mathbf{C}$ *is nonsingular, then*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_*) \to_d \operatorname{argmax}(v),$$

where

$$v(\mathbf{u}) = \mathbf{u}'\mathbf{T} - \frac{1}{2}\mathbf{u}'\mathbf{C}\mathbf{u} - \lambda_0 \sum_{j=1}^{p} [u_j sign(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)],$$

$\mathbf{T}$ *has a* $N(\mathbf{0}, \mathbf{C})$ *distribution, and* $\mathbf{u} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_*)$.

Proof:

Let $v_n(\mathbf{u}) = l(\boldsymbol{\beta}_* + \mathbf{u}/\sqrt{n}) - l(\boldsymbol{\beta}_*) - \lambda_n \sum_{j=1}^{p} [|\beta_j + u_j/\sqrt{n}| - |\beta_j|]$. As $l(\boldsymbol{\beta}_*)$ is a constant with respect to $\boldsymbol{\beta}$, $v_n(\mathbf{u})$ will be maximised when $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_n$. Rephrasing the result in terms of $\mathbf{u}$, $v_n(\mathbf{u})$ is maximised when $\mathbf{u} = \widehat{\mathbf{u}} = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_*)$.

To determine the limiting distribution of $v_n(\mathbf{u})$, consider a Taylor expansion of $l(\boldsymbol{\beta})$ about $\boldsymbol{\beta} = \boldsymbol{\beta}_*$:

$$l(\boldsymbol{\beta}) \approx l(\boldsymbol{\beta}_*) + l'(\boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_*)' l''(\boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*) + O_p(n^{-1/2}).$$

The $O_p(n^{-1/2})$ term follows from (6.4). Hence,

$$\begin{aligned} l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}_*) &\approx l'(\boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_*)' l''(\boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*) + O_p(n^{-1/2}) \\ &= l_n^{(1)}(\mathbf{u}) + l_n^{(2)}(\mathbf{u}) + O_p(n^{-1/2}), \end{aligned}$$

where

$$\begin{aligned} l_n^{(1)}(\mathbf{u}) &= \sum_{i=1}^{n} (y_i - \mu_i)\mathbf{x}_i' \frac{\mathbf{u}'}{\sqrt{n}} && (6.5) \\ l_n^{(2)}(\mathbf{u}) &= -\frac{1}{2n}\mathbf{u}' \sum_{i=1}^{n} (\mu_i \mathbf{x}_i \mathbf{x}_i')\mathbf{u}. \end{aligned}$$

From (6.2) and (6.3), the Central Limit Theorem can be applied to (6.5). Hence $l_n^{(1)}(\mathbf{u}) \rightarrow_d \mathbf{u}'\mathbf{T}$.

By (6.2), $l_n^{(2)}(\mathbf{u}) \rightarrow -1/2\mathbf{u}'\mathbf{C}\mathbf{u}$.

As in Knight & Fu (2000),
$\lambda_n \sum_{j=1}^p [|\beta_j + u_j/\sqrt{n}| - |\beta_j|] \rightarrow \lambda_0 \sum_{j=1}^p [u_j \mathrm{sign}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)]$.

Hence by Slutsky's Theorem, $v_n(\mathbf{u}) \rightarrow_d v(\mathbf{u})$. As in Knight & Fu (2000), v_n is convex and v has a unique maximum, so

$$\mathrm{argmax}(v_n) = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_*) \rightarrow_d \mathrm{argmax}(v).$$

6.3 Asymptotic Behaviour of Poisson PPM LASSO

I now focus on the key result of this Chapter, the asymptotic behaviour of a Poisson PPM as the expected number of presence locations $\mu_{\mathcal{A}}$ goes to infinity.

In this Section I assume that $\mathbf{y} = \{\mathbf{y}_P, \mathbf{y}_0\}$ is modelled as a Poisson PPM whose intensity $\mu_i = ae^{\mathbf{x}_i \boldsymbol{\beta}}$ is a log-linear function of environmental covariates $\mathbf{x}_i$ (Equation 3.4) with coefficients stored in $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\}$. Denote by $\boldsymbol{\beta}_*$ the vector of true parameter values.

For the asymptotic result that follows, I assume $a \rightarrow \infty$, which ensures that $\mu_{\mathcal{A}} \rightarrow \infty$ in such a way that $\mu_i \rightarrow \infty$ uniformly while $\boldsymbol{\beta}_*$ remains fixed. Geometrically, the spatial pattern of the intensity surface remains the same but the scale of the intensity increases to ∞ such that realisations of the point process generate an increasingly large number of points $\mathbf{y}_P$ placed in $\mathcal{A}$.

To fit a Poisson PPM with a LASSO penalty, once again the constrained likelihood is maximised:

$$l(\boldsymbol{\beta}; \mathbf{y}_P) = \sum_{i=1}^{m} \ln \mu(y_i) - \mu_{\mathcal{A}} - \lambda_m \sum_{j=1}^{p} |\beta_j|. \tag{6.6}$$

I denote the LASSO penalty parameter by λ_m in this Section to emphasise that it will be chosen as a function of m. Denote by $\widehat{\boldsymbol{\beta}}_m$ the vector of parameter values that maximises (6.6).

To show that $\widehat{\boldsymbol{\beta}}_m$ is $\sqrt{m}$-consistent, I will assume similar regularity conditions to those in Section 6.2:

$$\mathbf{C}_m = \frac{1}{\mu_{\mathcal{A}}} \int_{y \in \mathcal{A}} \mu(y) \mathbf{x}_i \mathbf{x}_i' dy \to \mathbf{C}, \tag{6.7}$$

where $\mathbf{C}$ is a nonnegative definite matrix,

$$\begin{aligned} \frac{1}{\mu_{\mathcal{A}}} \max_{y \in \mathcal{A}} e^{\mathbf{x}(y)\boldsymbol{\beta}} \mathbf{x}(y)' \mathbf{x}(y) &\to 0, \quad \text{and} && (6.8) \\ \frac{1}{m} l'''(\boldsymbol{\beta}) &\to_P \mathbf{K} && (6.9) \end{aligned}$$

for $\boldsymbol{\beta}$ in the neighbourhood of $\boldsymbol{\beta}_*$, where the elements of $\mathbf{K}$ are finite.

The following theorem shows that $\sqrt{m}$-consistency can be achieved with the proper choice of λ_m.

Theorem 6.2. *Assume that regularity conditions (6.7), (6.8), and (6.9) hold. If* $\lambda_m/\sqrt{m} \to \lambda_0 \geq 0$ *and* $\mathbf{C}$ *is nonsingular, then*

$$\sqrt{m}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_*) \to_d \operatorname{argmax}(v),$$

where

$$v(\mathbf{u}) = \mathbf{u}'\mathbf{T} - \frac{1}{2}\mathbf{u}'\mathbf{C}\mathbf{u} - \lambda_0 \sum_{j=1}^{p} [u_j sign(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)],$$

$\mathbf{T}$ *has a* $N(\mathbf{0}, \mathbf{C})$ *distribution, and* $\mathbf{u} = \sqrt{m}(\boldsymbol{\beta} - \boldsymbol{\beta}_*)$.

Proof:

Let $v_m(\mathbf{u}) = l(\boldsymbol{\beta}_* + \mathbf{u}/\sqrt{m}) - l(\boldsymbol{\beta}_*) - \lambda_m \sum_{j=1}^{p} [|\beta_j + u_j/\sqrt{m}| - |\beta_j|]$. From the same argument as given in the proof of Theorem 6.1, $v_m(\mathbf{u})$ is maximised when $\mathbf{u} = \widehat{\mathbf{u}} = \sqrt{m}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_*)$.

To determine the limiting distribution of $v_m(\mathbf{u})$, consider a Taylor expansion of $l(\boldsymbol{\beta})$ about $\boldsymbol{\beta} = \boldsymbol{\beta}_*$:

$$l(\boldsymbol{\beta}) \approx l(\boldsymbol{\beta}_*) + l'(\boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_*)' l''(\boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*) + O_p(m^{-1/2}).$$

The $O_p(m^{-1/2})$ term follows from (6.9). Hence,

$$\begin{aligned} l(\widehat{\boldsymbol{\beta}}_m) - l(\boldsymbol{\beta}_*) &\approx l'(\boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_*)' l''(\boldsymbol{\beta}_*)(\boldsymbol{\beta} - \boldsymbol{\beta}_*) + O_p(m^{-1/2}) \\ &= l_m^{(1)}(\mathbf{u}) + l_m^{(2)}(\mathbf{u}) + O_p(m^{-1/2}), \end{aligned}$$

where

$$\begin{aligned} l_m^{(1)}(\mathbf{u}) &= \frac{\mathbf{u}'}{\sqrt{m}} \left(\sum_{i=1}^{m} \mathbf{x}_i - \int_{y \in \mathcal{A}} \mathbf{x}(y)\mu(y)dy \right) \qquad (6.10) \\ l_m^{(2)}(\mathbf{u}) &= -\frac{1}{2m}\mathbf{u}' \int_{y \in \mathcal{A}} \mu(y)\mathbf{x}(y)'\mathbf{x}(y)dy\, \mathbf{u}. \end{aligned}$$

By (6.7), $l_m^{(2)}(\mathbf{u}) \to -1/2\mathbf{u}'\mathbf{C}\mathbf{u}$.

I will find the limit of (6.10) by splitting the region $\mathcal{A}$ into increasingly fine grid cells and expressing the integral as the limit of a Reimann sum of the volume under the intensity surface in each cell. Let $\mathbf{z} = \{z_1, \ldots, z_n\}$ for n grid cells, where $z_i = I(i \in \{1, \ldots, m\})$. z_i is therefore an indicator of whether cell i is one of the m presence-only locations. Let $w_i = |\mathcal{A}|/n$ be the area of each cell and $a/n \to k$, where $0 < k < \infty$. Letting a and n grow to ∞ at the same rate is the key innovation, as it permits the Central Limit Theorem to be applied as in the proof of Theorem 6.1.

$$\begin{aligned} l_m^{(1)}(\mathbf{u}) &= \frac{\mathbf{u}'}{\sqrt{m}} \left(\sum_{i=1}^{n} \mathbf{x}_i z_i - \lim_{n \to \infty} \sum_{i=1}^{n} \mathbf{x}_i \mu_i w_i \right) \\ &= \frac{\mathbf{u}'}{\sqrt{m}} \left(\lim_{n \to \infty} \sum_{i=1}^{n} \mathbf{x}_i (z_i - \mu_i w_i) \right) \\ &= \frac{\mathbf{u}'}{\sqrt{m}} \left(\lim_{n \to \infty} \sum_{i=1}^{n} \mathbf{x}_i (z_i - \frac{a}{n}|\mathcal{A}| e^{\mathbf{x}_i \boldsymbol{\beta}}) \right) \\ &= \frac{\mathbf{u}'}{\sqrt{\mu_{\mathcal{A}}}} \frac{\sqrt{\mu_{\mathcal{A}}}}{\sqrt{m}} \left(\lim_{n \to \infty} \sum_{i=1}^{n} \mathbf{x}_i (z_i - \frac{a}{n}|\mathcal{A}| e^{\mathbf{x}_i \boldsymbol{\beta}}) \right) \\ &\to_P \frac{\mathbf{u}'}{\sqrt{\mu_{\mathcal{A}}}} \left(\lim_{n \to \infty} \sum_{i=1}^{n} \mathbf{x}_i (z_i - k|\mathcal{A}| e^{\mathbf{x}_i \boldsymbol{\beta}}) \right) \\ &\to_d \mathbf{u}'\mathbf{T}. \qquad (6.11) \end{aligned}$$

(6.11) follows from the Central Limit Theorem. The z_i in different grid cells are independent with mean $\mu_i w_i$. Hence by the Central Limit Theorem for weighted sums (Fisher, 1992),

$$\begin{aligned}
\frac{1}{\sqrt{\sum_{i=1}^{n} \mu_i w_i \mathbf{x}_i \mathbf{x}_i'}} \sum_{i=1}^{n} \mathbf{x}_i \left(z_i - \mu_i w_i\right) &\to_d \quad N(\mathbf{0}, \mathbf{I}) \\
\frac{1}{\sqrt{\mu_{\mathcal{A}}}} \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{x}_i \left(z_i - \mu_i w_i\right) &\to_d \quad \mathbf{T}.
\end{aligned}$$

As in Knight & Fu (2000),
$\lambda_m \sum_{j=1}^{p}[|\beta_j + u_j/\sqrt{m}| - |\beta_j|] \to \lambda_0 \sum_{j=1}^{p}[u_j \text{sign}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)]$.

Hence by Slutsky's Theorem, $v_m(\mathbf{u}) \to_d v(\mathbf{u})$. As in Knight & Fu (2000), v_m is convex and v has a unique maximum, so

$$\text{argmax}(v_m) = \sqrt{m}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_*) \to_d \text{argmax}(v).$$

6.4 New method of choosing the LASSO penalty: MSI

The asymptotic result in Theorem 6.2 provides a guideline for developing LASSO penalty criteria that are $\sqrt{m}$-consistent. Hence I propose the MSI penalty for choosing the LASSO penalty, which is based on the smallest penalty that will fully shrink to the intercept model, $\lambda_{\max}$:

$$\lambda_{\text{MSI}} = \frac{\lambda_{\max}}{\sqrt{m}}. \tag{6.12}$$

The name MSI is proposed since $\lambda_{\max}$ is equal to the **M**aximum value of the **S**core functions of the **I**ntercept model. One noteworthy quality of the MSI penalty is that when $m = 1$, λ_{MSI} will shrink all coefficients to zero. This is appealing because there is insufficient information to construct a reasonable SDM with only one presence location.

I will now show that $\lambda_{\text{MSI}} = O(\sqrt{m})$.

The score functions are found by differentiating the unconstrained log-likelihood (Equation 3.7) with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} s(\boldsymbol{\beta};\mathbf{y}_P) &= \frac{\partial l(\boldsymbol{\beta};\mathbf{y}_P)}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{m} \frac{1}{\mu_i}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} - \int_{y\in\mathcal{A}} \frac{\partial \mu(y)}{\partial \boldsymbol{\beta}} dy \\ &= \sum_{i=1}^{m} \frac{1}{\mu_i}\mu_i \mathbf{x}_i - \int_{y\in\mathcal{A}} \mu(y)\mathbf{x}(y) dy \\ &= \sum_{i=1}^{m} \mathbf{x}_i - \int_{y\in\mathcal{A}} \mu(y)\mathbf{x}(y) dy. \end{aligned}$$

Let $\widehat{\boldsymbol{\beta}}_0 = (\beta_0, 0, \ldots, 0)$ be the parameters of the intercept model and $\widehat{\boldsymbol{\mu}}_0 = m/|\mathcal{A}|$ be the corresponding intensity. Let $\boldsymbol{\beta}_*$ and $\boldsymbol{\mu}_*$ be the true parameter values and intensity surface, respectively. Then, the score functions of the intercept model are:

$$\begin{aligned} s(\widehat{\boldsymbol{\beta}}_0;\mathbf{y}_P) &= \sum_{i=1}^{m} \mathbf{x}_i - \int_{y\in\mathcal{A}} \hat{\mu}_0(y)\mathbf{x}(y) dy \\ &\rightarrow_d \int_{y\in\mathcal{A}} \mu_*(y)\mathbf{x}(y) dy - \int_{y\in\mathcal{A}} \mu_0(y)\mathbf{x}(y) dy && (6.13) \\ &= \mu_{\mathcal{A}} \int_{y\in\mathcal{A}} \mathbf{x}(y) \left[\frac{\mu_*(y)}{\mu_{\mathcal{A}}} - \frac{\mu_0(y)}{\mu_{\mathcal{A}}} \right] dy, && (6.14) \end{aligned}$$

where $\mu_0 = \mu_{\mathcal{A}}/|\mathcal{A}|$. (6.13) follows from the GNZ formula (Equation 3.21). Because all terms in the integrand of (6.14) are constant with respect to a, the integral has order $O(1)$.

Consequently, the expression in (6.14) has order $O(m)$, and:

$$\begin{aligned} \lambda_{\text{MSI}} &= \frac{\max_j s(\widehat{\boldsymbol{\beta}}_0;\mathbf{y}_P)}{\sqrt{m}} && (6.15) \\ &= O(\sqrt{m}). \end{aligned}$$

6.5 Discussion

In this Chapter, I have established an asymptotic result for choosing the LASSO penalty for Poisson PPMs. While Theorem 6.2 establishes that choosing $\lambda_m =$

$O(\sqrt{m})$ leads to a $\widehat{\boldsymbol{\beta}}_m$ that is consistent in estimation, it does not guarantee that $\widehat{\boldsymbol{\beta}}_m$ is consistent in variable selection. In order for $\widehat{\boldsymbol{\beta}}_m$ to be consistent in variable selection, a further restriction known as the Strong Irrepresentable Condition (Zhao & Yu, 2006) must be imposed on $\mathbf{C}_m$ (Zhao & Yu, 2006; Zou, 2006; Yuan & Lin, 2007). Assume that q of the p environmental variables have nonzero coefficients and that the matrix of environmental variables $\mathbf{X}$ and coefficient vector $\boldsymbol{\beta}$ are arranged such that $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ and $\boldsymbol{\beta} = \{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\}$, where $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ contain coefficients that are nonzero and zero, respectively, and $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ contain the associated environmental variables. Then $\mathbf{C}_m$ can be represented as follows:

$$\mathbf{C}_m = \begin{pmatrix} \mathbf{C}_m^{11} & \mathbf{C}_m^{12} \\ \mathbf{C}_m^{21} & \mathbf{C}_m^{22} \end{pmatrix}.$$

In order for the LASSO estimate $\widehat{\boldsymbol{\beta}}_m$ to be consistent in variable selection, there must be a positive constant vector $\boldsymbol{\epsilon}$ such that $|\mathbf{C}_m^{21}(\mathbf{C}_m^{11})^{-1}\text{sign}(\boldsymbol{\beta}^{(1)})| \leq \mathbf{1} - \boldsymbol{\epsilon}$. Essentially, this puts a limit on the correlation between variables with nonzero coefficients and variables with zero coefficients. In theory, designing the matrix $\mathbf{X}$ to be orthogonal would ensure that the Strong Irrepresentable Condition holds. In SDM, however, this may cause the variables to lose interpretation.

However, Zou (2006) motivated the adaptive LASSO as an alternative to LASSO because it can be both consistent in estimation and variable selection. Theorem 6.1, which is very similar to the asymptotic results of Zou (2006), was used as an intermediate step to Theorem 6.2. Therefore, I expect that it would be possible to likewise extend the asymptotic behaviour of GLMs fitted with adaptive LASSO penalties to the setting of Poisson PPMs by phrasing the question in terms of m instead of n.

I developed MSI as an alternative to MAXENT's default penalty with Theorem 6.2 as the motivation, and in Section 6.4 I showed that the MSI penalty is therefore consistent in estimation. MAXENT software by default uses a penalty that can be understood to have the form $mf(m)$, where $f(m)$ is an *ad hoc* decreasing piecewise linear sequence of penalties (Phillips & Dudík, 2008). As a result, the

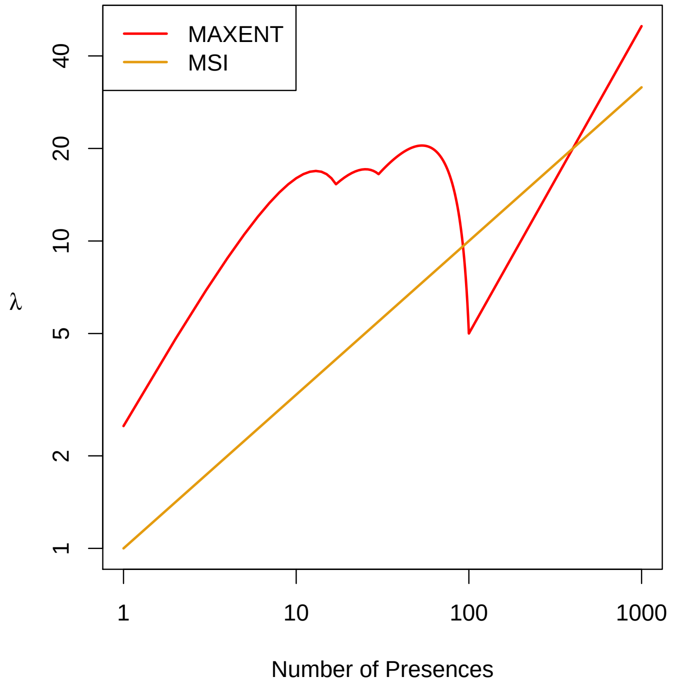

Figure 6.1: Comparison of MAXENT and MSI penalties. The MAXENT penalty does not increase smoothly as the number of presences m increases, while the MSI penalty increases smoothly at a rate of $\sqrt{m}$. The numerator of the MSI penalty is the maximum score function of the intercept model (6.15).

default MAXENT penalty does not increase smoothly as m increases, and in fact at times it decreases (Figure 6.1). For $m \geq 100$, the MAXENT penalty increases at a rate of m instead of $\sqrt{m}$, and therefore does not have order $O(\sqrt{m})$. Hence LASSO estimates chosen by the MAXENT penalty do not satisfy the conditions of Theorem 6.2 required for $\sqrt{m}$-consistency.

While this Chapter provides a theoretical comparison of the MAXENT and MSI penalties, in Chapter 7, I compare them from a data and simulation perspective.

Part III —

New Presence-Only Methods and Their Evaluation

Chapter 7

LASSO Penalty Choice for Species Distribution Models

7.1 Introduction

In Chapter 6, I considered the question of how to choose the LASSO penalty in Poisson PPMs from a theoretical perspective. In this Chapter I address the question from the perspective of real data and simulation by comparing predictive performance and variable selection across many different methods.

Most criteria, including those mentioned in Chapter 4, require a sequence of models with different LASSO penalties (a “regularisation path”) to be calculated to find the optimal point of the bias-variance tradeoff (Figure 7.1). MAXENT, on the other hand, proposes an *ad hoc* value for the penalty based solely on the number of presence cells and the type of variables included (Phillips & Dudík, 2008). These values were determined by optimising the predictive performance of a number of data sets (Phillips & Dudík, 2008) and assuming that the results applied universally. The default MAXENT penalty has yielded good predictive performance (Elith *et al.*, 2006), but there is reason to believe that even better performance is possible. In Chapter 5, non-linear GCV outperformed MAXENT in predictive performance for

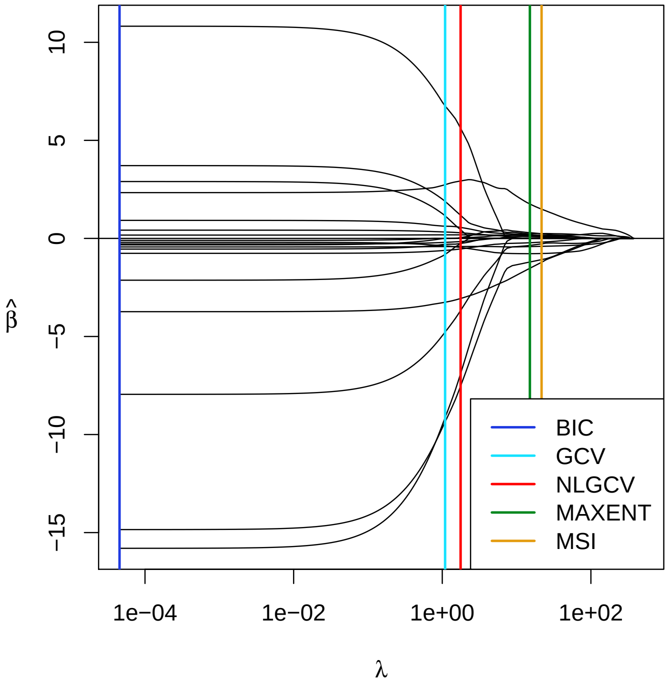

Figure 7.1: Example regularisation path required for most criteria for choosing the LASSO penalty λ. A sequence of models is fitted by varying λ, and the model which optimises the given criterion is chosen. In contrast, MAXENT and MSI only require a single model to be fitted at a value of λ chosen *a priori*, rather than estimation of an entire regularisation path.

Corymbia eximia. In Chapter 6, I developed a new criterion, MSI, which is consistent in estimation whereas MAXENT is not. Whether these advantages apply more generally will be explored in this Chapter.

Others are also considering the performance of MAXENT's default LASSO penalty. Using simulations from a "true" model fitted by MAXENT, Warren & Seifert (2011) showed that given a regularisation path of models, sample-size corrected AIC and BIC are superior in variable selection than criteria based on maximising AUC. Using presence-only data, Gastón & García-Viñas (2011) compared MAXENT's default penalty with three versions of logistic regression and found that penalised logistic regression has similar performance to MAXENT and that both penalised logistic regression and MAXENT exhibited superior performance to unpenalised logistic regression and forward-stagewise logistic regression. Moreover, they found that MAXENT's default penalty had the greatest performance advantage for small sample sizes. Anderson & Gonzalez (2011) segregated the training and test data geographically to account for spatial autocorrelation and found that the default MAXENT penalty achieved near-optimal predictive performance in predicting to the presence-only locations in the test set for small samples, but was generally too small and hence prone to overfitting for larger sample sizes.

The study design I use in this Chapter differs from those mentioned in the previous paragraph in numerous ways – I include additional methods of choosing the LASSO penalty, an additional implementation of LASSO (adaptive LASSO), models which account for interpoint dependence (area-interaction models), and an approach to evaluating model performance using separate presence/absence data that is robust to potential misspecification of point interactions. In each of these ways, this Chapter approaches the problem of LASSO penalty choice in a different way to what has been considered elsewhere in the ecology literature. In Section 7.2 I describe the methods that will be compared in the study, and in Section 7.3 I describe the data that is used in this aim. The metrics with which I compare performance of the different methods and results are presented in Section 7.4. Finally, some discussion follows in Section 7.5.

7.2 Methods of Choosing the LASSO Penalty

In this Section I describe the seven methods of choosing the LASSO penalty considered in this Chapter. These seven methods come from three different classes of methods for choosing the LASSO penalty: information criterion methods, cross validation methods and prevalence-based methods.

The information criterion methods evaluated are AIC (Akaike, 1974), BIC (Schwarz, 1978) and the Hannan-Quinn information criterion (HQC, Hannan & Quinn, 1979).

These methods all share the same form based on the log-likelihood $l(\boldsymbol{\beta}; \mathbf{y}_P)$ (Equation 3.7):

$$-2l(\boldsymbol{\beta}; \mathbf{y}_P) + Cv(\boldsymbol{\beta}),$$

where $v(\boldsymbol{\beta})$ is the effective degrees of freedom, estimated as the number of nonzero coefficient estimates $\sum_{j=1}^{p} I(|\beta_j| > 0)$. These methods vary only in the choice of constant C: $C = 2$, $C = \ln m$, and $C = 2\ln(\ln m)$ for AIC, BIC, and HQC, respectively.

The cross validation methods included are generalised cross validation (GCV, Craven & Wahba, 1979) and non-linear GCV (NLGCV, Fu, 2005). These methods are based on the deviance:

$$\frac{D(\boldsymbol{\beta}; \mathbf{y}_P)}{m(1 - v(\boldsymbol{\beta})/m)^2},$$

where $D(\boldsymbol{\beta}; \mathbf{y}_P)$ is the deviance of a Poisson PPM, m is the number of presence locations and $v(\boldsymbol{\beta})$ is the effective degrees of freedom.

For GCV, $v(\boldsymbol{\beta}) = (\mathbf{X}'\mathbf{W}\mathbf{X} + \lambda\mathbf{G}_{\boldsymbol{\beta}})^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}$, where $\mathbf{W} = \mathbf{w}\boldsymbol{\mu}$ for a vector of quadrature weights $\mathbf{w}$ and $\mathbf{G}_{\boldsymbol{\beta}} = \text{diag}(1/|\boldsymbol{\beta}|)$, replacing undefined elements of the diagonal with 0 when $\beta_j = 0$. For non-linear GCV, $v(\boldsymbol{\beta}) = pq(\boldsymbol{\beta})$, where $q(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j| / \sum_{j=1}^{p} |\hat{\beta}_{\text{GLM},j}|$.

The prevalence-based methods included are MAXENT's *ad hoc* method and MSI (Equation 6.12). MAXENT software uses a decreasing piecewise linear function $f(m)$ to determine the penalty when linear, quadratic, and interaction terms are used as in this Chapter:

$$f(m) = \begin{cases} -0.1m + 2.6 & : m \leq 17 \\ -\frac{7}{260}m + \frac{353}{260} & : 18 \leq m \leq 30 \\ -\frac{1}{140}m + \frac{107}{140} & : 31 \leq m \leq 100 \\ 0.05 & : m > 100. \end{cases}$$

From Chapter 5, the constrained entropy used by MAXENT $\Lambda(\boldsymbol{\beta}; \mathbf{y}_P)$ (Equation

5.7) can be related to the likelihood of a Poisson PPM $l(\boldsymbol{\beta}; \mathbf{y}_P)$ as follows:

$$\Lambda(\boldsymbol{\beta}; \mathbf{y}_P) = \frac{l(\boldsymbol{\beta}; \mathbf{y}_P)}{m} + C,$$

where C is a constant with respect to $\boldsymbol{\beta}$. Consequently,

$$\Lambda(\boldsymbol{\beta}; \mathbf{y}_P) - f(m) \sum_{j=1}^{p} |\beta_j| \propto l(\boldsymbol{\beta}; \mathbf{y}_P) - m f(m) \sum_{j=1}^{p} |\beta_j|.$$

Hence for MAXENT, the default penalty is $mf(m)$ in the context of Poisson PPMs, as shown in Figure 6.1.

7.3 Data Sets and Evaluation of Performance

In this Section I describe the data sets and metrics used to compare the methods for choosing the LASSO penalty.

To investigate the performance of the methods of choosing the LASSO penalty presented in Section 7.2, I chose to analyse both real species data and simulated data. The real species data, consisting of two pairs of data sets (a presence-only data set and an independently collected presence-absence data set), allow me to evaluate the predictive performance of these methods in practice. The simulations allow me to evaluate the performance of these methods in selecting the right subset of environmental variables and replicating the true intensity surface under a suite of assumptions about model structure and species abundance.

7.3.1 Real Species Data

The two pairs of real species data sets analysed represent two different families of plants in two different regions of New South Wales, Australia. The Blue Mountains pair tracks locations of 181 eucalypt species within 100 km of the Blue Mountains Region near Sydney (NSW Office of Environment and Heritage, 2012), whereas the Hunter Valley pair tracks locations of 31 fern species within 20 km of the Hunter

Valley (NSW Office of Environment and Heritage, 2012). In each pair there are two data sets – the first ("presence-only") consists only of reported presence locations $\mathbf{y}_P$ for the 181 eucalypt species (Blue Mountains) or 31 fern species (Hunter Valley), while the second ("presence-absence") consists of a systematic list of both presences and absences for the corresponding species at each of 12,504 (Blue Mountains) or 3,340 (Hunter Valley) pre-determined sites.

All models built for these data sets use a combination of environmental variables and variables measuring the accessibility of sites. For the Hunter Valley species, the environmental variables included minimum and maximum annual temperature (MNT and MXT) and annual rainfall (Rain). For the Blue Mountains species, a fourth environmental variable included was the number of fires recorded since 1943 (FC). The site accessibility variables included for both data sets were distance from main roads (D.Main) and distance from urban areas (D.Urb). I used a quadratic function of all variables, including interactions between the environmental variables and between the two accessibility variables as in Section 5.3. The functions I wrote for this purpose are available in the `ppmlasso` package (Chapter 8) for `R`. Quadrature points were chosen along a regular grid at the 1km × 1km spatial resolution following the method of Warton & Shepherd (2010).

My approach was to model the occurrence of each of the 181 Blue Mountains and 31 Hunter Valley species using the presence-only data and predict to new areas, measuring predictive performance using the presence-absence data. In order to construct "new areas" in which to predict, I applied a 5-fold spatial cross-validation scheme as in Section 5.4 using the presence-only data as training data and the presence-absence data as test data. Sites were grouped into square blocks that were randomly assigned to training and test samples. This scheme was implemented with blocks of size 64 × 64 km. If necessary, block length was repeatedly halved until no more than half of the presence-only locations fell into a single group. I then constructed a regularisation path of 202 models on the training blocks corresponding

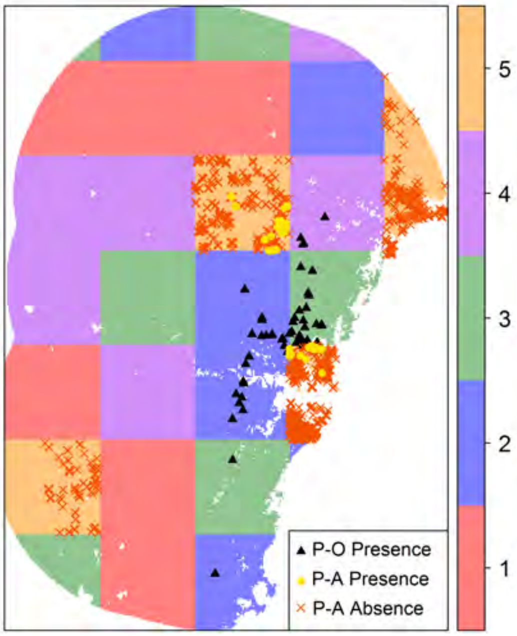

Figure 7.2: 5-fold spatial cross validation scheme used in analysis. Here, blocks in cross validation groups 1-4 are used to build models and blocks in cross validation group 5 are used to test models. Presence-only locations (▲) and quadrature points in cross validation groups 1-4 are used to build Poisson PPMs, which are then used to predict species intensities at the known presence (●) and absence (×) locations in cross-validation group 5.

---

to 200 LASSO penalties equally spaced on a logarithmic scale from e^{-10} to $\lambda_{\max}$ as well as the MAXENT and MSI penalties. The models corresponding to each of the criteria were used to calculate predicted intensities in the test blocks. Figure 7.2 illustrates the spatial cross-validation scheme to split both data sets into training and test blocks.

After applying this procedure to each cross validation group, I constructed a ROC curve from the predicted intensities and observed presence or absence of the species and measure the predictive performance by area under the curve (AUC). I repeated the randomisation procedure for assigning the data to test and training blocks 40 times.

The combination of using separate data sets for training and testing as well as applying spatial cross-validation represents a new approach to comparing methods of choosing the LASSO penalty. The studies of Gastón & García-Viñas (2011) and Anderson & Gonzalez (2011) applied a spatial cross-validation scheme, but only on a single data set, while the study of Elith *et al.* (2006) used separate data sets for training and testing but without any spatial cross-validation. The approach used in this Chapter reflects the notion that while the presence-only and presence-absence data sets were collected separately, this does not obviate the need to account for spatial autocorrelation.

I implemented this procedure across a $2 \times (1 + 3 \times 7)$ design (Figure 7.3). I considered two classes of model to evaluate: Poisson PPMs and area-interaction models with a species-specific interaction radius chosen to minimise the pseudolikelihood, ranging from 1 km to 10 km by increments of 0.1 km. For each class of model, I calculated the unpenalised model as well as regularisation paths from 3 implementations of LASSO – regular LASSO, adaptive LASSO with adaptive weights coming from the coefficients of the unpenalised solution (Ad-Unp LASSO), and adaptive LASSO with adaptive weights coming from the model coefficients determined by the seven criteria discussed in Section 7.2 (Ad-Method LASSO). For LASSO and Ad-Unp LASSO, I determined the models corresponding to each of the seven methods of Section 7.2. For Ad-Method LASSO, I determined the model that optimised the same criterion used to calculate the adaptive weights. These models were then used to calculate predicted intensities at the presence/absence locations in the test blocks.

7.3.2 Simulations

The design of simulations (Figure 7.3) largely mimics that of the real species data. Simulated presence locations were generated using realisations of either a Poisson PPM or area-interaction model with radius 2 km and interaction parameter of 2

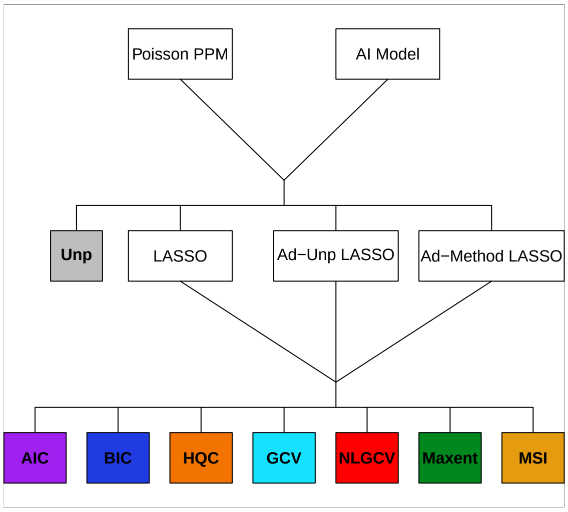

Figure 7.3: $2 \times (1 + 3 \times 7)$ design of models fitted and compared. There were two model types fitted: Poisson PPMs and area-interaction models. For both model types, I calculated (1) the unpenalised model, (2) a LASSO regularisation path, (3) an adaptive LASSO regularisation path using the unpenalised coefficients to determine initial weights (Ad-Unp LASSO), and (4) an adaptive LASSO regularisation path using the coefficients of the seven methods of choosing the LASSO penalty to determine initial weights (Ad-Method LASSO).

---

different strengths – weak ($0.1 \sum_{j=1}^{p} |\beta_j|$) and strong ($0.5 \sum_{j=1}^{p} |\beta_j|$). As the relative performance of the various methods of choosing the LASSO penalty may vary with true model complexity and magnitude of species prevalence, I generated 100 realisations across three levels of model sparsity and five levels of species abundance (Table 7.1) for each of six species using the `rpoispp` and `rmh` functions in the `spatstat` package in `R`. I calculated the relevant environmental variables at the simulated presence locations using bilinear interpolation (see Appendix B) from a regular 100m × 100m grid of environmental data.

Models corresponding to each method of choosing the LASSO penalty were com-

Table 7.1: Design of Simulations. For each of six species, I simulated from a Poisson PPM with one of three levels of sparsity (measured by the number of non-zero variables k out of 19) and abundance (measured by $\mu_{\mathcal{A}}$). Combinations of abundance and sparsity with check marks are presented in Figures 7.7-7.9.

		Sparsity		
		Full ($k = 19$)	Moderate ($k = 12$)	Sparse ($k = 6$)
Abundance	Super Abundant ($\mu_{\mathcal{A}} = 810$)	✓	✓	✓
	Abundant ($\mu_{\mathcal{A}} = 270$)			
	Moderate ($\mu_{\mathcal{A}} = 90$)	✓	✓	✓
	Rare ($\mu_{\mathcal{A}} = 30$)			
	Super Rare ($\mu_{\mathcal{A}} = 10$)	✓	✓	✓

pared in (1) selection of the correct subset of environmental variables, (2) accuracy of estimating regression coefficients, and (3) accuracy of predictions.

Variable selection performance was measured by counting both the number of variables correctly included in the model with the correct sign and the number of variables correctly shrunk to zero.

Accuracy in estimating regression coefficients ($\widehat{\boldsymbol{\beta}}$) was determined by calculating their mean squared error. Denote by $\widehat{\boldsymbol{\beta}}^{(i)}$ the predicted coefficients for the ith simulation and by $\boldsymbol{\beta}_*$ the vector of true coefficients. Given that there were 100 simulations, mean squared error was calculated by:

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) = \frac{1}{100}\sum_{i=1}^{100}(\widehat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}_*)'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\widehat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}_*), \tag{7.1}$$

where $\mathbf{W} = \text{diag}(\mathbf{w}\boldsymbol{\mu}_*)$ for a vector of quadrature weights $\mathbf{w}$ and a vector of true intensities $\boldsymbol{\mu}_*$. This is a scaled mean squared error, where $\widehat{\boldsymbol{\beta}}$ was rescaled by $\mathbf{X}'\mathbf{W}\mathbf{X}$ to account for correlation and differences in variability across regression coefficients. As usual, (7.1) can be decomposed into bias and variance components:

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) = \text{bias}^2(\widehat{\boldsymbol{\beta}}) + \text{var}(\widehat{\boldsymbol{\beta}}),$$

where

$$\begin{aligned}\text{bias}^2(\widehat{\boldsymbol{\beta}}) &= (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_*),\\ \text{var}(\widehat{\boldsymbol{\beta}}) &= \frac{1}{100}\sum_{i=1}^{100}(\widehat{\boldsymbol{\beta}}^{(i)} - \widetilde{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\widehat{\boldsymbol{\beta}}^{(i)} - \widetilde{\boldsymbol{\beta}}),\end{aligned}$$

and $\widetilde{\boldsymbol{\beta}} = \frac{1}{100}\sum_{i=1}^{100}\widehat{\boldsymbol{\beta}}^{(i)}$.

Predictive performance was measured using integrated mean squared error, defined as:

$$\text{IMSE}(\widehat{\boldsymbol{\eta}}) = \int_{y\in\mathcal{A}}(\eta_*(y) - \hat{\eta}(y))^2 dy,$$

where $\hat{\eta}(y) = \ln\hat{\mu}(y)$ and $\eta_*(y)$ is the true log intensity at location y. Given a set of quadrature points $\mathbf{y}_0 = \{y_{m+1}, \ldots, y_n\}$ and quadrature weights $\mathbf{w}$ as in Chapter 3, integrated mean squared error can be approximated as:

$$\text{IMSE}(\widehat{\boldsymbol{\eta}}) \approx \sum_{i=1}^{n} w_i(\eta_{*,i} - \hat{\eta}_i)^2,$$

where $\hat{\eta}_i$ and $\eta_{*,i}$ are respectively the predicted and true log intensities at location y_i. Kullback-Leibler divergence and correlation between the true intensity and predicted intensity were also calculated, but results are not presented because they are similar to those of integrated mean squared error.

The true parameters used to simulate presence locations for each combination of model sparsity and species abundance come from a regularisation path of Poisson PPMs for six eucalypt species in the Blue Mountains region – *Acmena smithii*, *Corymbia eximia*, *Eucalyptus canaliculata*, *Eucalyptus eugenioides Eucalyptus rubida*, and *Homoranthus cernuus*. These species were chosen as they occupy distinct environmental and geographic regions within the study area. For each species, I chose three models from the regularisation path to correspond to different model sizes (labelled “Full”, “Moderate”, and “Sparse”) and adjusted the intercept to correspond to one of five abundance levels (labelled “Super Abundant”, “Abundant”, “Moderate”, “Rare”, and “Super Rare”) as in Table 7.1.

Performing the calculations for this design was computationally intensive, requiring over 2 months of computation time on computational clusters of Dell PowerEdge M610 (Intel Xeon X5660, 2.80 GHz) processors.

7.4 Results

7.4.1 Blue Mountains and Hunter Results

The average area under the ROC curve (AUC) for the various criteria is presented in Table 7.2. As some information criterion methods performed similarly and both cross validation methods performed similarly, all subsequent results will include only the unpenalised model, BIC, non-linear GCV, MAXENT, and MSI, with AUC as the measure of predictive performance throughout. It appears that the *ad hoc* MAXENT method performed poorly, with worse average AUC than the unpenalised model and ahead of only GCV. Figure 7.4 illustrates the relationship between predictive performance and prevalence as estimated from a generalised additive model (Wood, 2011). Both the Blue Mountains and Hunter datasets exhibited similar patterns. MAXENT appeared to perform poorly for rare species but its performance generally improved as prevalence increased. MSI and non-linear GCV performed well for rare species, but due to high inter-species variability, it was difficult to ascertain significant differences among methods other than MAXENT for rare species.

When considering models shrunk by adaptive LASSO, there was no appreciable change in performance with the exception of MAXENT and non-linear GCV for rare species, which performed better with the Ad-Unp implementation but worse

Table 7.2: Average predictive performance (measured as average area under the ROC curve for 40 different 5-fold spatial cross-validation schemes) and standard error of different methods of choosing the LASSO penalty when predicting to a separate presence-absence dataset.

LASSO Penalty	**Blue Mountains**		**Hunter Valley**	
Criterion	**AUC**	**S.E.**	**AUC**	**S.E.**
Unpenalised	0.7356	0.0083	0.6492	0.0229
AIC	0.7393	0.0083	0.6530	0.0216
BIC	0.7467	0.0083	0.6600	0.0207
HQC	0.7407	0.0083	0.6580	0.0216
MAXENT	0.7276	0.0124	0.6288	0.0288
GCV	0.7199	0.0121	0.6258	0.0291
Non-Linear GCV	0.7686	0.0086	0.6832	0.0237
MSI	0.7799	0.0092	0.6753	0.0262

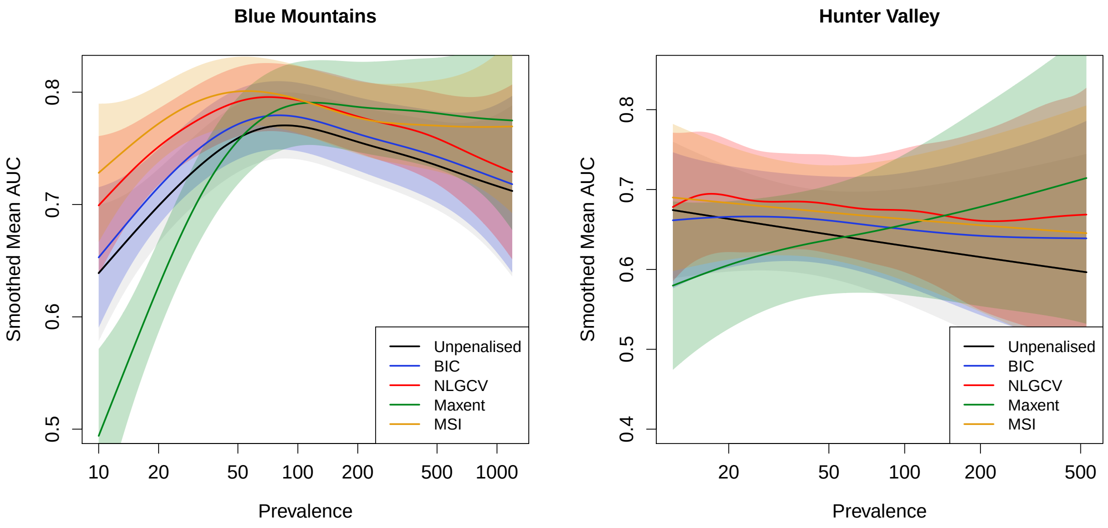

Figure 7.4: Predictive performance of different optimisation criteria applied to Poisson PPMs for (left) 181 eucalypt species in the Blue Mountains region near Sydney and (right) 31 fern species in the Hunter Valley. 95% confidence bands are shaded around each curve.

with the Ad-Method implementation (Figure 7.5). Applying an adaptive LASSO penalty generally reduced the difference in performance among the various criteria.

To investigate whether fitting area-interaction models improves predictive performance, I examined the inhomogeneous K-function of the 181 Blue Mountains species, as in Section 3.3. Among the 181 Blue Mountains species, 113 exhibited at least moderate deviation from the confidence bounds of 95% simulation envelopes and hence I fitted area-interaction models to these species. Figure 7.6 illustrates that there was no appreciable benefit in predictive performance when applying area-interaction models at any level of prevalence or interaction radius, and in fact was worse for MAXENT for moderately rare species. A notable difference is that the unpenalised model performed relatively better – while the performance of methods based on a LASSO penalty deteriorated as interaction radius grew larger, the unpenalised model did not lose any ground. For large radii, the unpenalised model had the highest average AUC, significantly better than MAXENT and BIC.

7.4.2 Simulation Results

The simulation results will be presented by examining (1) the selection of variables, (2) mean square error, and (3) predictive performance of each method. Appendix A also shows a comparison of the amount of shrinkage applied, as an alternative measure of overfitting or underfitting in comparison with Figures 7.7 and 7.8. Patterns were similar for all six species, so results are only shown for *Corymbia eximia*.

Figure 7.7 consists of double-sided barplots for the unpenalised model, BIC, nonlinear GCV, MAXENT, and MSI. The top part of each bar represents the number of parameters that should be in the model that were either falsely excluded or given the wrong sign, while the bottom part of each bar represents the number of parameters that were falsely included. Hence the overall size of the bar is the total number of variables misclassified and the location of the bar indicates whether methods tended

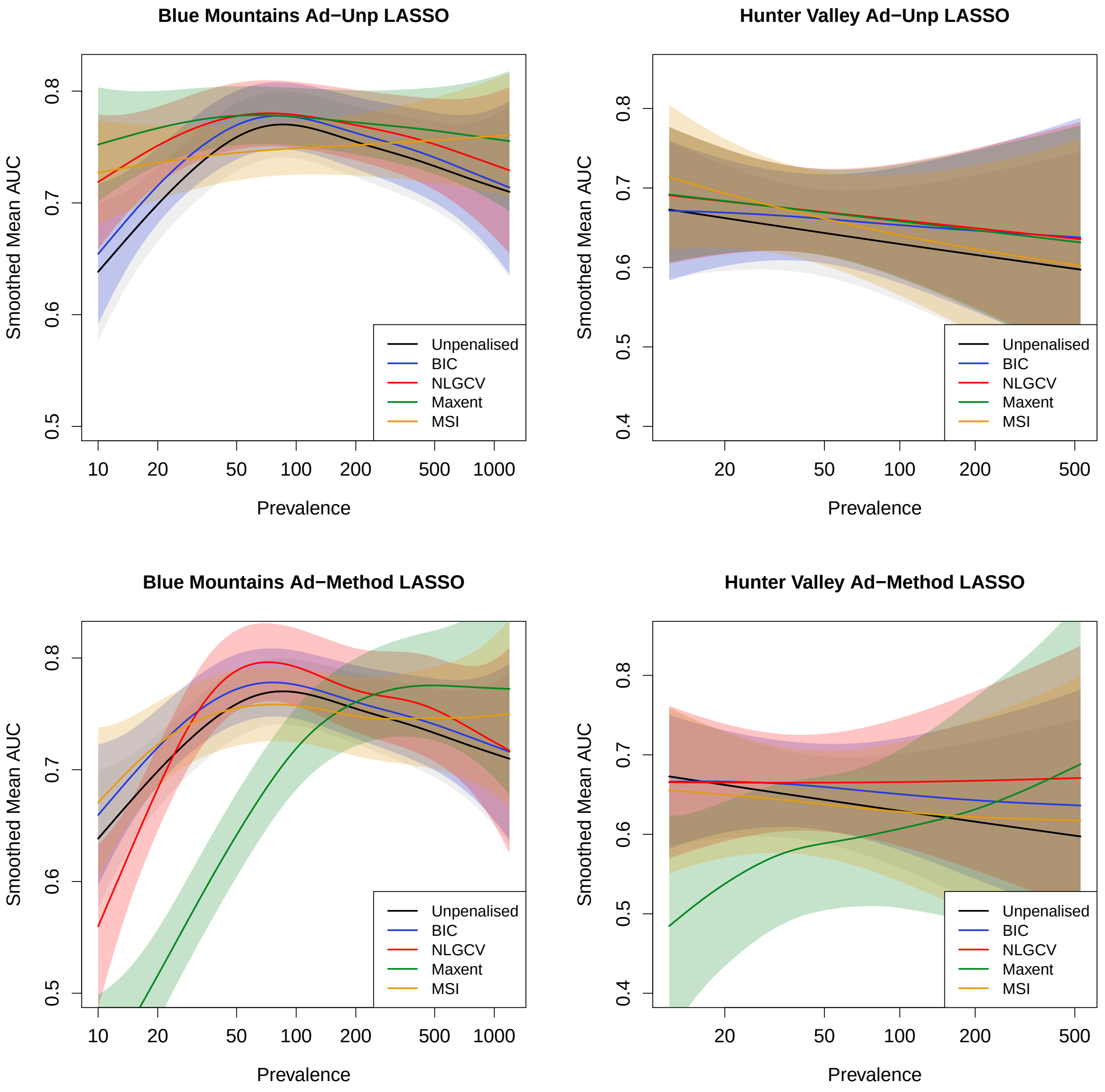

Figure 7.5: Predictive performance of different optimisation criteria applied to Poisson PPMs fitted with (top) Ad-Unp LASSO and (bottom) Ad-Method LASSO penalties. Most methods had comparable performance to their normal LASSO counterparts (Figure 7.4), with the exception of MAXENT and non-linear GCV for rare species, which performed worse for Ad-Method LASSO but better for Ad-Unp LASSO.

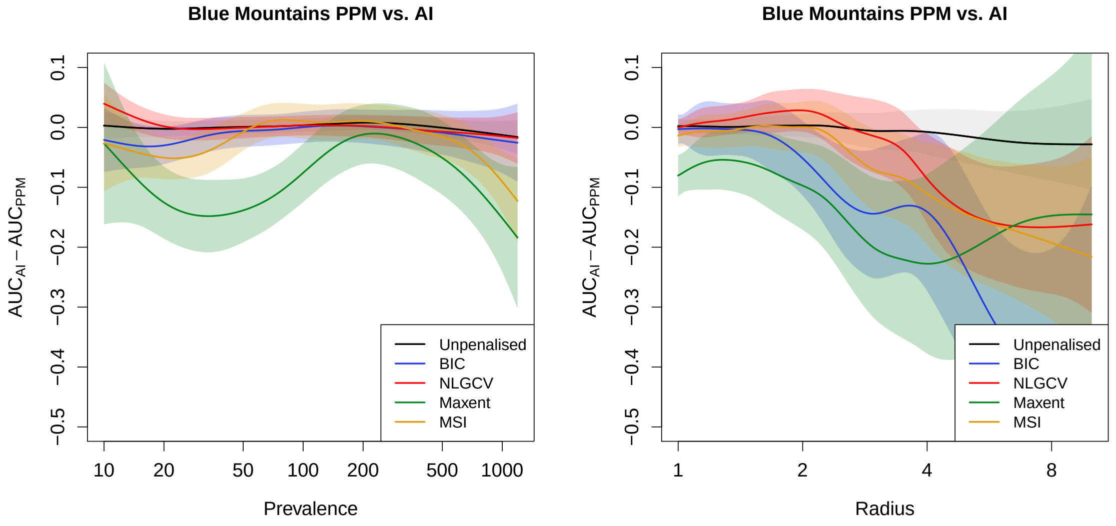

Figure 7.6: Predictive performance of different optimisation criteria applied to area-interaction models of 113 Blue Mountains species for which K-envelopes suggested the presence of interpoint interactions. There was no appreciable benefit in performance in fitting area-interaction models at either any level of prevalence (left) or interaction radius (right). For MAXENT, performance was worse when fitting an area-interaction model for moderately rare species. Performance of area-interaction models appeared to get worse as the interaction radius increased.

to overfit (bar is centred below zero) or underfit (bar is centred above zero).

The unpenalised model tended to overfit the data and hence falsely included the most variables, while non-linear GCV and MAXENT for super rare species tended to underfit and thus falsely exclude the most variables. BIC and MSI had the best overall performance. As abundance increased, each method improved overall variable selection regardless of model sparsity.

How variable selection impacted mean squared error (MSE) can be seen in Figure 7.8. The size of the bars correspond to the value of the MSE for each method. Each bar consists of a lighter portion and a darker portion, representing the proportion of MSE attributable to bias and variance, respectively. This was done to illustrate how each method judges the bias-variance tradeoff.

Overall MSE appeared to decrease as abundance increased and models became more sparse. Methods with high shrinkage such as non-linear GCV and MAXENT for rare species exhibited high bias and low variance. High shrinkage is desirable for rare species (where variance is high) and sparse models (where bias is low), whereas low shrinkage is desirable for abundant species (where variance is low) and full models (where bias can be high). BIC appeared to be severely punished with high variance for gross overfitting in a few outlying simulations of super rare species. Overall MSE was fairly similar across all methods, with the exception of BIC for super rare species as previously noted.

To compare the performance of the different methods in estimating the true intensity surface, I examine integrated mean square error in Figure 7.9.

It tells a broadly similar story to the real species data. The benefit of shrinkage increased as abundance decreased. When shrinkage was desired, MSI was competitive for all abundance levels as in the real species data, while MAXENT performed relatively worse for rare species, although not to nearly the same extent as for the real data as shown in Figure 7.4. Because MSI and MAXENT overshrank to the

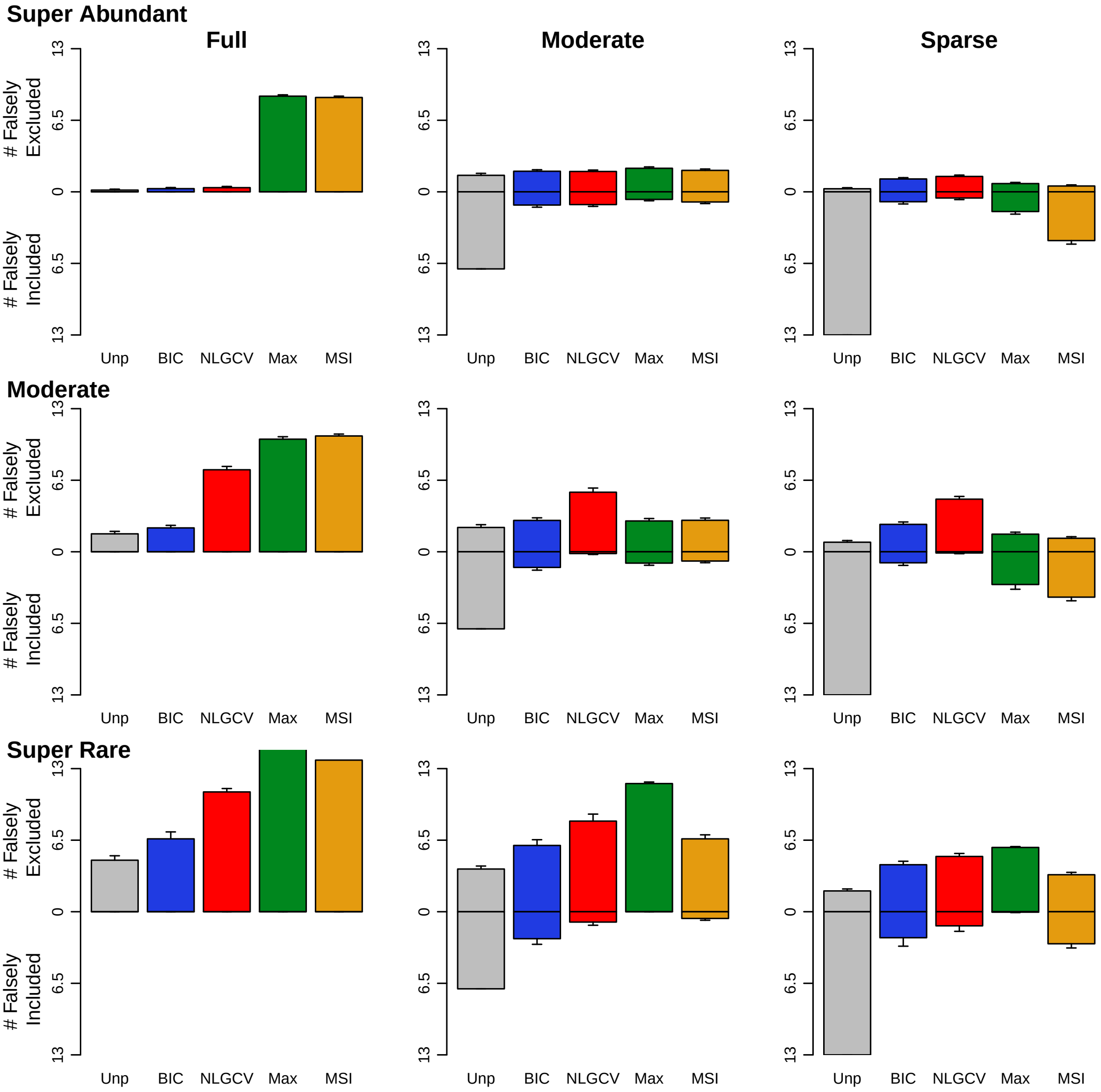

Figure 7.7: Variable selection of different methods of choosing the LASSO penalty. Bars above 0 indicate the number of variables in the true model that are incorrectly excluded or given the wrong sign. Bars below 0 indicate the number of variables equal to 0 in the true model but incorrectly included.

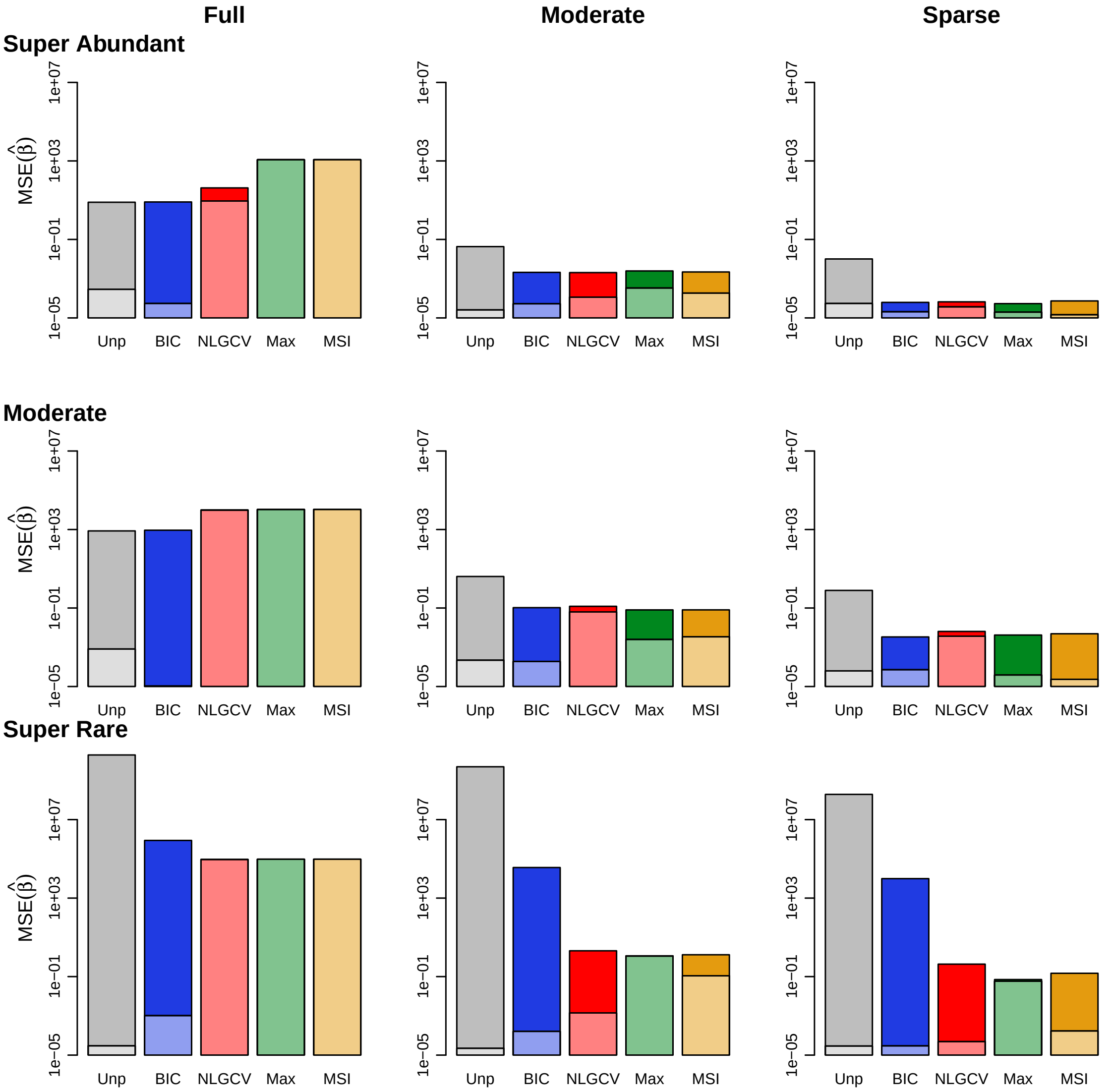

Figure 7.8: Mean Square Error of methods for choosing the LASSO penalty from 100 simulations at each of three different levels of model complexity (columns) and prevalence (rows). The lighter parts of the bars represent bias and the darker parts represent variance.

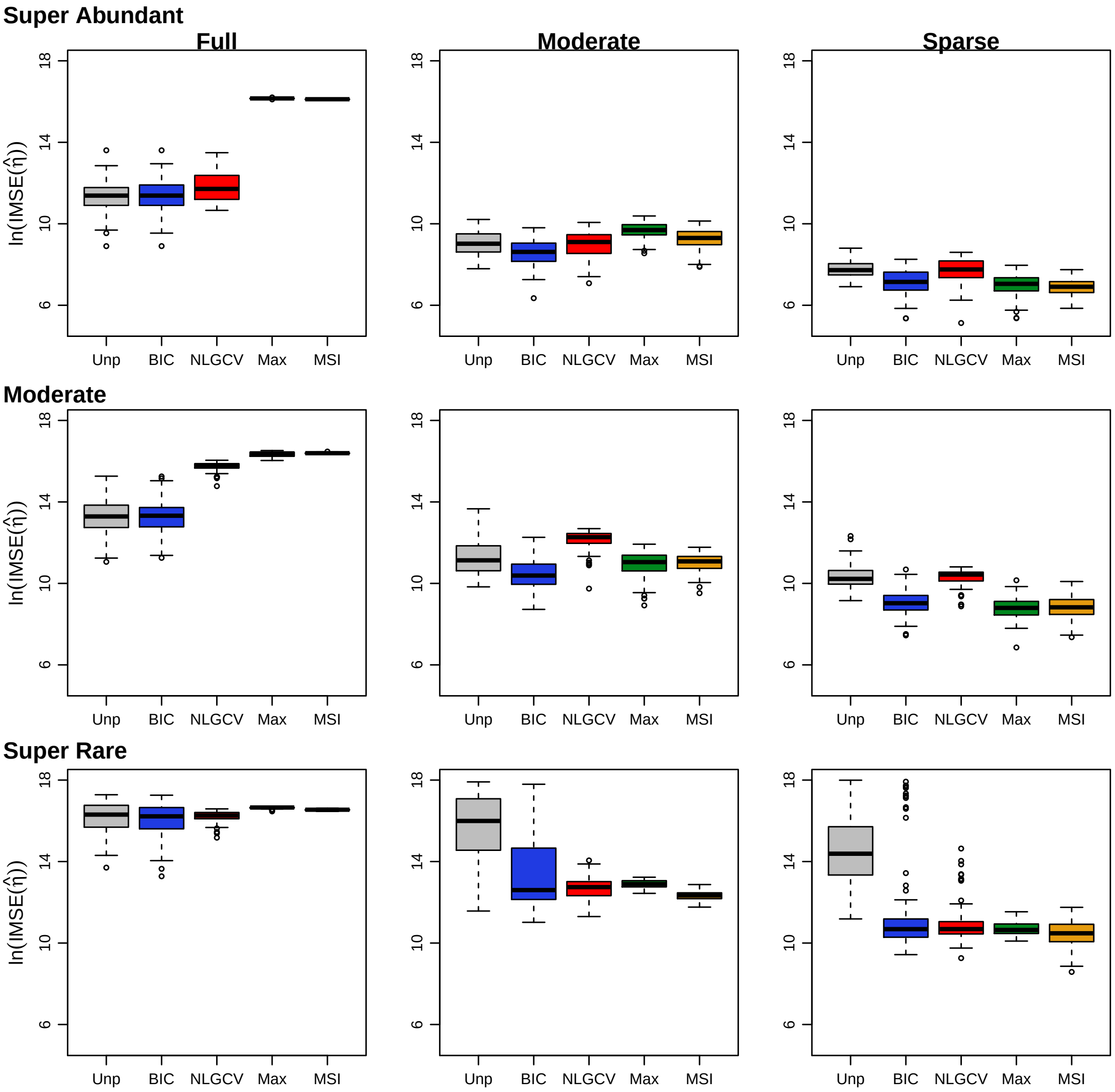

Figure 7.9: Integrated mean square error boxplots of methods for choosing the LASSO penalty from 100 simulations at each of three different levels of model complexity (columns) and prevalence (rows).

largest degree for full models (Figure 7.7), their predictive performance suffered for moderately abundant and super abundant species. The main differences in results from the real species data are that the relative predictive performance was better for BIC and worse for non-linear GCV in the simulated data, which indicates that the predictive cost of underfitting tended to outweigh the cost of overfitting.

Figure 7.10 compares the variable selection, MSE, and integrated mean squared error for LASSO and both adaptive LASSO implementations (Ad-Unp LASSO and Ad-Method LASSO). The plots correspond to moderate level sparsity and super rare abundance, as the patterns did not vary appreciably across different levels of sparsity and abundance. Applying adaptive LASSO does not generally change whether methods overfit or underfit. Overall MSE was generally similar regardless of the type of LASSO penalty applied, with the exception of MAXENT for Ad-Unp. The Ad-Unp implementation of adaptive LASSO generally reduced bias and increased variance. The reduction in bias seemed to allow MAXENT to close the gap in performance with MSI for Ad-Unp LASSO. This improvement was likewise apparent with the real species data (Figure 7.5, top row).

To examine the impact of incorrectly assuming independence among point locations in the existence of point interactions, I compared the performance of the different methods for choosing the LASSO penalty when fitted using the correct area-interaction model framework and the incorrect Poisson point process model framework. Figures 7.11 and 7.12 illustrate this comparison when the true interaction parameter was strong and weak, respectively. Because the pattern remained the same across sparsity levels, these plots correspond to moderately sparse models.

The results were somewhat complex and depended on species prevalence and the strength of the interaction parameter. When the true interaction parameter was strong (Figure 7.11), penalisation did not generally improve predictive performance, and in fact generally made it worse for rare species (Figure 7.11b). However, correctly fitting an area-interaction model was beneficial when data were sufficient

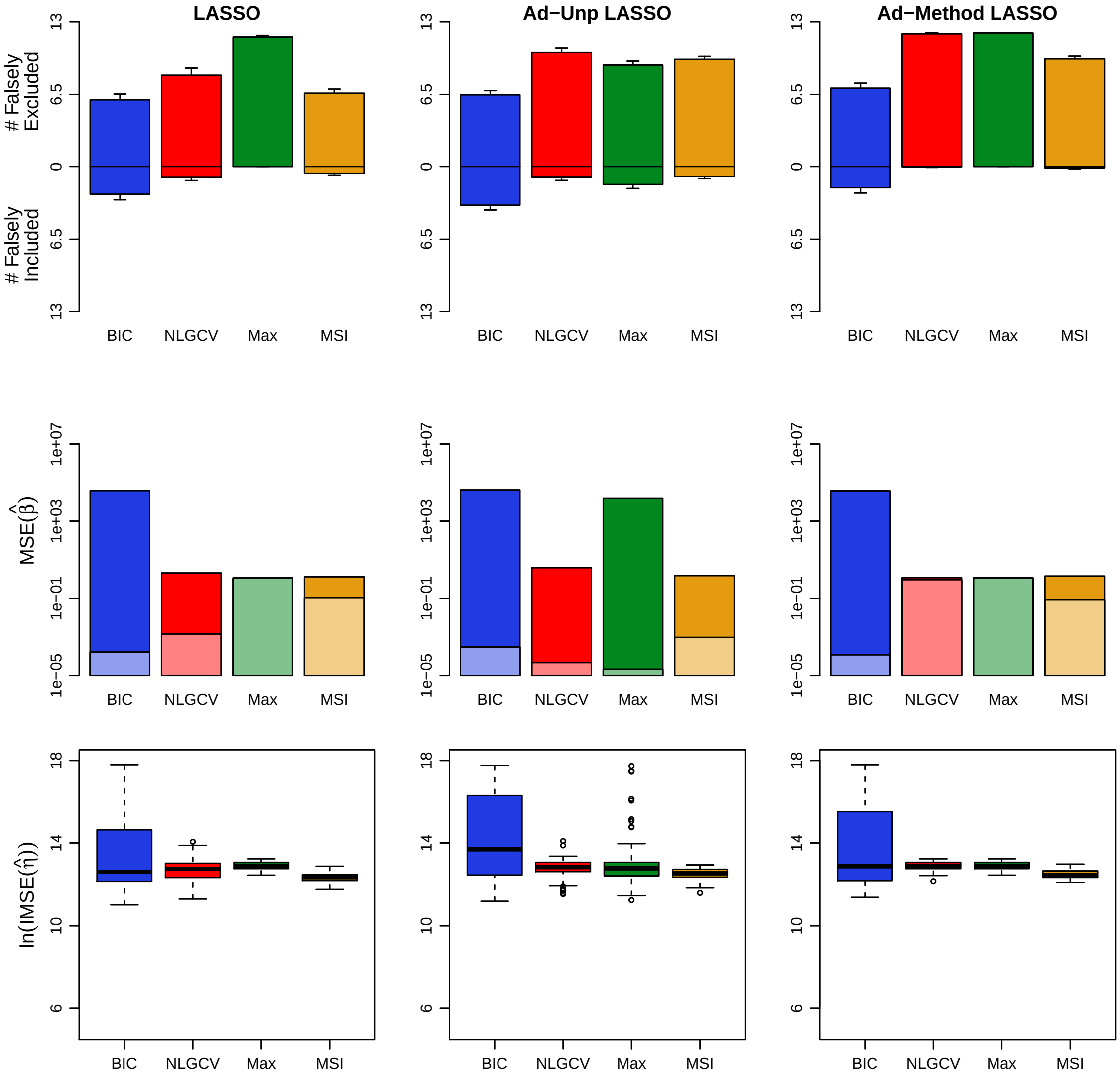

Figure 7.10: Comparison of methods for LASSO (left) and adaptive LASSO for super rare species. Adaptive weights for each method were chosen using coefficients of the unpenalised model (Ad-Unp LASSO, middle) or from LASSO estimates derived from the other methods (Ad-Method LASSO, right). MSE is generally unchanged by implementing an adaptive LASSO penalty, with the exception of MAXENT for Ad-Unp LASSO. The Ad-Unp implementation generally reduced bias but increased variance, allowing MAXENT to approach the predictive performance of MSI as measured by IMSE.

(Figure 7.11a).

When the true interaction parameter was weak (Figure 7.12), applying a LASSO penalty sometimes improved predictive performance, particularly for rare species. MSI had good performance regardless of prevalence. Unlike the real species data, MAXENT also performed well for both abundant and rare species. Correctly fitting an area-interaction model generally improved predictive performance, except in the case of MAXENT and MSI for abundant species (Figure 7.12a).

For both interaction strengths, all methods shrank environmental coefficients more when fitted as an area-interaction model than when fitted as a Poisson PPM (see details in Appendix A). In effect, the inclusion of an interaction term absorbed some of the signal from the environmental variables, thereby reducing their relative importance and making them more susceptible to elimination from the model by imposing a LASSO penalty. Subsequently, there were more environmental variables falsely excluded and fewer environmental variables falsely included in an area-interaction model.

7.5 Discussion

In this Chapter I have conducted a thorough comparison of methods of choosing the LASSO penalty that includes new methods (MSI) as well as comparisons to other LASSO implementations (adaptive LASSO) and models (area-interaction models) previously unconsidered in ecology. The results showed that there are a number of factors that influence the performance of various methods for choosing the LASSO penalty. For rare species, MSI tended to perform best, but when sufficient data were available, data-driven approaches like BIC and non-linear GCV were just as good (Figure 7.9). MAXENT tended to have the worst performance of all methods for rare species but had competitive performance for moderate and abundant species (Figure 7.9). The poor performance for rare species is problematic as there are typ-

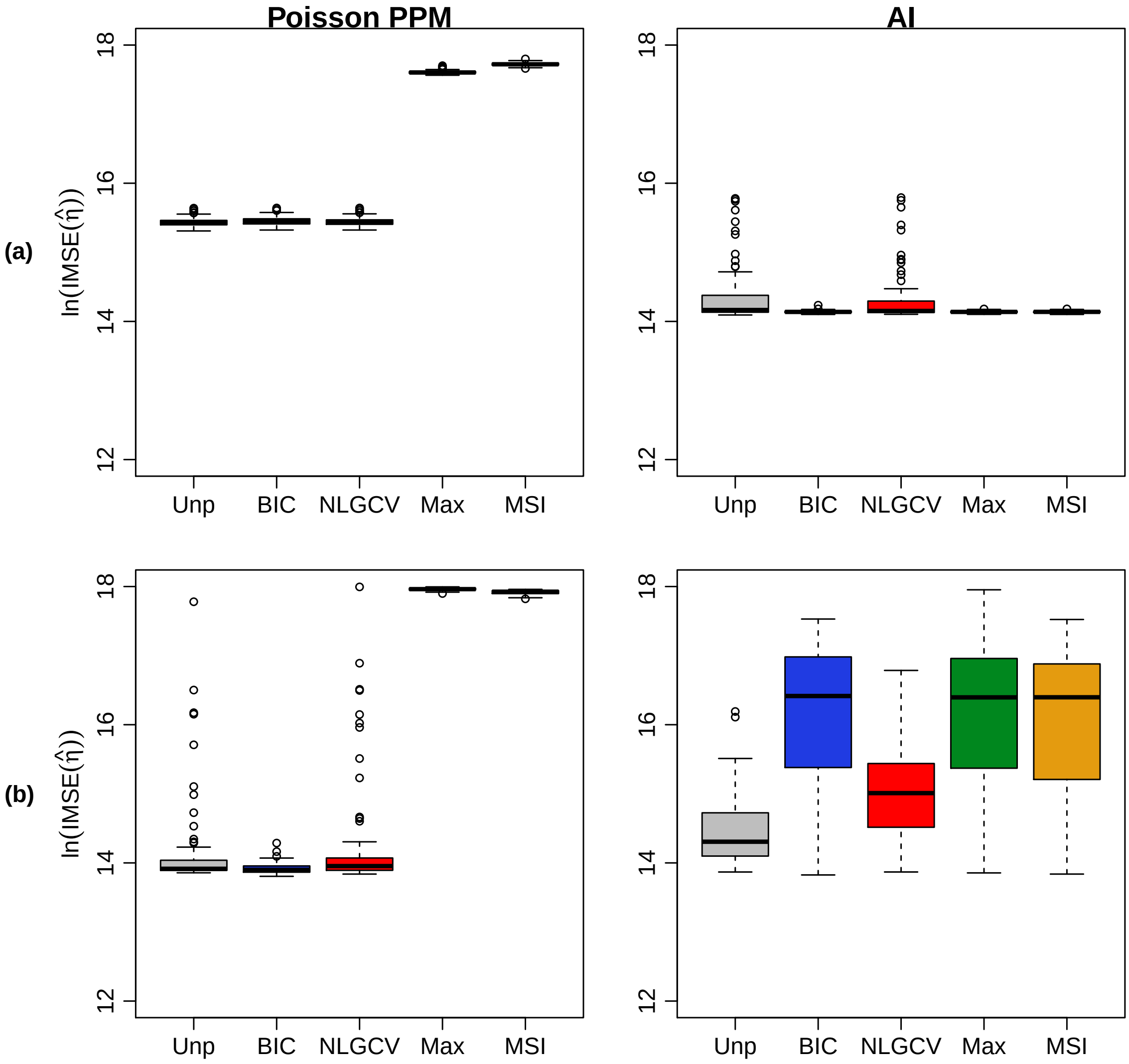

Figure 7.11: Integrated mean squared error for data generated from a moderately sparse area-interaction model with strong point interactions. A Poisson PPM and area-interaction model are both fitted. Correctly fitting an area-interaction model improved predictive performance for super abundant species (a), but not for super rare species (b). Fitting a LASSO penalty did not improve predictive performance for rare species (b) regardless of the method chosen.

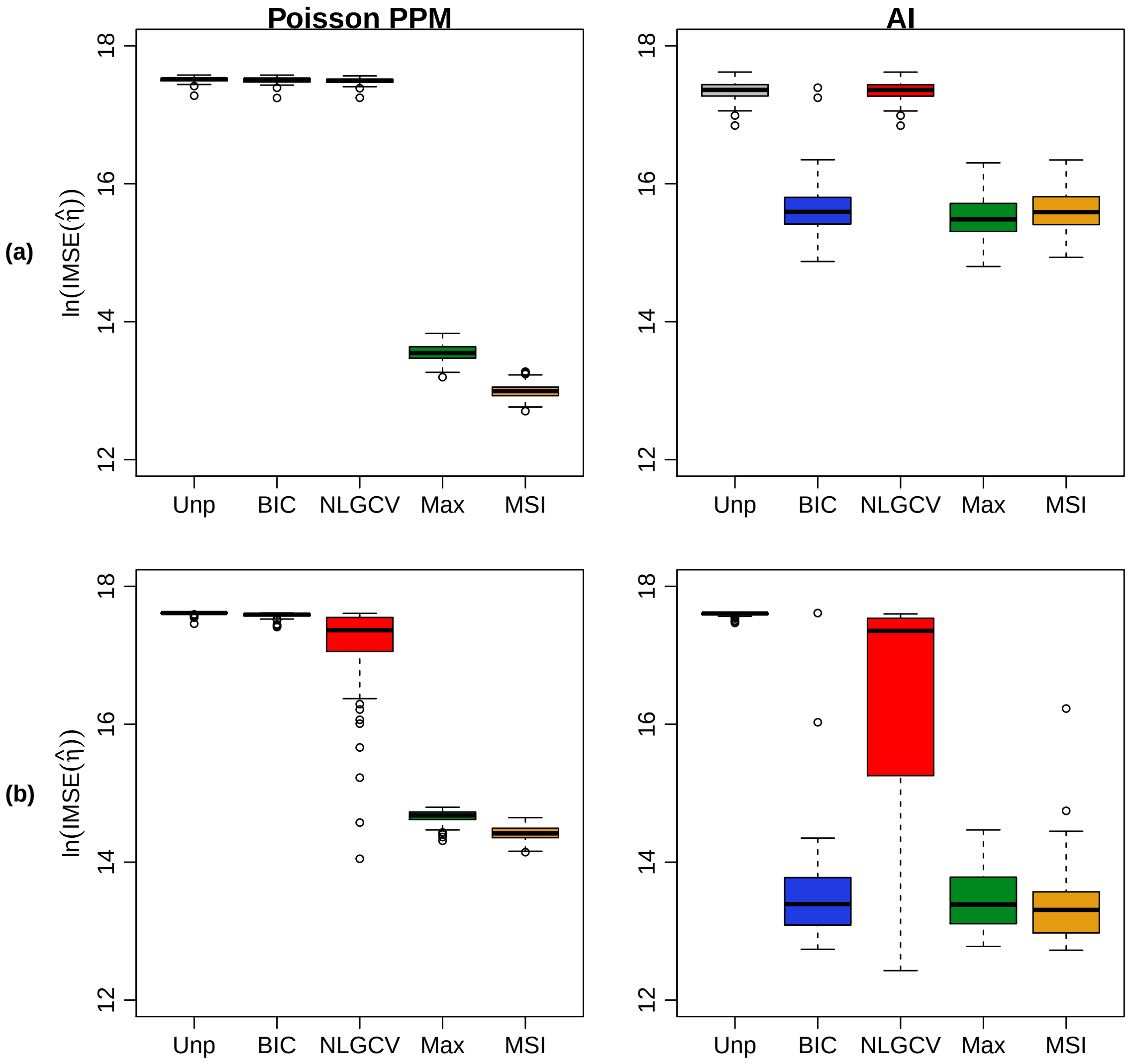

Figure 7.12: Integrated mean squared error for data generated from a moderately sparse area-interaction model with weak point interactions. A Poisson PPM and area-interaction model are both fitted. Fitting a LASSO penalty generally improved predictive performance for (a) super abundant and (b) super rare species. Correctly fitting an area-interaction model also generally improved predictive performance, except for MAXENT and MSI for super abundant species.

ically many rare species in a community leading to poorer performance (Figure 7.4 and bottom of Figure 7.9).

MSI was proposed as an alternative prevalence-based method to MAXENT. It generally outperformed MAXENT with respect to predictive performance (Figure 7.4) and variable selection (Figure 7.7). The differences were more subtle for simulations and were most notable for rare species, where MAXENT tended to underfit (bottom row of Figure 7.7) and hence coefficients were highly biased (bottom row of Figure 7.8) leading to slightly poorer predictive performance (bottom row of Figure 7.9). Comparing MAXENT to MSI as in Figure 6.1, λ chosen by MAXENT can often be expected to be too high for rare species. An additional difference that may be considered is that while MAXENT only uses the number of presences to determine the penalty, MSI uses the data in the form of $\lambda_{\max}$ for an initial estimate of sparsity. Despite the fact that $\lambda_{\max}$ is a rather crude estimate of sparsity, MSI was generally competitive with data-driven approaches which use sparsity directly in finding an optimal point along the bias-variance tradeoff, although these data-driven criteria were sometimes preferable for more abundant species, in particular, in situations where little shrinkage was required (Figure 7.9 top-left).

The results were generally consistent with other studies of the performance of the default MAXENT penalty. As also shown by Gastón & García-Viñas (2011) and Anderson & Gonzalez (2011), penalising parameter coefficients improved predictive performance, particularly for moderate or small levels of prevalence and sparsity (Figures 7.4 and 7.9). Moreover, the high predictive performance achieved by MAXENT's default penalty was matched and sometimes exceeded by other methods. Similar to the work of Warren & Seifert (2011), I found that the penalty for underfitting was stronger than for overfitting in terms of predictive performance (Figures 7.7 and 7.9). This is a well-known phenomenon (Hastie *et al.*, 2009).

In addition to proposing MSI as a criterion for choosing the LASSO penalty, there were three main points of difference in this Chapter from previous literature in ecology – implementing adaptive LASSO, including models that account for point interaction, and accounting for spatial autocorrelation using a spatial cross validation scheme. I will discuss new insights from these extensions in turn.

Applying an adaptive LASSO penalty instead of a LASSO penalty did not generally result in any appreciable improvement in predictive performance for either the real data (Figure 7.5) or simulations (Figure 7.10), with the exception of MAXENT and non-linear GCV (to a lesser extent) for rare species when adaptive weights were chosen from the unpenalised model. Although these methods still underfitted the data (top row of Figure 7.10), there was a reduction in bias from overshrinking (middle row of Figure 7.10), resulting in a slight improvement in predictive performance (bottom row of Figure 7.10).

Results for area-interaction models suggest that conventional thought on the advantages of implementing a LASSO penalty may not apply. When the interaction parameter was strong and data were sparse, applying a LASSO penalty led to substantially poorer predictive performance (Figure 7.11b). This appeared to also be the case in analysis of the Blue Mountains dataset, where predictive performance deteriorated rapidly for all LASSO approaches under increasing interaction radius (Figure 7.6, right). The inclusion of an interaction term itself shrank the coefficients of the environmental variables, which is some cases obviated the need for a LASSO penalty at all. On the other hand, when data were prevalent (Figure 7.11a) or when interactions were weak (Figure 7.12), there were still gains to be made from using the LASSO. Perhaps this issue could be resolved by using a different method to choose the LASSO penalty which was more attentive to the presence of point interactions.

Traditional applications of cross validation are known to be prone to optimistic estimates of performance when predicting in other locations (Wenger & Olden, 2011). Incorporating a spatial cross validation scheme in this Chapter, as opposed to randomly selecting points for cross validation, therefore ensured that methods which tended to overfit the data would be accordingly penalised. However, as MAXENT tended to underfit, the results of this Chapter would present the MAXENT penalty in a more favourable light, if anything. Hence its poor performance, particularly for rare species, illustrates the drawback of the *ad hoc* nature with which it was derived.

Chapter 8

The `ppmlasso` `R` package

In Chapter 5, I established the equivalence of MAXENT and Poisson PPMs and demonstrated a number of advantages incurred from expressing the model as a PPM. In Chapter 7, I found that applying a LASSO penalty can improve predictive performance. While there are `R` packages available to fit PPMs (Baddeley & Turner, 2005) and packages available to fit models with a LASSO penalty (Friedman *et al.*, 2010), there are no packages that fit a regularisation path of PPMs with a LASSO penalty. I have developed the `ppmlasso` package for this purpose, and in this Chapter I describe its functions and features.

Starting with a list of locations of *Corymbia eximia* (`sp.xy`), a matrix of environmental data at the $500m \times 500m$ resolution (`backg`), and a binary matrix of locations throughout the study region indicating whether a location is available or not (`availability`), I will demonstrate the functions of the `ppmlasso` package. With these functions, it is possible to:

- Prepare data for model fitting given a geo-referenced grid of environmental data and a list of species presence locations. This involves creating a regular grid of quadrature points at the desired spatial resolution (`sample.quad`), extracting environmental data to species presence locations (`env.var`), and constructing the design matrix and observation weights used in the model (`ppm.dat`).

- Calculate point interactions for use in an area-interaction model (`point.interactions`).
- Construct a regularisation path of Poisson PPMs or area-interaction models and choose an optimal model based on the various criteria described in Chapter 7 (`ppmlasso`).

Section 8.1 describes the main features of the `ppmlasso` function, while Section 8.2 describes functions that are useful for reanalysis. Detailed descriptions of the arguments and output of each function are presented in Appendix B.

8.1 Fitting a Regularisation Path of Point Process Models: `ppmlasso`

The `ppmlasso` function fits a regularisation path of Poisson PPMs or area-interaction models. Users can control the type of penalty (LASSO, adaptive LASSO or elastic net), the number of fitted models in the regularisation path, and the criterion to be method of choosing the LASSO penalty λ. It relies on the `sample.quad`, `env.var`, `ppm.dat`, and (for area-interaction models) `point.interactions` functions to set up the design matrix. These functions are described in detail in Section 8.2.

In Chapters 5 and 7, I fitted Poisson PPMs and area-interaction models to *C. eximia* using minimum and maximum annual temperature (MNT and MXT), annual rainfall (Rain), the number of fires recorded since 1943 (FC) as well as distance from main roads (D.Main) and distance from urban areas (D.Urb) to account for observer bias. The design matrix consisted of a quadratic function of all variables, including interactions between the four environmental variables and between the two accessibility variables. To set up the design matrix and response vector, the command was:

```
> ppm.form = Pres/wt ~ cbind(poly(FC, MNT, MXT, Rain, degree = 2),
poly(D.Main, D.Urb, degree = 2)).
```

The command for fitting a regularisation path of 200 Poisson PPMs and choosing the model that minimised non-linear GCV as in Chapter 5 was:

```
> species.fit = ppmlasso(ppm.form, sp.xy = sp.xy, env.grid = backg,
sp.scale = 0.8, criteria = "nlgcv").
```

`ppmlasso` can also fit regularisation paths of other types of model with the inclusion of a few additional arguments:

- **Area-interaction models**: Set `family = "area.inter"` and provide an interaction radius `r` (*e.g.* `r = 5` for *C. eximia*).
- **Adaptive LASSO**: Provide the coefficients `init.coef` and exponent `gamma` to calculate the adaptive weights (*e.g.* `init.coef = species.fit$beta` and `gamma = 1`).
- **Elastic Net**: Provide `alpha` for an elastic net penalty of the form $\alpha\lambda\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\lambda\sum_{j=1}^{p}(\beta_j)^2$. Note that the default of `alpha` $= 1$ fits a LASSO penalty, while `alpha` $= 0$ fits a ridge regression penalty.

A description of all arguments of `ppmlasso` appears in Appendix B.

Table 8.1 contains the parameters of Poisson PPMs fitted with LASSO, adaptive LASSO, and elastic net penalties as well as area-interaction models that minimise non-linear GCV. Note that the coefficients tend to be smaller for the AI-LASSO model - this is consistent with the results of Chapter 7 where I found that area-interaction models tended to underfit.

Each model in the regularisation path is fitted by extending the Osborne descent algorithm described in Section 4.3 to GLMs with penalised iteratively reweighted

Table 8.1: Coefficients of the fitted models that minimise non-linear GCV. The PPM-LASSO and AI-LASSO models were fitted with LASSO penalties to Poisson PPMs and area-interaction models, respectively. The Adaptive model was fitted with an adaptive LASSO penalty with adaptive weights $w_j = 1/|\hat{\beta}_{\text{NLGCV},j}|$. The Elastic Net model was fitted with $\alpha = 0.7$.

Parameter	**PPM-LASSO**	**AI-LASSO**	**Adaptive LASSO**	**Elastic Net**
`Intercept`	-15.12	-9.72	-16.38	-14.62
`FC`	0.56	0.22	0.43	0.58
`FC`^{2}	-0.41	-0.17	-0.34	-0.43
`MNT`	2.93	0.21	2.89	2.84
`FC*MNT`	0.02	-0.09	0	0
`MNT`^{2}	-3.06	-0.22	-2.87	-3.29
`MXT`	5.62	0	8.41	4.52
`FC*MXT`	0.04	-0.03	0	0.12
`MNT*MXT`	0.44	-0.39	0	1.15
`MXT`^{2}	-7.53	-0.12	-10.12	-7.13
`Rain`	-0.27	-0.01	0	-1.00
`FC*Rain`	-0.26	-0.20	-0.13	-0.20
`MNT*Rain`	1.18	0	0.81	1.92
`MXT*Rain`	-6.90	0	-9.33	-6.75
`Rain`^{2}	-3.68	0	-4.44	-3.80
`D.Main`	-0.46	0	-0.58	-0.43
`D.Main`^{2}	0.29	0.38	0.11	0.31
`D.Urb`	-0.54	-0.15	-0.39	-0.58
`D.Main*D.Urb`	-0.16	-0.12	0	-0.19
`D.Urb`^{2}	0.18	0.14	0.08	0.19
`Interaction`	NA	1.47	NA	NA

least squares. As in Section 4.3, let $\boldsymbol{\sigma}$ denote the indices of the nonzero parameters. Given a current estimate $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\sigma}}^{i-1}$ with corresponding fitted intensities $\widehat{\boldsymbol{\mu}}^{i-1}$, the proposed update in Step 1 is $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\sigma}}^{i} = (\mathbf{X}_{\boldsymbol{\sigma}}'\mathbf{W}^{i-1}\mathbf{X}_{\boldsymbol{\sigma}})^{-1}[\mathbf{X}_{\boldsymbol{\sigma}}'\mathbf{W}^{i-1}\{\mathbf{z} - \lambda \text{sign}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\sigma}}^{i-1})\}]$, where for a Poisson GLM, $\mathbf{W}^{i-1} = \text{diag}(\mathbf{w}\widehat{\boldsymbol{\mu}}^{i-1})$, $\mathbf{w}$ is a vector of quadrature weights, and $\mathbf{z} = \ln \widehat{\boldsymbol{\mu}}^{i-1} + (\mathbf{y} - \widehat{\boldsymbol{\mu}}^{i-1})/\widehat{\boldsymbol{\mu}}^{i-1}$. The rest of the algorithm proceeds in the same way, with the score equations of Step 4 calculated as $s(\widehat{\boldsymbol{\beta}}^{i}; \mathbf{y}_P) = \mathbf{X}'\mathbf{W}^{i}(\mathbf{y} - \widehat{\boldsymbol{\mu}}^{i})/\widehat{\boldsymbol{\mu}}^{i}$.

8.2 Functions for Reanalysis

In order to fit regularisation paths, `ppmlasso` calls on other functions to create a set of quadrature points, extract environmental data to species locations, and set up a data matrix with observation weights. For a path of area-interaction models, point interactions are also calculated. These can be time-consuming processes, so for analyses in which some of these steps are repeated (*e.g.* interpolating environmental variables for building models of the same species at different spatial resolutions), it is worthwhile to use these functions independently of `ppmlasso`. This way, their output can be directly supplied to `ppmlasso` without the need to repeat these preliminary steps.

8.2.1 Generating Quadrature Points: `sample.quad`

The `sample.quad` function creates a matrix of quadrature points and associated environmental data for a user-defined spatial resolution. It requires only a georeferenced matrix of environmental grids. The `sample.quad` function exploits the fact that environmental data usually come at locations along regular grids by quickly subsetting the reference matrix selecting the rows that coincide with the nominated spatial resolution. To generate a set of quadrature points `quad.1` at a spatial resolution of 1 km × 1 km from the matrix `backg` and save the output as the file `"Quad1.RData"`, the command is:

```
> quad.1 = sample.quad(env.grid = backg, sp.scale = 1, file = "Quad").
```

Supplying a vector of resolutions will generate matrices of quadrature points for all elements of `sp.scale`:

```
> quad.1 = sample.quad(env.grid = backg, sp.scale = c(1, 2, 3, 4, 5),
file = "Quad").
```

8.2.2 Interpolating Environmental Data to Species Locations: `env.var`

Given a matrix of quadrature points and a list of species presences, the function `env.var` extracts environmental data to presence locations using bilinear interpolation. Details of how bilinear interpolation is implemented are in Appendix B.

To generate the environmental data `species.env` for the *C. eximia* presence locations stored in the matrix `sp.xy` using a spatial resolution of $500m$ and quadrature points stored in `backg` and save the output in the file `"C Eximia Env.RData"`, the command is:

```
> species.env = env.var(sp.xy, env.grid = backg, env.scale = 0.5,
file.name = "C Eximia Env").
```

8.2.3 Setting Up the Data for Model Fitting: `ppm.dat`

The `ppm.dat` function prepares the data for model fitting. In particular, the `ppm.dat` function calculates observation weights (Chapter 3) and returns a matrix `dat.ppm` ready for use in the `ppmlasso` function for fitting a regularisation path of Poisson PPMs or area-interaction models.

To set up a design matrix `species.ppm` from species locations stored in `sp.xy` at a spatial resolution of $1km$ using background grid of environmental data `backg` and save into the file `"C Eximia PPM.RData"`, the command is:

```
> species.ppm = ppm.dat(sp.xy = sp.xy, env.grid = backg, sp.scale = 1,
file.name = "C Eximia PPM").
```

The above call of the `ppm.dat` function will apply the `sample.quad` and `env.var` functions to generate quadrature points and interpolate environmental data. Alternatively, the user can input files containing both the species data and quadrature points to obviate the need to call these functions:

```
> species.ppm = ppm.dat(sp.xy = "C Eximia", sp.scale = 1,
quad.file = "Quad", file.name = "C Eximia PPM").
```

8.2.4 Calculating Point Interactions: `point.interactions`

In order to fit an area-interaction model, point interactions must first be calculated at both presence locations and quadrature points using the `point.interactions` function. The `point.interactions` function requires only a data matrix generated using the `ppm.dat` function and the radius `r` of interactions. For study regions that have inaccessible areas (*e.g.* urban areas or ocean), the user may also supply a binary matrix called `availability` which indicates whether locations are available (`availability = 1`) or not (`availability = 0`). If not supplied, `availability` is automatically generated at a spatial resolution equal to half of the radius `r`, with all values set to 1. A detailed description of how point interactions are estimated is in Appendix B.

To calculate point interactions of radius $r = 5$ km for *C. eximia* at the locations in the matrix `species.ppm` using availability matrix `availability`, the command is:

```
> species.int = point.interactions(species.ppm, 5, availability).
```

This function was used to calculate point interactions in Section 3.2 and Chapters 5 and 7.

8.2.5 Example Use of Functions for Reanalysis

Consider the four fitted regularisation paths of Section 8.1. Although each of these paths used the same data, each call to `ppmlasso` in Section 8.1 used the species locations `sp.xy` and grid of environmental data `backg` to generate the same matrix `data` using the `ppm.dat` function. A more time-efficient strategy is to generate this matrix first with a single call to the `ppm.dat` function and supply it to each call to the `ppmlasso` function, as follows:

```
> species.ppm = ppm.dat(sp.xy = sp.xy, env.grid = backg,
sp.scale = 0.8, file.name = "C Eximia PPM").
> lasso.fit = ppmlasso(ppm.form, data = species.ppm, criteria = "nlgcv").
> ad.lasso.fit = ppmlasso(ppm.form, data = species.ppm, criteria = "nlgcv",
init.coef = lasso.fit$beta, gamma = 1).
> e.net.fit = ppmlasso(ppm.form, data = species.ppm, criteria = "nlgcv",
alpha = 0.7).
> ai.fit = ppmlasso(ppm.form, data = species.ppm, criteria = "nlgcv",
family = "area.inter", r = 5).
```

To generate the `species.ppm` matrix in the first line of code above required 3.75 minutes of computation time. Hence supplying `species.ppm` to the `data` argument of the four `ppmlasso` function calls of Section 8.1, rather than calculating it for each call, saved 11.25 minutes of computation time.

Computation gains can likewise be made in other instances of repeated analysis. When fitting models to multiple species at the same spatial resolution, the same set of quadrature points needed for each species can be generated with a single call to the `sample.quad` function. When fitting models to the same species at different spatial resolutions (*e.g.* to create the likelihood plot in Figure 5.4a), the species environmental data required at each resolution can be interpolated just once with a single call to the `env.var` function. When fitting multiple area-interaction model regularisation paths for the same species and spatial resolution, the vector of point interactions necessary for each path can be calculated through a single call to the `point.interactions` function.

8.3 Conclusion

With the `ppmlasso` package, users can now fit PPMs that implement LASSO penalties optimised by appropriate methods in a single function call. At the time of writing, the `ppmlasso` package contains functions that can reproduce all of the analysis of the entire thesis with the exception of the simulations performed in Chapter 7 and the goodness-of-fit tests and plots in Chapters 3 and 5, which were performed using `spatstat`. I aim to make the `ppmlasso` package dependent on `spatstat` so that the fitted models may be diagnosed, plotted, and simulated without further user manipulation.

Chapter 9

Discussion

9.1 Summary

The work presented in this thesis stands at the intersection of statistics and ecology and has used strengths of both fields to improve the way presence-only analysis is done in practice. In Chapter 5 I linked MAXENT, one of the most popular methods in use today for presence-only analysis in ecology, and Poisson PPMs (Chapter 3), a method with a long history in statistics but only recently proposed for SDM. This equivalence extends the benefits of PPMs to users of MAXENT, allowing them to think more critically about the models they fit. The point process framework facilitates the data-based method of choosing the spatial resolution and opens up MAXENT to a suite of diagnostic tools (Cressie, 1993; Diggle, 2003; Baddeley & Turner, 2005; Baddeley *et al.*, 2005) to assess model assumptions.

The equivalence result also extends benefits to the way PPMs are currently fitted. MAXENT implements a LASSO penalty (Chapter 4), which has been shown to improve predictive performance. This motivates the use of shrinkage methods such as LASSO with PPMs to improve their ability to predict to new data (Chapter 7). In Chapter 6, I established an asymptotic result for Poisson PPMs that offers insight into how the LASSO penalty could be chosen. I noted that MAXENT's default penalty increases too quickly to satisfy the conditions of $\sqrt{m}$-consistency. This result motivated the development of a novel criterion for determining the LASSO penalty, MSI, which I demonstrated in Chapter 7 to be generally superior to MAXENT's default penalty in an extensive comparison with many assorted criteria. Data-driven methods such as BIC and non-linear GCV performed well for more prevalent species.

Hence I have used the equivalence result of Chapter 5 to combine the advantages inherent to Poisson PPMs and the application of a LASSO penalty to boost predictive performance. The outcome is PPM-LASSO, a presence-only SDM method with rigorous statistical foundations and competitive predictive performance. I also developed the `ppmlasso` package (Chapter 8) for `R`, which means that all of the advantages of this cross-disciplinary work can be realised by practitioners.

One particular advantage of PPM-LASSO is that it is flexible and can adapt to key properties of the data. Rather than coerce the data into a pre-determined set of rules for fitting and regularising the model as in the case of MAXENT, PPM-LASSO uses the data to inform the choice of model, spatial resolution, and LASSO penalty.

9.2 Future Extensions

This work has the potential to be extended in a number of ways.

Elith *et al.* (2006) found that methods that can be applied to multiple species simultaneously ("communities") often performed better than when applied to species individually, such as a community-level implementation of MARS (Elith & Leathwick, 2007). Hence, one opportunity is to extend PPM-LASSO to the community level. Borrowing the strength of multiple species to inform the predicted distribution of single species could further amplify the benefits of PPM-LASSO demonstrated in this thesis, in particular in modelling observer bias, whose contributing factors are likely to be similar across species.

The asymptotic results established in Chapter 6 could be used to cultivate new methods for choosing the LASSO penalty. The MSI penalty is one such example, based on the maximum score of the intercept model. However, any criteria that is of order $O(\sqrt{m})$ or smaller could be posed as an alternative. Indeed, MSI exhibits a tendency to overshrink for moderate and abundant species. Perhaps a more modest penalty of order $O(\sqrt{m})$ such as the mean score of the intercept model may yield even better performance.

An additional extension is in considering how other classical fixed-n asymptotic results extend to the PPM framework. For example, traditional criteria for choosing the LASSO penalty such as BIC were developed for designs with a fixed sample size n. As in Chapter 6, an asymptotic approach could be used to assess whether these criteria should be modified for PPMs, and whether the "replace n with m" result of Chapter 6 applies more generally.

The framework of the simulations performed in Chapter 7 could be used to motivate and develop variable selection strategies for area-interaction models and other Gibbs processes. The `ppmlasso` package is the first package that permits regularisation of area-interaction models, and the simulation results presented in Chapter 7 and Appendix A suggest that traditional methods of choosing the LASSO penalty may be too harsh without some modification.

I have addressed the challenge of spatial autocorrelation using Gibbs processes, and by implementing a spatial cross-validation scheme in model validation, but there are other strategies. Hierarchical approaches such as Chakraborty *et al.* (2011) can also be used to offset spatial autocorrelation. A comparison of various methods of accounting for spatial autocorrelation could further improve the application of PPM-LASSO.

It is commonplace in SDM to publish results demonstrated solely through analysis of real and/or simulated data (Anderson & Gonzalez, 2011; Gastón & García-Viñas, 2011; Warren & Seifert, 2011). The work of this thesis, however, has been enabled through a close examination of theory underpinning the methods of analysis in addition to analysis of real and simulated data. The methodological advances presented in this thesis illustrate the merit of this approach. Such a process is not only applicable to species distribution modelling, and indeed is a natural way to drive methodological research in other contexts.

Appendix A

Extended Simulation Results of Adaptive LASSO and Area-Interaction Models

This Appendix provides further simulation results relevant to comparing methods for choosing the LASSO penalty described in Chapter 7.

A.1 LASSO

Figure A.1 depicts the proportion of shrinkage applied by BIC, MAXENT, non-linear GCV, and MSI at different levels of sparsity (columns) and abundance (rows). This proportion $q(\widehat{\boldsymbol{\beta}})$ is defined as in Section 7.2. Because the true coefficients $\boldsymbol{\beta}_*$ are known, $q(\boldsymbol{\beta}_*)$ provides a measure of how much shrinkage is ideal.

For full models whereby no variables should be shrunk to zero (left column), all methods except BIC shrank too much for moderate and super rare abundances. As abundance increases, non-linear GCV did not overshrink as much, while MAXENT and MSI continued to shrink too much regardless of abundance. For moderate and sparse models (middle and right columns), each method also applied less shrinkage as

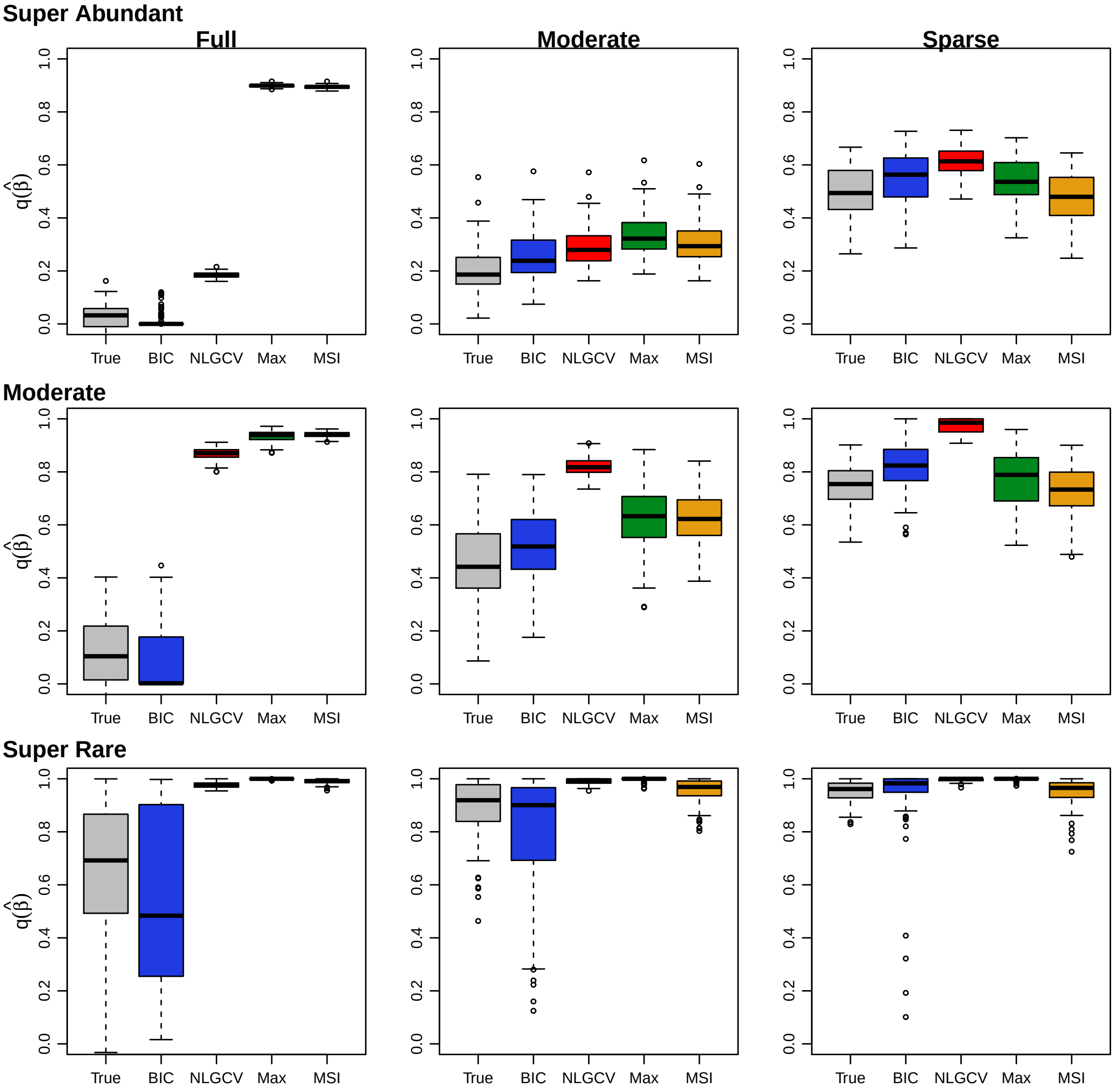

Figure A.1: Proportion of shrinkage of different methods of choosing the LASSO penalty. MAXENT tended to overfit while non-linear GCV tended to oversrhink. BIC and MSI generally applied an optimal amount of shrinkage, although MSI overshrank somewhat for moderately sparse models.

---

abundance increased, with MAXENT and non-linear GCV generally overshrinking, while BIC and MSI came closest to the true level of shrinkage required, although MSI tended to overshrink somewhat for moderately sparse models.

A.2 Area-Interaction Models

This Section contains results of simulations from an area-interaction model when data were fitted using both Poisson PPMs and area-interaction models.

Figures A.2 and A.3 compare the selection of environmental variables when the interaction parameter is strong and weak, respectively. Regardless of the strength of the interaction coefficient or abundance level, area-interaction models tended to falsely exclude more environmental variables and falsely include fewer environmental variables, indicating higher shrinkage than Poisson PPMs. Rare species likewise tended to falsely exclude more environmental variables and falsely include fewer environmental variables than abundant species. In contrast to when the true model is a Poisson PPM, non-linear GCV tended to shrink the least.

Figures A.4 and A.5 compare the bias, variance, and mean squared error for when the interaction parameter is strong and weak, respectively. Fitting an area-interaction model generally led to a reduction of overall mean squared error regardless of the method of choosing λ. This was generally achieved by a reduction in bias. By not including an interaction coefficient, Poisson PPMs overly inflated coefficients to account for the spatial trend due to point interactions.

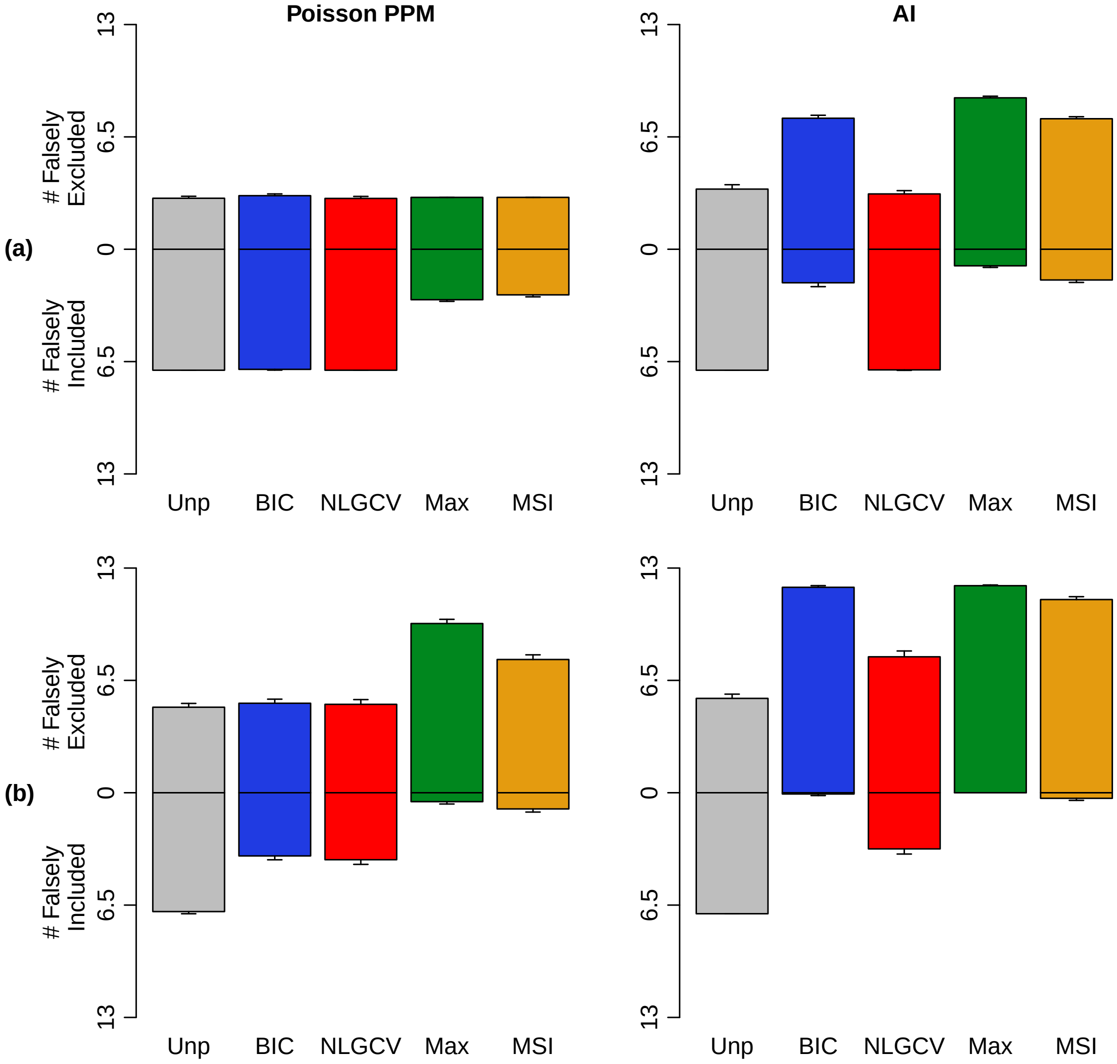

Figure A.2: Variable selection when a Poisson PPM and area-interaction model are both fitted to data generated from a moderately sparse area-interaction model with strong point interactions. Correctly fitting an area-interaction model tended to reduce the amount of overfitting for (a) super abundant and (b) super rare species.

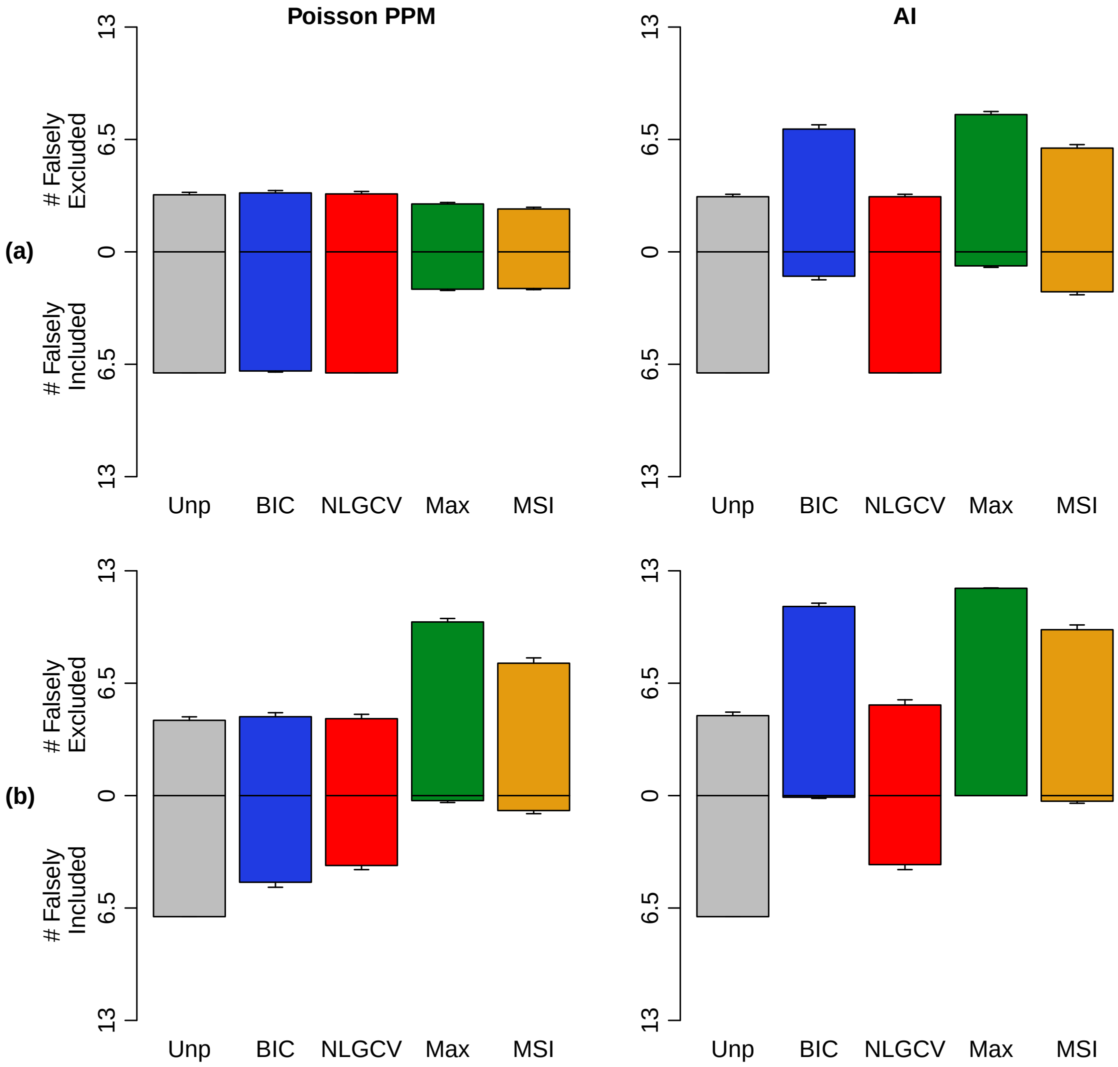

Figure A.3: Variable selection when a Poisson PPM and area-interaction model are both fitted to data generated from a moderately sparse area-interaction model with weak point interactions. Correctly fitting an area-interaction model tended to reduce the amount of overfitting for (a) super abundant and (b) super rare species.

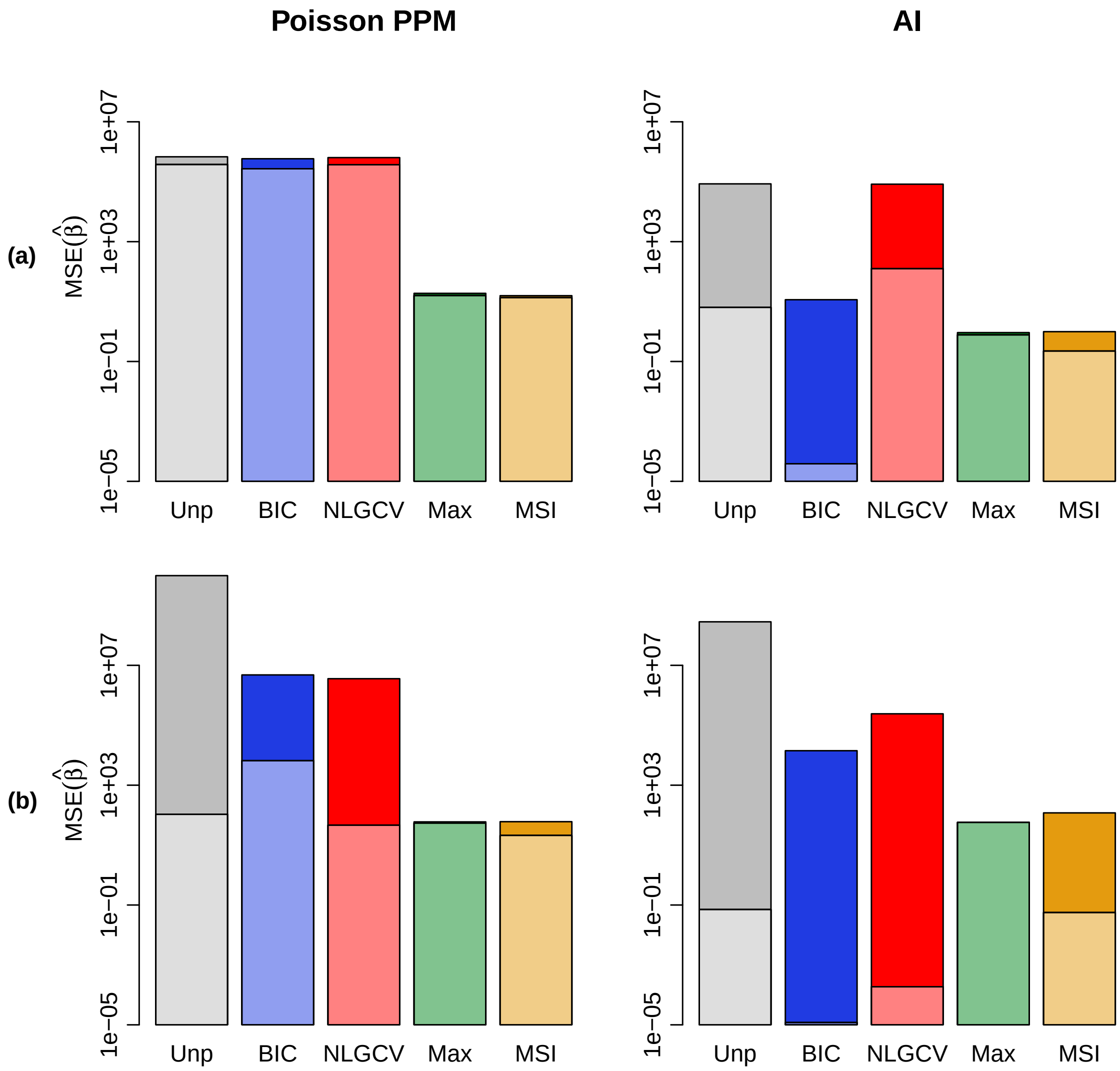

Figure A.4: Mean squared error when a Poisson PPM and area-interaction model are both fitted to data generated from a moderately sparse area-interaction model with strong point interactions. Correctly fitting an area-interaction model reduced bias and therefore reduced overall MSE for (a) super abundant and (b) super rare species.

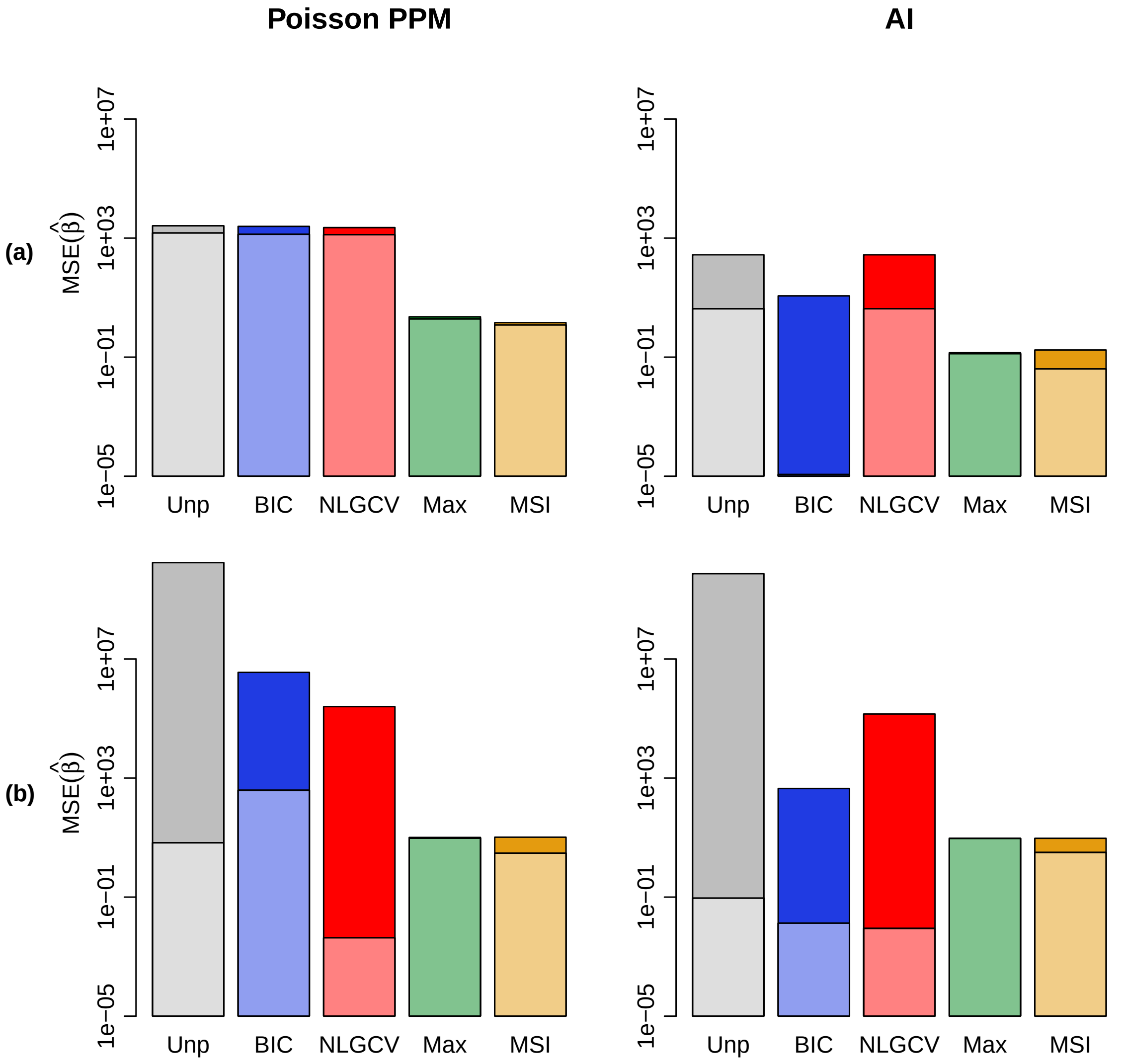

Figure A.5: Mean squared error when a Poisson PPM and area-interaction model are both fitted to data generated from a moderately sparse area-interaction model with weak point interactions. Correctly fitting an area-interaction model reduced bias and therefore reduced overall MSE for (a) super abundant and (b) super rare species.

Appendix B

Details of `ppmlasso` package functions

B.1 `ppmlasso`

B.1.1 Description

The `ppmlasso` function fits a regularisation path of Poisson PPMs or area-interaction models using the `single.lasso` function with either a sequence of LASSO, adaptive LASSO or elastic net penalties.

B.1.2 Usage

```
ppmlasso(formula, sp.xy, env.grid, sp.scale, data, lamb = NA, n.fits =
200, criteria = "bic", family = "poisson", r = NA, interactions = NA, ...)
```

B.1.3 Arguments

`formula` The formula of the fitted model. For a PPM, the correct form is `Pres/wt` $\sim$ `variables`.

`sp.xy` A matrix of species locations $\mathbf{y}_P$ containing at least one column representing longitude and one column representing latitude. Environmental variables are interpolated to the locations of `sp.xy` using the `env.var` function, unless the `data` argument is supplied.

`env.grid` The geo-referenced matrix of environmental grids. This matrix is used to generate quadrature points using the `sample.quad` function, interpolate environmental data to the species locations of `sp.xy` using the `env.var` function, and calculate observation weights using the `ppm.dat` function, unless the `data` argument is supplied. This creates a data matrix `data` which provides the variables for the `formula` argument.

`sp.scale` The spatial resolution at which to define the regular grid of quadrature points. `sample.quad` will subsample from the rows of `data` that coincide with a regular grid at a resolution of `sp.scale`.

`data` An optional data matrix generated from the `ppm.dat` function. Supplying a matrix to `data` is an alternative way of providing the environmental variables used in the `formula` argument, instead of specifying `sp.xy` and `env.grid`.

`lamb` A vector of penalty values that will be used to create the regularisation path. If `lamb = NA`, `ppmlasso` automatically determines the penalty values from the data and the `n.fits` argument.

`n.fits` The number of models fitted in the regularisation path. If `lamb = NA`, the `n.fits` penalty values will be equally spaced on a logarithmic scale from e^{-10} to $\lambda_{\max}$, the smallest penalty that shrinks all parameter coefficients to zero.

`criteria` The penalisation criteria to be optimised by the regularisation path. The options include `"aic"`, `"bic"`, `"hqc"`, `"gcv"`, `"nlgcv"` and `"msi"`, all of which are described in Chapter 7.

`family` The family of models to be fitted – `family = "poisson"` for Poisson PPMs or `family = "area.inter"` for area-interaction models.

`r` The radius of point interactions, required if `family = "area.inter"`.

`interactions` A vector of point interactions calculated from the `point.interactions` function necessary for fitting area-interaction models. If `interactions = NA` and `family = "area.inter"`, point interactions will be automatically calculated for radius `r` to the locations of `data`.

... Further arguments passed to the `single.lasso`, `sample.quad`, and `point.interactions` functions.

B.1.4 Value

The output of `ppmlasso` is a list with the following components:

`betas` A matrix of fitted coefficients of the `n.fits` models.

`lambdas` A vector containing the `n.fits` penalty values.

`likelihoods` A vector containing the likelihood of `n.fits` fitted models.

`pen.likelihoods` A vector containing the penalised likelihood of `n.fits` fitted models.

`beta` A vector containing the coefficients of the model that optimises the criteria specified by the `criteria` argument.

`lambda` The penalty value of the model that optimises the criteria specified by the `criteria` argument.

`likelihood` The likelihood of the model that optimises the criteria specified by the `criteria` argument.

`mu` A vector of fitted values from the model that optimises the criteria specified by the `criteria` argument.

`criteria.matrix` A matrix with `n.fits` rows corresponding to the observed values of AIC, BIC, HQC, GCV, and non-linear GCV.

`family` The `family` argument supplied to `ppmlasso`.

`criteria` The `criteria` argument supplied to `ppmlasso`.

B.2 `single.lasso`

B.2.1 Description

The `single.lasso` function fits a single LASSO-regularised model using the descent algorithm of Osborne *et al.* (2000*b*) and passes information to the `ppmlasso` function.

B.2.2 Usage

```
single.lasso(max.it = 25, tol = 1.e-9, gamma = 0, init.coef = NA, alpha
= 1, mu.min = 1.e-16, mu.max = 1.e16, standardise = TRUE, ...)
```

B.2.3 Arguments

`max.it` The maximum number of iterations of the descent algorithm for fitting the model.

`tol` The convergence threshold for the descent algorithm. The algorithm continues for a maximum of `max.it` iterations until the difference in likelihood between successive fits falls below `tol`.

`gamma` The exponent of the adaptive weights for the adaptive LASSO penalty. The default value `gamma = 0` corresponds to a normal LASSO penalty.

`init.coef` The initial coefficients used for an adaptive LASSO penalty.

`alpha` The elastic net parameter. The form of the penalty is $\alpha\lambda\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\lambda\sum_{j=1}^{p}(\beta_j)^2$. The default value `alpha = 1` corresponds to a LASSO penalty, while `alpha` $= 0$ corresponds to a ridge regression penalty.

`mu.min` The threshold for small fitted values. Any fitted value less than the threshold is set to `mu.min`.

`mu.max` The threshold for large fitted values. Any fitted value larger than the threshold will be set to `mu.max`.

`standardise` A logical argument indicating whether the environmental variables should be standardised to have mean 0 and variance 1.

... Other arguments inherited from `ppmlasso`.

B.2.4 Value

The output of `single.lasso` is a list with the same components as `ppmlasso`.

B.3 sample.quad

B.3.1 Description

The `sample.quad` function creates a matrix of quadrature points at a given spatial resolution.

B.3.2 Usage

```
sample.quad(env.grid, sp.scale, coord = c("X", "Y"), file = "Quad")
```

B.3.3 Arguments

`env.grid` The geo-referenced matrix of environmental grids, as in `ppmlasso`.

`sp.scale` The spatial resolution at which to sample quadrature points, as in `ppmlasso`.

`coord` A vector containing the names of the longitude and latitude coordinates.

`file` An optional argument containing the prefix of the name of the saved file. The default is `"Quad"` so that a matrix generated at a spatial resolution of 1 would be saved in the file `"Quad1.RData"`. A file is saved for every resolution given in `sp.scale`.

B.3.4 Value

The output of `sample.quad` is a matrix of quadrature points at the spatial resolution supplied to `sp.scale`. If a vector of resolutions is supplied, the output is a list of file names containing the saved matrices of quadrature points stored as `dat.quad`.

B.4 `env.var`

B.4.1 Description

The `env.var` function uses bilinear interpolation to extract environmental data from a matrix of quadrature points to a list of species locations.

B.4.2 Usage

```
env.var(sp.xy, env.grid, env.scale, coord = c("X", "Y"), file.name = NA)
```

B.4.3 Arguments

`sp.xy` A matrix of species locations containing at least one column representing longitude and one column representing latitude, as in `ppmlasso`.

`env.grid` The geo-referenced matrix of environmental grids, as in `ppmlasso`.

`env.scale` The spatial resolution used for interpolating environmental data. At a given species location, the environmental data will be interpolated from the four points that form a square of side length `env.scale` that contains the location (Figure B.1).

`coord` A vector containing the names of the longitude and latitude coordinates, as in `sample.quad`.

`file.name` An optional argument containing the name of the saved file. Setting `file.name = "Sp Env"` will save a matrix `sp.dat` containing the species presence locations and the interpolated environmental data to the file `"Sp Env.RData"`.

B.4.4 Details

At a given species location with coordinates (x, y), the interpolated value of the environmental variable z is calculated as a weighted average of z at four reference quadrature points $(x^{(1)}, y^{(1)})$, $(x^{(1)}, y^{(2)})$, $(x^{(2)}, y^{(1)})$ and $(x^{(2)}, y^{(2)})$ that form a square of nominated side length `env.scale` surrounding (x, y) (Figure B.1). Each reference weight is then calculated as the area of a rectangle with diagonal formed by the reference point and the species location:

$$w(x^{(i)}, y^{(j)}) = |(x^{(i)} - x)(y^{(j)} - y)|. \tag{B.1}$$

Hence the interpolated value $z(x, y)$ is:

$$z(x, y) = \frac{\sum_{i=1}^{2}\sum_{j=1}^{2} w(x^{(i)}, y^{(j)}) z(x^{(i)}, y^{(j)})}{\sum_{i=1}^{2}\sum_{j=1}^{2} w(x^{(i)}, y^{(j)})}.$$

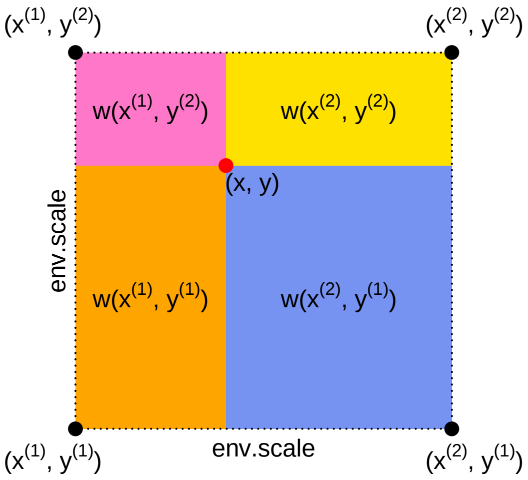

Figure B.1: Bilinear interpolation of environmental data at a species location with coordinates (x, y) is a weighted average of the environmental variable at four reference quadrature points $(x^{(1)}, y^{(1)})$, $(x^{(1)}, y^{(2)})$, $(x^{(2)}, y^{(1)})$ and $(x^{(2)}, y^{(2)})$. The weight of each reference point (Equation B.1) is equal to the area of a rectangle with diagonal formed by the reference point and the species location.

B.4.5 Value

The output of `env.var` is a matrix containing locations of species presences in the first two columns and the interpolated environmental data in the remaining columns.

B.5 `ppm.dat`

B.5.1 Description

The `ppm.dat` function calculates observation weights and prepares a data matrix for use in the `ppmlasso` function.

B.5.2 Usage

```
ppm.dat(sp.xy, env.grid, sp.scale, coord = c("X", "Y"), quad.file = NA,
file.name = NA)
```

B.5.3 Arguments

`sp.xy` A matrix of species locations $\mathbf{y}_P$ containing at least one column representing longitude and one column representing latitude, as in `ppmlasso`.

`env.grid` The geo-referenced matrix of environmental grids, as in `ppmlasso`.

`sp.scale` The spatial resolution at which to sample quadrature points, as in `ppmlasso`.

`coord` A vector containing the names of the longitude and latitude coordinates, as in `sample.quad`.

`quad.file` The name of a file containing the quadrature points created from the `sample.quad` function. If `quad.file = NA`, the `sample.quad` function is called to generate quadrature points at the nominated resolution of `sp.scale` from the `env.grid` matrix.

`file.name` An optional argument containing the name of the saved file, as in `env.var`.

B.5.4 Value

The output of `ppm.dat` is a matrix `dat.ppm` containing columns representing locations and their associated environmental data, a column `Pres` indicating whether a location is a presence location (`Pres = 1`) or quadrature point (`Pres = 0`), and a column `wt` of observation weights.

B.6 `point.interactions`

B.6.1 Description

The `point.interactions` functions calculates point interactions necessary to fit a regularisation path of area-interaction models.

B.6.2 Usage

```
point.interactions(dat.ppm, r, availability = NA)
```

B.6.3 Arguments

The `r` argument is the same as that in `ppmlasso`. The additional arguments are as follows:

`dat.ppm` A design matrix generated using the `ppm.dat` function.

`r` The radius of point interactions, as in `ppmlasso`.

`availability` An optional binary matrix used in calculating point interactions indicating whether locations are available (1) or not (0). If no such matrix is provided, `availability` is automatically generated with all values set to 1 at a special resolution of half of `r`. This is useful for study regions that have inaccessible areas due to the presence of water or urban areas.

B.6.4 Details

Theoretically, the point interaction $t(y)$ is calculated as the proportion of available area in a circular region Y of radius r centred at y that overlaps with circles of radius r centred at the other presence locations $\mathbf{y}_P \backslash \{y\}$, as in Section 3.2. The `point.interactions` function discretises the study region at the same spatial resolution as `availability` by defining the matrix `occupied`, a fine grid of locations spanning the study region initialised to zero. The values of `occupied` within a distance of `r` of each presence location $y \in \mathbf{y}_P$ are then augmented by 1, such that `occupied` then contains the total number of presence locations with which each grid location interacts. To prevent unavailable areas from being included in the calculation of point interactions, the values of `occupied` at grid locations for which `availability` $= 0$ are set to zero.

$t(y)$ is then estimated as the proportion of available grid locations within Y that overlap circular regions around other presence locations $\mathbf{y}_P \setminus \{y\}$:

$$t(y) = \begin{cases} \frac{\sum_{i \in Y} I(\texttt{occupied[i]}>0 \ \& \ \texttt{availability[i]}>0)}{\sum_{i \in Y} I(\texttt{availability[i]}>0)} & : y \notin \mathbf{y}_P \\ \frac{\sum_{i \in Y} I(\texttt{occupied[i]}>1 \ \& \ \texttt{availability[i]}>0)}{\sum_{i \in Y} I(\texttt{availability[i]}>0)} & : y \in \mathbf{y}_P \end{cases} \tag{B.2}$$

Figure B.2 illustrates how the point interaction $t(y)$ is estimated. Available land area is depicted in dark green while ocean area is represented by blue, discretised into available (orange and green dots) and unavailable (blue dots) grid locations. $t(y)$ is estimated according to (B.2) as the proportion of available grid locations within Y (the black circle) that are occupied (orange): $t(y) = 10/13$. Finer resolutions of the `availability` matrix will yield more precise estimates but at a cost of greater computation time.

B.6.5 Value

The output of `point.interactions` is a vector of point interactions corresponding to the locations contained in the `dat.ppm` argument.

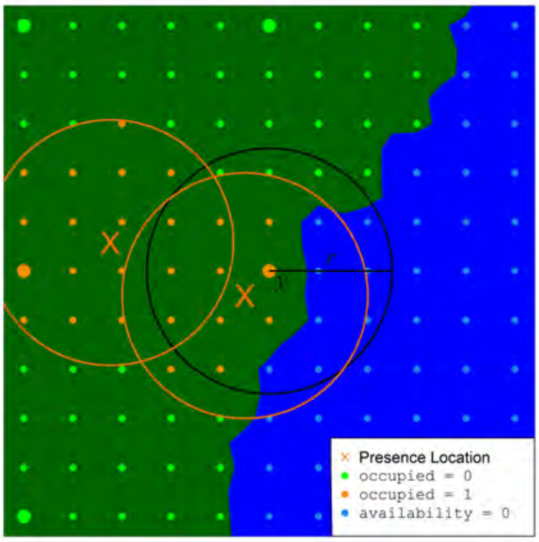

Figure B.2: Calculating point interactions with the `point.interactions` function. Land is represented by dark green and ocean is represented by blue. The point interaction $t(y)$ is the proportion of the land area within the black circle of radius r centred at y that overlaps the orange circles of radius r around the presence locations. In `ppmlasso`, the region is discretised and $t(y)$ is calculated according to Equation B.2. $t(y)$ is therefore estimated as the number of occupied and available grid locations (orange) within the black circle divided by the total number of available grid locations (orange and green) within the black circle: $t(y) = 10/13$.

Bibliography

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

Agterberg, F. P. (1974). Automatic contouring of geological maps to detect target areas for mineral exploration. *Journal of the International Association for Mathematical Geology* **6**, 373–395.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Anderson, R. P. & Gonzalez, I. (2011). Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with maxent. *Ecological Modelling* **222**, 2796–2811.

Assunção, R. & Guttorp, P. (1999). Robustness for inhomogeneous Poisson point processes *Annals of the Institute of Statistical Mathematics* **51**, 657–678.

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

Baddeley, A. J. & Silverman, B. W. (1984). A cautionary example on the use of second-order methods for analyzing point patterns. *Biometrics* **40**, 1089–1093.

Baddeley, A. J. & Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Australian and New Zealand Journal of Statistics* **42**, 283–322.

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

Baddeley, A. J. & Turner, R. (2006). Modelling spatial point patterns in R. In: Baddeley, A., Gregori, P., Mateu, J., Stoica, R., & Stoyan, D. (Eds.), *Case Studies in Spatial Point Pattern Modelling*, Lecture Notes in Statistics, pp. 23–74. Springer-Verlag, New York.

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

Beguin, J., Martino, S., Rue, H., & Cumming, S. G. (2012). Hierarchical analysis of spatially autocorrelated ecological data using integrated nested Laplace approximation. *Methods in Ecology and Evolution* **3**, 921–929.

Berman, M. (1986). Testing for spatial association between a point process and another stochastic process. *Journal of the Royal Statistical Society, Series C* **35**, 54–62.

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.

Besag, J., Milne, R, & Zachary, S. (1982). Point process limits of lattice processes. *Journal of Applied Probability* **19**, 210–216.

Bonham-Carter, G. (1994). *Geographic Information Systems for geoscientists: modelling with GIS*. Volume 13. Access Online via Elsevier.

Boyd, S. P. & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press, Cambridge, UK.

Breiman, L. (2001). Random forests. *Machine learning* **45**, 5–32.

Brillinger, D. R. (1978). Comparative aspects of the study of ordinary time series and of point processes. In: Krishnaiah, P. R. (Ed.), *Developments in Statistics*, pp. 33–133. Academic Press, New York, London.

Brix, A. & Møller, J. (2001). Space-time multitype log Gaussian Cox processes with a view to modelling weed data. *Scandinavian Journal of Statistics* **28**, 471–488.

Busby, J. R. (1991). BIOCLIM-a bioclimate analysis and prediction system. *Plant Protection Quarterly* **6**, 8–9.

Carpenter, G., Gillison, A. N., & Winter, J. (1993). DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation* **2**, 667–680.

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.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., & Silander Jr, J. A. (2010). Modeling large scale species abundance with latent spatial processes. *Annals of Applied Statistics* **4**, 1403–1429.

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

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377–403.

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.

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

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features on random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**, 380–393.

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

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

Dutta, M. (1966). On maximum (information-theoretic) entropy estimation. *Sankhya: The Indian Journal of Statistics, Series A* **28**, 319–328.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–451.

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

Elith, J. & Leathwick, J. R. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* **13**, 265–275.

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., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology* **77**, 802–813.

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.

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.

Fan, J. & Li, R. (2005). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of Statistical Planning and Inference* **131**, 333–347.

Fisher, E. (1992). A Skorohod representation and an invariance principle for sums of weighted i.i.d. random variables. *Rocky Mountain Journal of Mathematics* **22**, 169–179.

Fithian, W. & Hastie, T. (in review). Statistical models for presence-only data: Finite-sample equivalence and addressing observer bias. *arXiv preprint arXiv:1207.6950*.

Frank, I. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–135.

Franklin, J. (2009). *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press, Cambridge, UK.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

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

Gastón, A. & García-Viñas, J. I. (2011). Modelling species distributions with penalised logistic regressions: A comparison with maximum entropy models. *Ecological Modelling* **222**, 2037–2041.

Georgii, H.-O. (1976). Canonical and grand canonical Gibbs states for continuum systems. *Communications in Mathematical Physics* **48**, 31–51.

Good, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Annals of Mathematical Statistics* **34**, 911–934.

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.

Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* **41**, 190–195.

Harte, J., Smith, A. B., & Storch, D. (2009). Biodiversity scales from plots to biomes with a universal species–area curve. *Ecology Letters* **12**, 789–797.

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

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

Hengl, T., Sierdsema, H., Radovic, A., & Dilo, A. (2010). Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. *Ecological Modelling* **220**, 3499–3511.

Hernandez, P. A., Franke, I., Herzog, S. K., Pacheco, V., Paniagua, L., Quintana, H. L., Soto, A., Swenson, J. J., Tovar, C., Valqui, T. H., Vargas, J., & Young, B. E. (2008). Predicting species distributions in poorly-studied landscapes. *Biodiversity and Conservation* **17**, 1353–1366.

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and $\ell 1$ penalized regression: A review. *Statistics Surveys* **2**, 61–93.

Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology* **83**, 2027–2036.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

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

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Review* **106**, 620–630.

Jensen, J. L. & Møller, J. (1991). Pseudolikelihood for exponential family models of spatial point processes. *Annals of Applied Probability* **1**, 445–461.

Jensen, J. L. & Künsch, H. R. (1994). On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Annals of the Institute of Statistical Mathematics* **46**, 475–486.

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* **14**, 401–413.

Kelly, F. P. & Ripley, B. D. (1976). A note on Strauss's model for clustering. *Biometrika* **63**, 357–360.

Knight, K. & Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics* **28**, 1356–1378.

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley & Sons, New York.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* **5**, 369–411.

Lawson, A. (1988). On tests for spatial trend in a non-homogeneous Poisson process. *Journal of Applied Statistics* **15**, 225–234.

Lobo, J. M., Jimenez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography* **33**, 103–114.

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

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

NSW Office of Environment and Heritage (2010). Atlas of NSW Wildlife database. Data accessed 20/04/2010.

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

Osborne, M. R., Presnell, B., & Turlach, B. A. (2000*a*). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20**, 389–403.

Osborne, M. R., Presnell, B., & Turlach, B. A. (2000*b*). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9**, 319–337.

O'Sullivan, D. & Unwin, D. J. (2010). *Geographic Information Analysis*. Second edition. John Wiley & Sons, Hoboken.

Papangelou, F. (1974). The conditional intensity of general point processes and an application to line processes. *Probability Theory and Related Fields* **28**, 207–226.

Park, M. Y. & Hastie, T. (2007). L1-regularization algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B* **69**, 659–677.

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

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

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, R. (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, Austria. ISBN 3-900051-07-0.

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.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability* **13**, 255–266.

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

Royle, J. A., Chandler, R. B., Yackulic, C., & Nichols, J. D. (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution* **3**, 545–554.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In: Denison, D. D., Hansen, M. H., Holmes, C., Mallick, B., & Yu, B. (Eds.), *Nonlinear Estimation and Classification*, Lecture Notes in Statistics, pp. 149–172. Springer-Verlag, New York.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Shipley, B., Vile, D., & Garnier, E. (2006). From plant traits to plant communities: A statistical mechanistic approach to biodiversity. *Science* **314**, 812–814.

Stockwell, D. (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* **13**, 143–158.

Strauss, D. J. (1975). A model for clustering. *Biometrika* **62**, 467–475.

Thullier, W., Albert, C., Araújo, M. B., Berry, P., Cabeza, M., Guisan, A., Hicker, T., Midgely, G., Paterson, J., Schurr, F., Sykes, M., & Zimmermann, N. (2008). Predicting global change impacts on plant species distributions: Future challenges. *Perspectives in Plant Ecology, Evolution and Systematics* **9**, 137–152.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B* **67**, 91–108.

Tukey, J. W. (1972). Discussion of paper by F. P. Agterberg and S. C. Robinson. *Bulletin of the International Statistical Institute* **44**, 596.

Volkov, I., Banavar, J. R., Hubbell, S. P., & Maritan, A. (2009). Inferring species interactions in tropical forests. *Proceedings of the National Academy of Sciences* **106**, 13854–13859.

Waagepetersen, R. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics* **63**, 252–258.

Waagepetersen, R. (2008). Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika* **95**, 351–363.

Walker, P. A. & Cocks, K. D. (1991). HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters* **1**, 108–118.

Waller, L. A., Turnbull, B. W., Clark, L. C., & Nasca, P. (1992). Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and TCE-contaminated dumpsites in upstate New York. *Environmetrics* **3**, 281–300.

Warren, D. L. & Seifert, S. N. (2011). Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications* **21**, 335–342.

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.

Wenger, S. J. & Olden, J. D. (2011). 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**, 1670–1684.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B* **73**, 3–36.

Xanh, N. X. & Zessin, H. (1979). Integral and differential characterizations of the Gibbs process. *Mathematische Nachrichten* **88**, 105–115.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Yuan, M. & Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society, Series B* **69**, 143–161.

Zhang, Y., Li, R., & Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* **105**, 312–323.

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**, 2541–2563.

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.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the degrees of freedom of the lasso. *Annals of Statistics* **35**, 2173–2192.