Fast methods for fitting log-Gaussian Cox Process models in Ecology

Elliot Dovers

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Mathematics and Statistics

Faculty of Science

May 2021

Thesis Title

Fast methods for fitting log-Gaussian Cox Process models in Ecology

Thesis Abstract

Log-Gaussian Cox processes (LGCPs) offer a framework for regression-style modelling of point patterns that can accommodate latent effects. These latent effects can be used to account for missing predictors or other sources of clustering that could not be explained by a Poisson process. Such models are important in ecology where point patterns arise in the form of presence-only data – records of species' locations – and used to construct Species Distribution Models (SDMs) as a function of environmental variables. Fitting LGCP models can be difficult and time consuming and, as a result, limits the ability of researchers to flexibly analyse presence-only data. In this thesis, we develop novel methodology and software for fitting LGCP models, as well as demonstrating how to incorporate presence-only and other data sources jointly into SDMs.

Fitting LGCPs quickly is challenging due to their intractable marginal likelihood which involves a high dimensional integral to account for the latent Gaussian field – leading to large spatial variance-covariance matrices. In this thesis we address these using a novel combination of variational approximation and reduced rank interpolation. Additionally, we implement automatic differentiation that enables us to obtain exact gradient information rapidly for computationally efficient optimisation and inference. We demonstrate the method's performance through both simulations and a real data application, with promising results in terms of computational speed and accuracy compared to that of existing approaches.

We then extend our novel method to combine presence-only data with that obtained through scientific surveys to improve SDM in what is called data integration. We demonstrate scenarios in which sharing both the latent influence and the species' response to environment across each data set can improve upon results achieved by modelling each individually – both via simulation and using real data involving several species of flora in NSW, Australia.

Within this thesis we also illustrate the use of software developed to implement these advances via the freely available R package – scampr. The package allows users to fit likelihood-based LGCP to presence-only data swiftly and with a formula interface familiar to those with experience in other regression-style modelling frameworks implemented in R.

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.

COPYRIGHT STATEMENT

AUTHENTICITY STATEMENT

☑ I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

Thesis submission for the degree of Doctor of Philosophy

Thesis Title and Abstract | Declarations | Inclusion of Publications Statement | Corrected Thesis and Responses

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in the candidate's thesis in lieu of a Chapter provided:

- The candidate contributed **greater than 50%** of the content in the publication and are the "primary author", i.e. they were responsible primarily for the planning, execution and preparation of the work for publication.
- The candidate has obtained approval to include the publication in their thesis in lieu of a Chapter from their Supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis.

☑ The candidate has declared that **their thesis contains no publications, either published or submitted for publication**.

Candidate's Declaration

I declare that I have complied with the Thesis Examination Procedure.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, David Warton for sharing his expertise, advice, and support throughout the duration of my PhD. I have learnt a lot from him — in fields, and aspects of life, that go beyond statistics and ecology. I would also like to extend this gratitude to my co-supervisor Gordana Popovic for her knowledge, guidance, and encouragement — leaning on her recent experience as a PhD candidate helped me to remain focused on, and believe in, a path to completion. David and Gordana have both put a great deal of time and effort into the development of my thesis, and I am extremely thankful for the support and understanding I have received throughout the entire process. I can't imagine getting through this journey without them.

Special thanks go to my partner Tivoli for being kind and understanding, even when I haven't been at my best or even able to focus on anything other than this thesis. I can't praise my family enough for their love and support — something I've been lucky enough to always count on. Thank you, Mum, Dad, Remy and Rob, for your encouragement and comfort. To my friends and extended family who helped me get my mind off my work every now and then, albeit via video calls most recently — Peter, Claire, Dan, Jen, Tyler, Mic, Ro, and Tango, I appreciate your friendship over the years immensely.

Thank you to the Eco-Stats Group — Robert, Maeve, Michelle, and other members past and occasional. Our insightful weekly meetings and the invaluable feedback to drafts and presentations have always greatly enhanced my experience.

I would like to recognise and thank the University of New South Wales, particularly the School of Mathematics and Statistics and the support staff who have helped me

with tech support and admin during my studies.

This research is supported by an Australian Government Research Training Program (RTP) Scholarship. I greatly appreciate receiving this and the top-up scholarship from the school. These allowed me to fully dedicate myself to this research.

Finally, thank you to a well-timed La Niña weather cycle, for making sure there were less beautiful sunny days for surfing and diving that may have otherwise distracted me during an important period of my research and write-up.

Contents

1 Introduction **1**

1.1 Example Datasets 5

1.1.1 Gorilla Nesting Data 6

1.1.2 Flora Presence Locations 8

2 Fast, Likelihood-based approximation for log-Gaussian Cox Processes **13**

2.1 Introduction 14

2.2 Existing Methods 15

2.3 Proposed Methodology 20

2.3.1 Approximate Marginalisation 21

2.3.1.1 Variational approximation 21

2.3.1.2 Laplace approximation 22

2.3.2 Rank Reduction 22

2.3.3 Automatic Differentiation 28

2.4 Simulation Study 29

2.4.1 Simulation Results 31

2.5 Application: Gorilla Nesting Locations 36

2.5.1 Methods 36

2.5.2 Results 38

2.6 Discussion 44

3 Extending Data Integration Methods in Ecology **49**

3.1 Introduction 50

3.2 Existing Method . . . 53
3.3 Proposed Extension . . . 55
3.4 Simulation Study . . . 57
3.4.1 Simulation Results . . . 59
3.5 Application: Flora in the Greater Blue Mountains . . . 62
3.5.1 Methods . . . 62
3.5.2 Results . . . 64
3.6 Discussion . . . 65

4 `scampr` R Package: Spatially Correlated, Approximate Modelling of Presences in R 75
4.1 Introduction . . . 76
4.2 Fitting the LGCP Model . . . 77
4.3 Integrated Data Models . . . 82
4.4 Model Diagnostics and Inference . . . 87
4.5 Fine Tuning `scampr()` . . . 90
4.5.1 Basis Functions . . . 91
4.5.2 Speed Control . . . 96
4.6 Discussion . . . 98

5 Final Remarks 101
5.1 Summary . . . 102
5.2 Future Research . . . 103

A Additional results for Chapter 2 109
A.1 Variational approximation for a LGCP Model . . . 109
A.1.1 Constrained VA likelihood . . . 114
A.1.2 Profiled VA log-likelihood . . . 115
A.1.3 VA as a penalised likelihood . . . 116
A.2 Complete Simulation Results . . . 117

B Additional results for Chapter 3 121
B.1 Complete Flora Data Integration Results . . . 121

C Additional Details for Chapter 4 **129**

C.1 Data Requirements 129

C.1.1 Interpolation of Predictors 129

C.1.2 Quadrature (or Background) Points 130

List of Figures

1.1 Gorilla Nesting Data 6
1.2 Inhomogeneous K Functions: Gorilla Nesting Locations 8
1.3 Greater Blue Mountains World Heritage Area 9
1.4 Flora Datasets 10
1.5 Flora Data: Predictor Variables 11
1.6 Inhomogeneous K Functions: Flora Locations 12

2.1 1D Example of Local Bi-square Basis Functions 25
2.2 2×2 Simulation Design 30
2.3 Simulation Results: Computation Times 32
2.4 Simulation Results: Point Estimation 33
2.5 Simulation Results: Interval Estimation 35
2.6 Spatial Block Cross-Validation and INLA Mesh 37
2.7 Basis Function Configuration Search 39
2.8 Real Data Results: Fitted Intensities 41
2.9 Real Data Results: Fixed Effect Estimates 42

3.1 Simulation Design Example 69
3.2 Simulation Results: Predictive Accuracy 70
3.3 Simulation Results: Coefficient Recovery 71
3.4 Real Data Spatial Cross-Validation 72
3.5 Distances Between Data Locations 73
3.6 Real Data Results: Predictive Accuracy 74

4.1 `scampr` Plots 80

4.2 Inhomogeneous K Functions Envelopes in `scampr` 89
4.3 Basis Functions Available to `scampr` 92
4.4 Basis Function Search with `scampr` 94

List of Tables

1.1 Flora Data: Relative Presences 9

2.1 Simulation Results: Basis Function Configurations 34
2.2 Prediction Results: Gorilla Nesting Data 39
2.3 Computation Times for Method Components 43

3.1 Model Accuracy: Integrated vs. Presence/absence 60
3.2 Model Accuracy: Integrated vs. Presence-only 61

4.1 Defaults for `scampr()` 90

A.1 Simulation Results Summaries: Scenario S,S 118
A.2 Simulation Results Summaries: Scenario S,W 118
A.3 Simulation Results Summaries: Scenario W,S 119
A.4 Simulation Results Summaries: Scenario W,W 119

B.1 Results: Integrated Data Model — *C. eximia* 121
B.2 Results: Presence-only Data Model — *C. eximia* 122
B.3 Results: Presence/absence Data Model — *C. eximia* 123
B.4 Results: Integrated Data Model — *E. canaliculata* 123
B.5 Results: Presence-only Data Model — *E. canaliculata* 124
B.6 Results: Presence/absence Data Model — *E. canaliculata* 124
B.7 Results: Integrated Data Model — *H. cernuus* 124
B.8 Results: Presence-only Data Model — *H. cernuus* 125
B.9 Results: Presence/absence Data Model — *H. cernuus* 125
B.10 Results: Integrated Data Model — *E. sparsifolia* 126

B.11 Results: Presence-only Data Model — *E. sparsifolia* 127
B.12 Results: Presence/absence Data Model — *E. sparsifolia* 128

Chapter 1

Introduction

Point patterns are data representing the location of events within a domain of interest. In this thesis, we assume a spatial point pattern. Modelling such data with spatial predictors is useful in a variety of fields for insight into the relationship between occurrence of point events and the observed environment or other characteristics of the domain. This is also useful in predicting where and how abundant these events might be. In application, point events assume a variety of roles as diverse as, for example, earthquakes (Ogata, 1988), financial market transactions (Bowsher, 2007), crime (Mohler et al., 2011), conflicts in war (Zammit-Mangion et al., 2012) and species locations (Renner et al., 2015).

We will focus on spatial point patterns in ecology, where the point events are recorded locations of a species (or evidence of their behaviour) — called presence-only data. These are used to produce species distribution models (SDMs), widely used in the ecological literature to relate the spatial distribution of species to their environment. In turn, these models are used to: guide conservation efforts through environmental planning; identify risk factors to vulnerable populations and; undertake stock assessments for commercial collection/hunting efforts (Elith and Leathwick, 2009). Presence-only data can be found historically in museum records and, for plant species, within herbaria around the globe (Pearce and Boyce, 2006). More recently, online stores such as the Global Biodiversity Information Facility, iNaturalist, e-Bird, Pl@ntNet, and others, have led to an increase in the volume of presence-only data available to researchers — in many instances collected by people without training in ecology, in what is termed "citizen science" (Chandler et al., 2017). A common trait shared by both recent and historic sources of these data is that their collection is often *ad hoc* or opportunistic, which introduces bias that must be accounted for when modelling.

Point processes provide a framework that is a natural approach to modelling presence-only data as it avoids the problem of a lack of true absences, and hence vagueness in selecting pseudo-absences to perform, say, binary regression (Warton and Shepherd, 2010; Chakraborty et al., 2011). Likewise avoided is the information loss that occurs when binning data into grids to be treated as independent Poisson random variables. Point process models have been used in ecology to model presence-only

data for some time (see Renner et al., 2015, for a comprehensive review). Most commonly, these models take the form of an inhomogeneous Poisson process (IPP) which considers the point pattern to be the Poisson realisation of a spatially varying intensity/rate, characterised by some log-linear combination of predictor variables. Renner and Warton (2013) and Hastie and Fithian (2013) showed that this is equivalent to MAXENT, one of the more popular procedures used in the ecological literature for modelling presence-only data. For example, the original paper describing the MAXENT procedure has been cited over $14,000$ times (Phillips et al., 2006), and is currently accumulating about 600 additional citations each year.

Implicit in the IPP framework is the assumption that each unique point event (or collections of point events within non-overlapping regions) is independent, arising from their shared intensity function which is, in turn, conditional on the model predictors. Reframed, this assumes that the predictors in the model account for all of the clustering or repulsion found in the resulting point pattern. In many practical applications however, presence-only data will exhibit additional clustering (or repulsion) due to some unobserved or unmeasured covariates. Cox, Neyman-Scott, Hawkes, and Gibbs processes are all examples of point process frameworks that can involve additional clustering or spatial correlation; or can induce it through point interactions (Daley and Vere-Jones, 2007). Of these, Cox processes are better equipped to model point patterns arising from environmental phenomena compared to those driven by point interactions (Diggle et al., 2013). Cox processes have a stochastic, spatially varying intensity which can be used to provide the model a hierarchy to include additional correlation structures. This is particularly true of the log-Gaussian Cox process (LGCP) of Møller et al. (1998) where the log-intensity of the process is a Gaussian random field (GRF), the correlation structure of which provides the additional spatial correlation. The GRF can be modelled such that it plays the role of missing predictors.

Despite the appealing features of LGCP in modelling presence-only data in ecology, model complexity and long computation times are a barrier to their widespread use. Much existing software adopts Markov Chain Monte Carlo (MCMC) sampling procedures to model the GRF precisely (see *e.g.* Taylor et al., 2013; Diggle et al.,

2013) which can be computationally costly, and scales poorly with sample size. Rue et al. (2009) introduced a framework for approximately fitting latent Gaussian models — a class into which LGCP models can be non-trivially coerced following the likes of Illian et al. (2012) or Bachl et al. (2019). In either case, modelling tends to be approached from a Bayesian standpoint, for which full posterior distributions on model parameters are required — this too can contribute to lengthy computations.

The aim of this thesis is to develop fast, maximum likelihood-based, approximation methods for fitting LGCPs to point patterns, and illustrate how this framework can be used to model ecological presence-only data in a variety of ways. Additionally, the thesis introduces software written in `R` (R Core Team, 2020) that implements these advances, with the goal of making LGCP models faster and simpler to implement — providing researchers better access to tools that can fit these spatial models.

In Chapter 2 we propose novel methodology for fitting LGCP to point pattern data that uses a combination of variational approximation (VA), reduced rank "kriging" (a term used in spatial statistics for interpolation, see Cressie, 1993) and automatic differentiation (AD). We approach this in a frequentist, maximum likelihood setting — something not common in the LGCP literature. We examine simulations to test how fast and accurately our proposed model performs against a widely used, approximate, Bayesian approach to fitting LGCP models: Integrated Nested Laplace Approximations (INLA; Rue et al., 2009). We further highlight contrasts between the methods when applied to an ecological dataset.

Chapter 3 focuses on data integration in ecology — the process of combining presence-only data with presence/absence data from rigorous scientific surveys — to improve SDMs. We adopt elements of the methodology proposed in Chapter 2 to extend an existing data integration framework (as in Fithian et al., 2015) to account for spatial dependence between the two datasets that may arise due to missing or unaccounted-for covariates. We use simulations to assess the performance of our method, comparing it to modelling the data separately as well as the original integration framework (Fithian et al., 2015). We analyse several examples of real-world data to further illustrate our proposed method.

In Chapter 4 we introduce an `R` software package, `scampr`, that implements the advances of the previous two chapters, providing a user-friendly interface. We demonstrate the code and functionality through examples. The thesis is concluded with final remarks in Chapter 5.

We intend to submit the body of work presented in this thesis to peer-reviewed journals in the near future. We envisage three manuscripts arising from the various chapters. First, Chapter 2 will form a methodological paper presenting our novel approach to fitting LGCP models. Second, we intend Chapter 3 to form a paper submitted to an ecological modelling journal, where there is an existing body of literature on data integration techniques. Finally, we hope to publish Chapter 4 as a software paper.

We conclude this chapter by introducing the motivating datasets and notation used throughout this thesis.

1.1 Example Datasets

We will use several example datasets as both motivation for, and illustration of, the methodologies proposed in this thesis. These comprise presence-only data — presence locations of the events of interest — along with various predictor variables representing characteristics of the domain in which they occur. Throughout the thesis we will denote the domain by $\mathcal{D}$ which is continually indexed by spatial parameter $\boldsymbol{s}$. We can then write the n presence records as $S_n = \{\boldsymbol{s}_i\}$ for $i = 1, \ldots, n$. In our examples $\boldsymbol{s}$ is vector of coordinates and $\mathcal{D} \subseteq \mathbb{R}^2$, however the methods presented apply more generally. Other variables or processes are assumed to be likewise continually indexed over the domain. We use $\boldsymbol{X}(\boldsymbol{s})$ to represent p predictor variables at location $\boldsymbol{s}$, forming a row vector. For brevity we write $\boldsymbol{X}$ to mean an $n \times p$ matrix representing the predictors at each of the n locations (these should be clear given the context). We refer to an entire predictor "field" using $X(\mathcal{D})$, however in practice the data will often be a geo-referenced grid of values. In the subsequent sections of this chapter we perform exploratory analysis on datasets to highlight the need for a LGCP framework to model them.

1.1.1 Gorilla Nesting Data

The first dataset we use contains the locations of gorilla nesting sites in Kagwene Gorilla Sanctuary in Cameroon as provided within the `R` package `inlabru` (Bachl et al., 2019) and `spatstat` (Turner and Baddeley, 2005). The data consists of 640 (non-duplicate) nesting sites — defined as a location containing one to six nests (Funwi-Gabga and Mateu, 2012). These are located in an irregularly shaped, two dimensional domain reflecting the Kagwene Gorilla Sanctuary $|\mathcal{D}| \approx 20\text{km}^2$ in size. The region and corresponding nesting locations are shown in Figure 1.1a. Also included are predictors comprising two covariates and one factor. These are elevation above sea level (m); distance to fresh water source (m) and; average temperature category — Coolest, Moderate and Warmest. See Figure 1.1b-d respectively. These come discretised into a fine grid of 25,380 squares each approximately 800m^2.

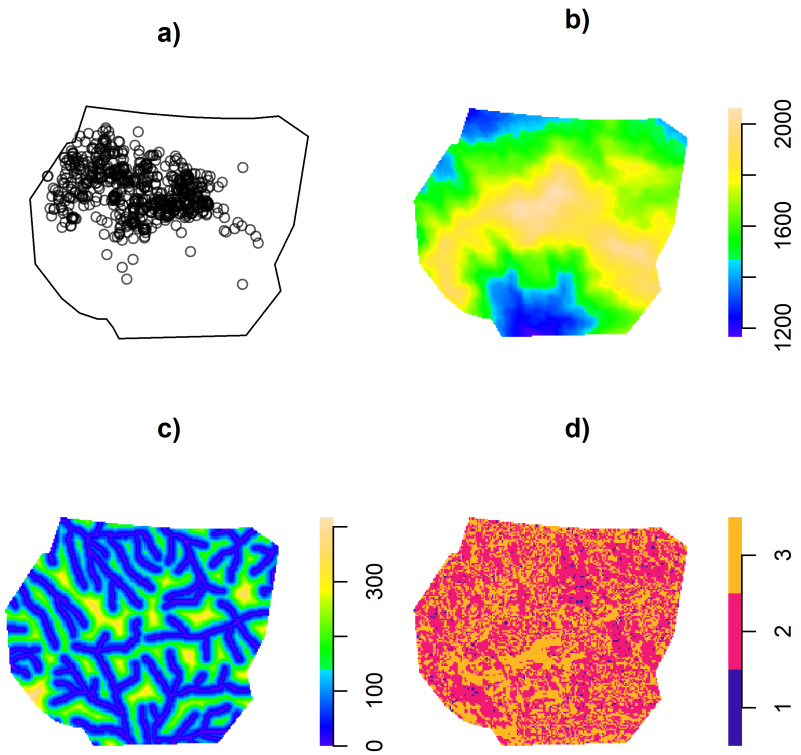

Figure 1.1: a) Nesting locations for gorillas in the Cameroon sanctuary. b) Spatial covariate describing elevation in meters above sea level. c) Spatial covariate describing distance in meters to nearest water source. d) Factor describing heat category: 1 = Coolest, 2 = Moderate, 3 = Warmest.

While LGCPs offer a flexible framework for modelling a point pattern like the gorilla

nesting data presented here, it would be far simpler to fit an inhomogeneous Poisson process (IPP) and be done with it. Hence we need to determine whether an IPP is adequate for the modelling task at hand. One way to check this is using the inhomogeneous K function (K_{inhom}; Baddeley et al., 2000) — a generalisation of Ripley's K function for stationary point processes, also known as the reduced second order moment function. This can be loosely interpreted as counting the number of points within certain distances from one another in the point pattern (weighted by the localised intensity and perhaps with edge correction). Hence K_{inhom} can be used to examine if clustering is present within a point pattern beyond that accounted for given a spatially varying intensity surface. That is, if the intensity surface is parameterised by predictor variables — as is the case when we use the fitted intensity from an IPP — we can identify when additional spatial clustering exists, perhaps due to missing or unaccounted-for covariates. This can be done by comparing the observed K_{inhom} for our point pattern against those calculated on point patterns simulated from the fitted IPP intensity. Here we compare the observed K_{inhom} to a simulation envelope constructed from 1000 point patterns simulated in this way.

We want to construct a simulation envelope in a way that controls for the functional nature of K_{inhom}, such that the envelope provides global control of Type I error, rather than pointwise control. This is non-trivial but can be achieved using the methods described in Myllymäki et al. (2017), and available on `CRAN` in the `GET` package (Myllymäki and Mrkvička, 2019).

The results for the gorilla nesting data example can be found in Figure 1.2, for both an IPP and LGCP model (top and bottom panel respectively) that regresses the locations against all the available predictors. As we can see, there is evidence ($\alpha = 0.05$) of violation of the Poisson assumption (top panel), with additional spatial clustering at all inter-point distances less than a kilometre. The LGCP model seems to adequately account for this additional clustering. The confidence bounds in the IPP case (top panel) seem strikingly narrow, which arises because the fitted intensity for this IPP is particularly flat (as seen later in Section 2.5).

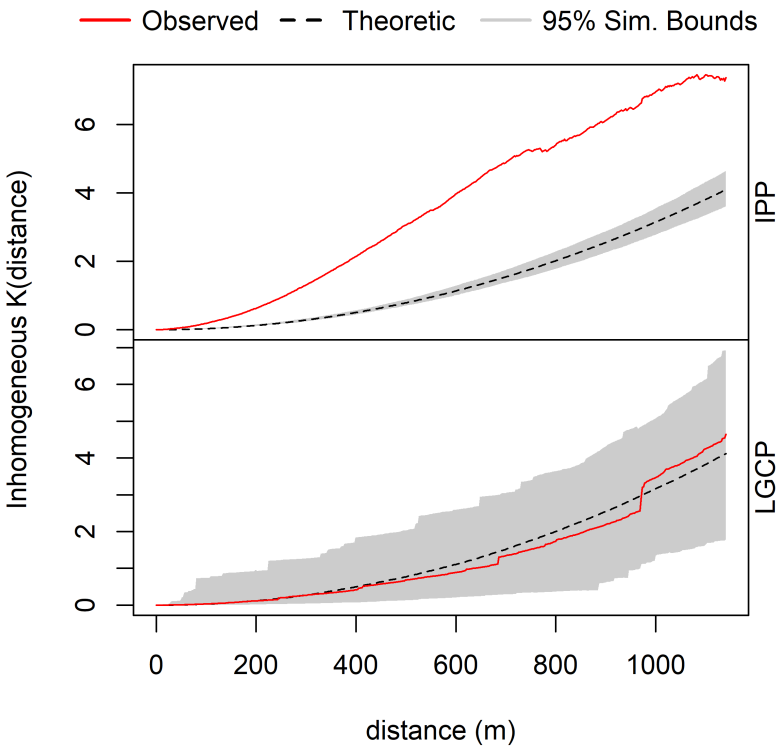

Figure 1.2: The inhomogeneous K functions for the gorilla nesting data under an IPP model (top panel) and LGCP model (bottom panel) — both use the three fixed effect predictors. All K functions use a border correction as in Baddeley and Turner (2000). Shaded regions are global 95% confidence bounds based on 1000 simulated point patterns from the fitted intensity, and the central dashed line shows the equation $K(d) = \pi d^2$ that represents the theoretical K function for the fitted IPP. There is evidence of additional spatial clustering to that accounted for by the IPP model, not so in the LGCP where the latent field seems to adequately account for the clustering. Note that the observed function far exceeds the 95% global simulation bounds in the IPP case. Models fit here are found in Chapter 2.

1.1.2 Flora Presence Locations

The other datasets used in this thesis are presence-only data for four species of flora within the Greater Blue Mountains World Heritage Area (GBMWHA) in NSW, Australia (Figure 1.3). These data are useful to our current aims as we additionally have presence/absence data for these species — meaning we can study methods that integrate these two data types within a single model. Presence/absence data comprise of a vector of zero or one responses, $\boldsymbol{y}$ of length n_{survey}, logically representing whether or not the species was present at each site, $\left\{\boldsymbol{s}_i^{\text{PA}}\right\}_{i=1}^{n_{\text{survey}}} \in \mathcal{D}$, from detailed scientific surveys. In the flora survey described here we have $n_{\text{survey}} = 8,223$.

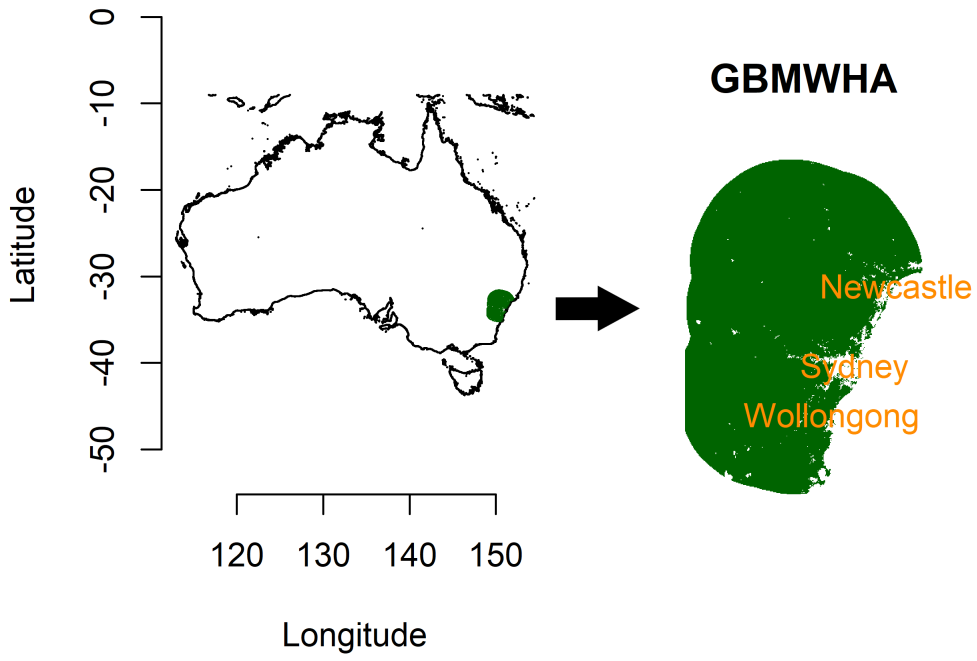

Figure 1.3: The domain of interest for the flora datasets, the Greater Blue Mountains World Heritage Area (GBMWHA) in N.S.W, Australia.

The species are *Corymbia eximia*, *Eucalyptus sparsifolia*, *Eucalyptus canaliculata* and *Homoranthus cernuus*. The first three species are large trees while *H. cernuus* is a small shrub, and they were chosen as species highly endemic to this study region that differ in their spatial distribution and range. *E. canaliculata* and *H. cernuus* are species with a very restricted range, only found in a small subregion of the GBMWHA. The locations of both the presence-only and presence/absence data for each species is found in Figure 1.4. Table 1.1 shows the relative numbers of presences found in each dataset. As may be expected, the two species with restricted distributions are the least frequently observed. In these data, we find all species except *H. cernuus* to be more prevalent within the survey data than in the presence-only records.

Table 1.1: Number of presences of each species in the presence/absence data (also referred to as survey data; PA, top row) and presence-only data (PO, bottom row).

	C. eximia	*E. canaliculata*	*H. cernuus*	*E. sparsifolia*
PA	324	78	9	618
PO	242	38	11	194

We have several predictor variables measured throughout the GBMWHA as a geo-

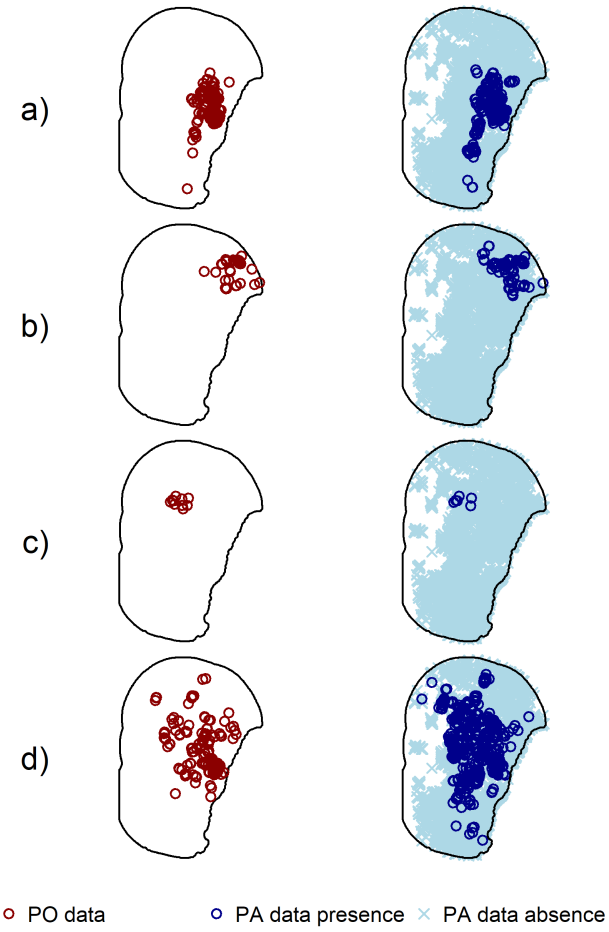

Figure 1.4: Locations of both presence records (presence-only, left column) and survey sites (presence/absence, right column) in the GBMWHA for each of the four species. a) *Corymbia eximia*. b) *Eucalyptus canaliculata*. c) *Homoranthus cernuus*. d) *Eucalyptus sparsifolia*.

referenced regular grid, at a spatial resolution of 1km^2 which we can use to interpolate values to the presence/absence and presence-only datasets. Predictors include two environmental variables — average annual minimum (MNT) and maximum (MXT) temperatures (^{o}C) — and two variables that may affect how the presence-only data were collected, we call these biasing predictors — distances (km) to main road (D.Main) and urban areas (D.Urb). These are displayed in Figure 1.5a-d respectively. As in the previous data example, we can fit the IPP model and compare the observed K_{inhom} to that of simulated point patterns from the fitted intensity. In Figure 1.6 we see there is evidence ($\alpha = 0.05$) of additional clustering in three of the four species presented here. The species that lacked evidence of spatial clustering was one of the range-restricted species with a small number of presence records.

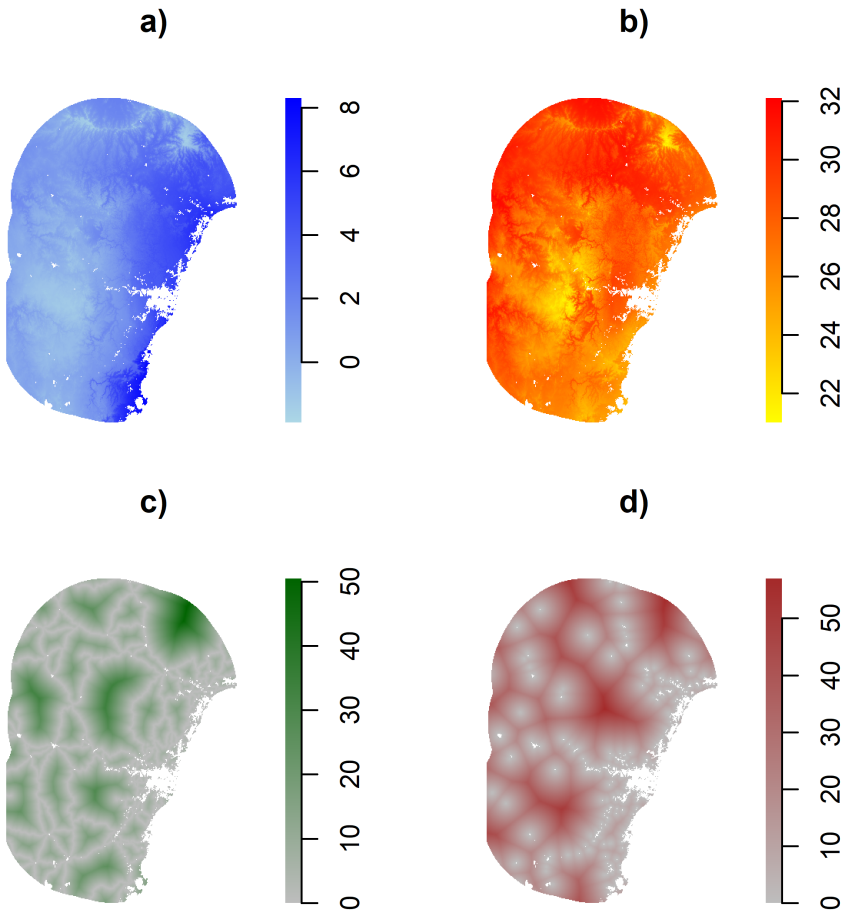

Figure 1.5: Predictor variables used in the analysis of the flora datasets. a) Minimum average annual temperature (^{o}C). b) Maximum average annual temperature (^{o}C). c) Distance from a main road (km). d) Distance from urban area (km).

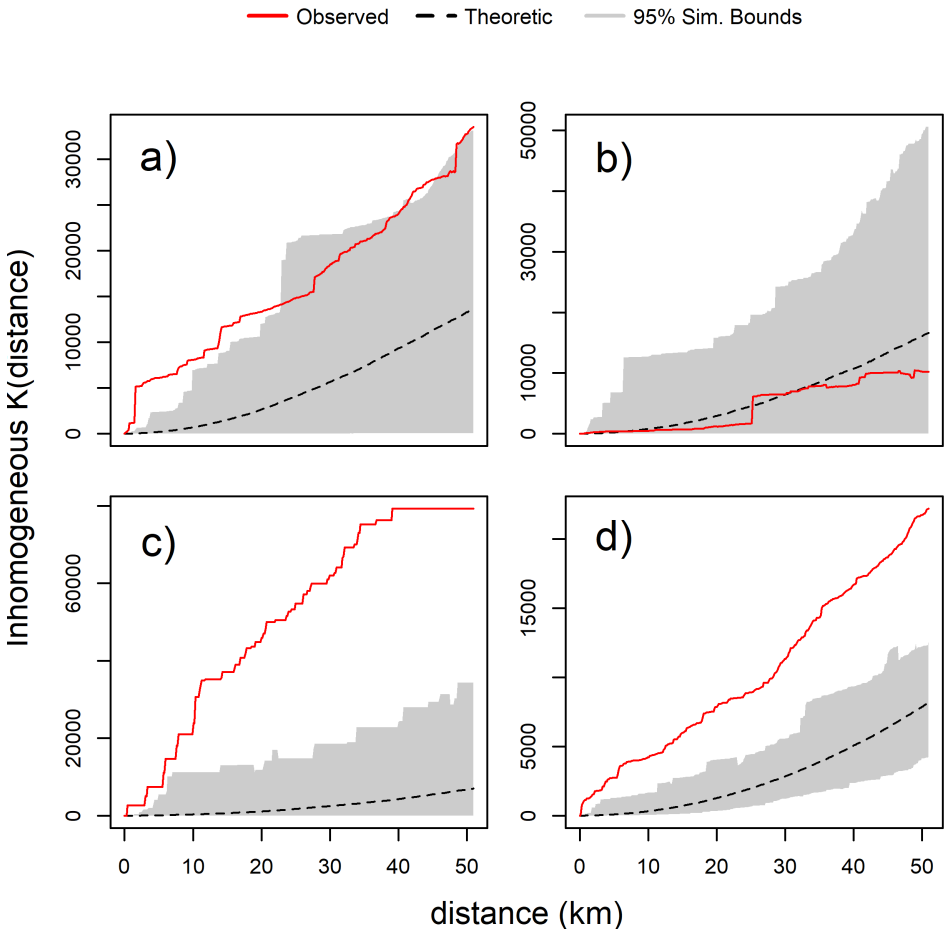

Figure 1.6: The inhomogeneous K functions for each species of flora locations within the presence-only datasets. a) *Corymbia eximia*. b) *Eucalyptus canaliculata*. c) *Homoranthus cernuus*. d) *Eucalyptus sparsifolia*. Each uses IPP models as found in Section 3.5 and K functions and confidence bounds constructed as in Figure 1.2 (top panel).

Chapter 2

Fast, Likelihood-based approximation for log-Gaussian Cox Processes

2.1 Introduction

Point patterns in ecology often exhibit additional spatial clustering to that accounted for by environmental predictors that are available to a researcher. This is exemplified in the previous chapter for the gorilla nesting data in Figure 1.2 (top panel) and three of the four species of flora in Figure 1.6. These plots of inhomogeneous K functions provide evidence that the underlying inhomogeneous Poisson process (IPP) framework is inadequate at modelling the drivers of these point patterns. The log-Gaussian Cox Process (LGCP; Møller et al., 1998) offers a way to incorporate such additional spatial clustering into point process models. This is achieved by including a Gaussian random field (GRF) to induce additional spatial correlation between observations — effectively acting as a spatially correlated error term in the model. LGCP models are particularly appropriate in instances where clustering arises from missing or unmeasured environmental processes/phenomena, as opposed to those in which clustering/dispersal is due to interactions between the point events themselves. Distinguishing between these processes empirically can be difficult or impossible (Diggle et al., 2013). So while we can test for the presence of clustering with the inhomogeneous K function as described in Chapter 1, deciding on the appropriate modelling mechanism will more likely come from *a priori* research or hypotheses. In either case, more accurate modelling of the underlying drivers of point patterns enables researchers to make more accurate inference and predictions about them.

Fitting LGCP models often takes a long time. In previous literature, model fits were typically performed in a Bayesian context where full posterior distributions on parameters are estimated. This accounts for some of the long computation. In particular, MCMC sampling scales poorly as the size of the point pattern increases. Even one of the fastest approximation methods, Integrated Nested Laplace Approximations (INLA; Rue et al., 2009), can take a prohibitive time to fit the model — particularly when making predictions at many locations, one of the main motivations for using a model with spatially correlated errors. A maximum likelihood approach to LGCP has the potential to speed up analysis, with a focus on point es-

timation and access to the likelihood-based statistical toolkit; including the various information criteria and likelihood ratio testing.

In this chapter we propose a fast, novel maximum likelihood approach to fitting LGCP to point pattern data, involving three innovations. First, variational approximation (VA) permits a closed form approximation to the marginalised log-likelihood. Second, we use a rank reduced approximation to the large spatial variance-covariance matrices that arise and are otherwise very computationally demanding. Finally, automatic differentiation is used to quickly obtain gradient information for efficient optimisation and inference. Performance in fitting LGCP is also examined, trialled against a leading alternative (INLA) in a simulation study. For motivation and illustration we analyse locations of gorilla nesting sites in a sanctuary in Cameroon (Section 1.1.1)

2.2 Existing Methods

The target of inference for point process models is their spatially varying intensity function $\lambda(\boldsymbol{s})$, which describes the limiting number of point events per unit area in some infinitesimally small region around the point $\boldsymbol{s}$. We assume the log-intensity is a linear combination of predictor variables, $\boldsymbol{X}(\boldsymbol{s})$ and corresponding effects, $\boldsymbol{\beta}$. For conciseness, we include any intercept term in $\boldsymbol{\beta}$ with its corresponding indicator variable in $\boldsymbol{X}(\boldsymbol{s})$. Cox processes are an example of this broader class, characterised by the stochastic nature of their intensity function. As a specific example of these, Møller et al. (1998) introduced the log-Gaussian Cox Process (LGCP), for which the (log-)intensity function inherits stochasticity from a Gaussian process, $\xi(\boldsymbol{s})$, often called a Gaussian random field (GRF). Note this model assumes that there are two stochastic outcomes to the observed point pattern $S_{N=n}$, *i.e.* the realisation of the Gaussian field and the realisation of the Poisson process. The zero mean GRF may be characterised by some covariance function, f_ξ, in turn governed by parameters $\boldsymbol{\theta}_\xi$. The LGCP intensity is given by

$$\ln \lambda(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \xi(\boldsymbol{s}). \tag{2.1}$$

This highlights a key aspect of the LGCP, that conditional on ξ, the process is an IPP with fixed mean, $\int_{\mathcal{D}} \lambda(t)\,\mathrm{d}t$.

It can be shown that the point pattern has (conditional) probability density function

$$\begin{aligned}\pi(S_n|\boldsymbol{\xi}) &= \left\{\prod_{i=1}^{n} \exp\{\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta} + \xi(\boldsymbol{s}_i)\}\right\} \exp\left\{|\mathcal{D}| - \int_{\mathcal{D}} \lambda(t)\,\mathrm{d}t\right\} \\ &\propto \left\{\prod_{i=1}^{n} \exp\{\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta} + \xi(\boldsymbol{s}_i)\}\right\} \exp\left\{-\int_{\mathcal{D}} \lambda(t)\,\mathrm{d}t\right\} \qquad (2.2)\end{aligned}$$

with respect to a unit rate Poisson process, see Daley and Vere-Jones (2007) for details. Note that we wrote the probability density as $\pi(S_n|\boldsymbol{\xi})$ – we will use $\pi(\cdot)$ throughout this thesis to denote the probability density function of a random variable (or stochastic process) in general. Different variables will have different probability density functions and these are distinguished by the arguments of $\pi(\cdot)$.

The spatial integral that describes the mean, $\int_{\mathcal{D}} \lambda(t)\,\mathrm{d}t$, can be approximated using numerical quadrature, *i.e.* discretising $\mathcal{D}$ into q regions represented by points $S_q = \{\boldsymbol{s}_i\}_{i=n+1}^{n+q}$, so that each $j = 1, \ldots, q$ quadrat has area $w_j \equiv |\boldsymbol{s}_j|$ such that $|\mathcal{D}| = \sum_{j=1}^{q} w_j$. This gives us the quadrature approximation

$$\int_D \lambda(t)\,\mathrm{d}t \approx \sum_{j=1}^{q} w_j \exp\{\boldsymbol{X}(\boldsymbol{s}_j)\boldsymbol{\beta} + \xi(\boldsymbol{s}_j)\}. \qquad (2.3)$$

As the discretisation becomes fine enough, the approximation converges to the true value of the integral (Davis and Rabinowitz, 2007). This numerical quadrature approach is commonly used to fit IPP to point patterns (as in Berman and Turner, 1992).

The quadrature approximation means that the GRF has an $n+q$ dimensional realisation, *i.e.* $\boldsymbol{\xi} = \{\xi(\boldsymbol{s}_i)\}_{i=1}^{n+q}$. As such, model fitting procedures will involve the corresponding $(n+q)$-variate normal distribution, $\pi(\boldsymbol{\xi}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$, where $\boldsymbol{C}$ is the variance-covariance matrix induced by the covariance function, $f_{\boldsymbol{\xi}}$. For likelihood-based fitting in a frequentist paradigm this occurs within the marginal log-likelihood

of the point pattern, given by

$$\ell(\boldsymbol{\beta}) = \log \int \pi(S_n|\boldsymbol{\xi})\,\pi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}. \tag{2.4}$$

Computations and storage involving the resulting variance-covariance matrix from $\pi(\boldsymbol{\xi})$ is a major source of computational burden — a common problem in spatial statistics. This is particularly true for MCMC methods, which require the high dimension GRF at each sampling iteration. Taylor et al. (2013) provide an `R` package that fits a LGCP via MCMC methods. There are many other examples of sampling routines, including Metropolis-Hastings algorithms via Gibbs samplers (Taylor et al., 2015) and the Metropolis-adjusted Langevin algorithm (Møller et al., 1998; Brix and Diggle, 2001; Diggle et al., 2013). These strategies take a long time to fit; scale poorly as the size of point patterns increases, and require thorough tuning and careful checking of mixing and convergence properties (Diggle et al., 2013). The efficiency of many of these MCMC routines rely upon the power of computation at the time. We note there may be improvement found in advances in machine usage. For example, the advances of the GRETA package of Golding (2019) could provide faster ways of implementing the methods described above via Google's TensorFlow, however this is not explored here.

INLA (Rue et al., 2009) is an approximate Bayesian inference scheme that uses a Gaussian Markov random field in place of any latent Gaussian process component of a regression-type model. This results in a sparse precision matrix that assists in avoiding the problematic computations involving the variance-covariance matrix. Additionally, Laplace approximations are used for posterior densities of hyperparamters of the model. Taylor and Diggle (2014) performed a simulation study into the comparable speeds and efficacy of INLA and MCMC routines to fitting LGCP and found INLA to be generally less accurate but much faster. Illian et al. (2012) provide a detailed overview of fitting complex spatial point patterns with INLA using LGCPs, additionally demonstrating the ability to include pointwise interactions. This approach uses a regular lattice of cells containing point counts to approximate the LGCP with the resulting collection of Poisson variables. A fine

spatial scale of the lattice is required for the approximation to converge as shown by Waagepetersen (2004). However, the advances of Simpson et al. (2016) exploiting the equivalence of the Gaussian Markov random field and the solution to a stochastic partial differential equation highlighted by Lindgren et al. (2011) permits a reduced dimension approximation to integrate the field rather than relying on the fine scale lattice. Shirota and Gelfand (2017a) developed a similar approximate Bayesian method that incorporates a pseudo-marginal MCMC routine and does not have the limitation on the number of hyperparameters found within the INLA scheme. More recently, Bachl et al. (2019) have created a wrapper-package for INLA (`inlabru`) to improve the usability and lower the bar-of-entry to non-specialist users of LGCP models. We use INLA as a benchmark for the methodology proposed here. Others combine some rank reduction techniques within MCMC routines to try and improve computation speed, such as Chakraborty et al. (2011), using the predictive process of Banerjee et al. (2008).

The GMRF approach of Lindgren et al. (2011) can also be used outside the `INLA` software framework to incorporate GRFs into spatial models in a frequentist paradigm. The R package, `VAST` (Thorson, 2019) provides such a framework for a variety of ecological models (excluding LGCP models that are the focus of this thesis). The software uses `INLA` to construct the GMRF mesh structure that is then passed to more general statistical computing routines for model fitting. Applications of `VAST` (or its precursor software) include: Thorson et al. (2015) who analyse ground fish abundance using delta generalised linear mixed models; Thorson et al. (2016) for fitting joint species distribution models that have one or more latent GMRFs. In the latter example the authors report that models including one to six latent GRF can be fitted in a matter of hours.

In this thesis we are proposing a novel methodology that applies a combination of modern techniques and software tools to the problem of fitting LGCP models efficiently. In the remainder of this section we we broadly introduce and review some of these tools.

Variational Approximation (VA) is a method for approximating intractable integrals arising, for example, from marginalising random (latent) components of a joint

probability density function. Simply put, this is done by substituting the unknown probability density of the latent effects with a candidate class of (variational) density functions that permit a closed form solution to the intractable integral. This closed form solution is then optimised with respect to the candidate variational densities - motivated by the notion of minimising the Kullback-Leibler divergence — Appendix A.1 further elucidates this. Ormerod and Wand (2010) provide a detailed summary of VA, defining a range of candidate variational density classes. The most commonly used class are termed *product* density VA which effectively assumes independence between random components permitting the unknown density to be factorised by some product of simpler density functions. Another popular class is termed by Ormerod and Wand (2010) as *parametric* density VA as this involves assuming the unknown density belongs to a particular parametric family whose density function yields a closed form solution to the particular problem at hand.

VA has been around in the fields of physics for some time (Cooper et al., 1986) and more recently machine learning (Opper and Archambeau, 2009), but is still relatively under-utilised in the statistical literature. This is particularly true of the parametric version, specifically in a frequentist context (Ormerod and Wand, 2010). Hence there is, at this stage, a somewhat case-by-case understanding of the asymptotics of VA estimators. Bickel et al. (2013) show asymptotic normality of estimators arising from a product density transform-type variational approximation to stochastic blockmodels and more recently Wang and Blei (2019) show consistency for frequentist estimators again using the product density version of VA. Hall et al. (2011) develop asymptotic theory for Gaussian VA for Poisson Mixed models. Hui et al. (2019) show consistency and asymptotic normality of Gaussian VA estimates for fitting semi-parametric models involving several of the exponential family of distributions. This is particularly relevant to the method proposed here as there are many similarities between the approximate marginal likelihoods of the Poisson generalised additive model (GAM) and that of the approximate LGCP we propose here — *i.e.* we approximate the latent field in our LGCP using what could be considered as a 2D penalised smoother in the GAMs terminology (Hastie and Tibshirani, 1993).

Another key technique used in this thesis is basis function approaches to approxi-

mating Gaussian processes (or fields). Including latent Gaussian random fields is a common practice in spatial analysis, to represent unobserved processes that are important to models with a variety of target responses. It is also common to encounter a need to address the computational burden caused by the latent field having a prohibitively large dimension. There are a variety of potential strategies to address this, including: low-rank approximations; enforcing sparsity in the resulting covariance matrices; and exploiting parallel computing. Heaton et al. (2019) provides a good illustration and comparison of methods that are used to compute (directly observed) Gaussian processes, pitted against one another for predictive accuracy. Due to the latency of the process we are trying to characterise in modelling point pattern data, we focus on low-rank approximations that do not prioritise the precise estimation of the field's characteristics. Specifically, we look towards basis function approximations, as commonly used in spatial (and kernel) smoothing. Examples of these include fixed rank kriging (Cressie and Johannesson, 2008), predictive processes (Banerjee et al., 2008) and GAMs (Hastie and Tibshirani, 1993). It should be noted that other approaches like nearest-neighbour Gaussian processes (Datta et al., 2016) show promise in approximating non-latent spatial processes (Heaton et al., 2019). The previously mentioned GMRF approach of Lindgren et al. (2011) is another example that is widely used. A key element that improves computation for this method of modelling ξ is that it yields sparse precision matrices. However, we note that the range parameter that informs the sparsity must be estimated, at some computational cost.

2.3 Proposed Methodology

Our proposed strategy is to maximise an approximation to the likelihood (Equation 2.4), designed for fast computation, at minimal cost in terms of accuracy. This approach requires addressing several challenges that we outline below.

2.3.1 Approximate Marginalisation

The first major problem to be overcome in our proposed approach is that the integral in the marginal likelihood, Equation (2.4), is intractable. We examine two approaches to approximate it.

2.3.1.1 Variational approximation

The first is to use Gaussian variational approximation (VA), a parametric density transform (Ormerod and Wand, 2010) which we implement in a frequentist paradigm. Specifically, we replace the conditional or "posterior" density of the latent field, $\pi(\boldsymbol{\xi}|S_n)$, with some multivariate Gaussian density function, $\pi_{\text{VA}}(\boldsymbol{\xi})$. The integral in Equation (2.4) then has a closed form solution. We denote the mean of this posterior as $\boldsymbol{m}_{\text{VA}}$ and the variance-covariance matrix as $\boldsymbol{C}_{\text{VA}}$, *i.e.* $\pi_{\text{VA}}(\boldsymbol{\xi}) \sim \mathcal{N}(\boldsymbol{m}_{\text{VA}}, \boldsymbol{C}_{\text{VA}})$. While the Gaussian posterior assumption might not be exactly satisfied, it is a plausible approximation since the "prior" on the random field is Gaussian, $\pi(\boldsymbol{\xi}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C})$.

To optimise the approximation we minimise the Kullback-Leibler divergence between $\pi_{\text{VA}}(\boldsymbol{\xi})$ and $\pi(\boldsymbol{\xi}|S_n)$. This is achieved by simply estimating the variational parameters $\boldsymbol{m}_{\text{VA}}$ and $\boldsymbol{C}_{\text{VA}}$ that maximise the variational approximation to the marginal log-likelihood of the point pattern, given by

$$\underline{\ell}_{\text{VA}}(\boldsymbol{\beta}) = \int \pi_{\text{VA}}(\boldsymbol{\xi}) \log \left[\frac{\pi(S_n|\boldsymbol{\xi})\,\pi(\boldsymbol{\xi})}{\pi_{\text{VA}}(\boldsymbol{\xi})} \right] \mathrm{d}\boldsymbol{\xi} \tag{2.5}$$

so that the parameter estimates are

$$\left\{ \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{m}}_{\text{VA}}, \hat{\boldsymbol{C}}_{\text{VA}}, \hat{\boldsymbol{C}} \right\} = \underset{\{\boldsymbol{\beta}, \boldsymbol{m}_{\text{VA}}, \boldsymbol{C}_{\text{VA}}, \boldsymbol{C}\}}{\arg\max}\, \underline{\ell}_{\text{VA}}$$

That is, we find estimates that simultaneously maximise $\underline{\ell}_{\text{VA}}$ for the fixed effects, $\boldsymbol{\beta}$, as well as the variational parameters, $\boldsymbol{m}_{\text{VA}}$ and $\boldsymbol{C}_{\text{VA}}$, and the prior variance-covariance matrix of $\boldsymbol{\xi}$, $\boldsymbol{C}$. The integral in Equation (2.5) is simply an expectation with respect to the variational density. We show later, for log-Gaussian Cox process likelihoods, that if $\pi_{\text{VA}}(\boldsymbol{\xi})$ is Gaussian, the integral will have a closed form,

considerably simplifying parameter estimation.

2.3.1.2 Laplace approximation

The second approach we use to approximate the intractable marginalisation, as an alternative to VA, is using Laplace approximation. This uses the Laplace formula for which software and literature abounds — see for example Wolfinger (1993) and Kristensen et al. (2016). Laplace approximations of high dimensional integrals can be poor (Shun and McCullagh, 1995) and so within our model the dimension of $\boldsymbol{\xi}$ (*i.e.* $n+q$) can be problematic since q is often large, as can be the size of the point pattern, n. However, in Section 2.3.2 we propose a rank reduction approach where we effectively restrict this dimension to $k \ll n+q$ which will, provided k is kept small, largely mitigate this issue. If we re-express the marginal likelihood in Equation (2.4) as

$$\ell(\boldsymbol{\beta}) = \log\left[\text{const.} \cdot \int \exp\{f(\boldsymbol{\xi})\}\, \mathrm{d}\boldsymbol{\xi}\right]$$

then its Laplace approximation is given by

$$\underset{\sim}{\ell}_{\text{Laplace}}(\boldsymbol{\beta}) \approx \log\left[\text{const.} \cdot \frac{(2\pi)^{\frac{k}{2}}}{|-H_f(\boldsymbol{\xi}_0)|^{\frac{1}{2}}} \exp\{f(\boldsymbol{\xi}_0)\}\right] \tag{2.6}$$

where H_f is the Hessian matrix of f. The approximation is centred at $\boldsymbol{\xi}_0$ which maximises f and so in this case is the mode of the joint distribution, $\pi(S_n, \boldsymbol{\xi}|\boldsymbol{\beta})$.

In summary, Laplace approximation assumes that the *integrand* is Gaussian and centres the approximation about the *mode* of the joint probability distribution, while Gaussian VA assumes the *posterior* distribution of the latent field (given the point pattern) is Gaussian, with the approximation centred about the *mean*.

2.3.2 Rank Reduction

The second major challenge of fitting a LGCP is that the number of points at which the latent field $\boldsymbol{\xi}$ is represented often needs to be large, and a curse of dimensionality applies. Specifically, large $\boldsymbol{\xi}$ leads to large variance-covariance matrices $\boldsymbol{C}$ and $\boldsymbol{C}_{\text{VA}}$

in Equation (2.5) and a large Hessian matrix $H(\boldsymbol{\xi})$ in Equation (2.6). To compute a likelihood approximation, we are required to invert these matrices and/or compute their determinant, which are computationally prohibitive operations when the dimension is large.

We have already mentioned that this type of issue is common in spatial statistics, but is particularly problematic here because the dimension of $\boldsymbol{\xi}$ is not just a function of the number of point events n, but also the number of quadrature points q used to approximate the spatial integral in Equation (2.3). The number of quadrature points needed is necessarily large when the intensity surface is not expected to be smooth, irrespective of the number of presence points n. Note for example that the predictors used in the flora analyses vary considerably over fine spatial scales (see Figure 1.5), so considerable fine-scale variation in intensity can be expected, and values of q of at least 10,000 are commonly recommended (Renner et al., 2015).

In this thesis we will use fixed rank kriging (FRK; Cressie and Johannesson, 2008) as an approximation to large spatial processes — the term "kriging" comes from the geostatistical sciences and essentially means the predictive interpolation of a process (Cressie, 1993). Some $k \ll n+q$ basis functions $\boldsymbol{Z}(\boldsymbol{s}) = (Z_1(\boldsymbol{s}), \ldots, Z_k(\boldsymbol{s}))$ are chosen to capture spatial dependence so that at any location, the latent field's value is a linear combination of the basis functions with random coefficients $\boldsymbol{u} = (u_1, \ldots, u_k)^{\mathrm{T}}$, so the approximation is given by

$$\xi(\boldsymbol{s}) \approx \boldsymbol{Z}(\boldsymbol{s})\,\boldsymbol{u} \tag{2.7}$$

This reduces the dimension of the marginalising integral from $n+q$ to k, *i.e.* we integrate out the $\boldsymbol{u}$ rather than the latent field — effectively assuming that all stochasticity in the LGCP intensity is attributed to these random coefficients. We assume $u_r \sim \mathcal{N}\left(0, \sigma^2_{\text{prior}}\right)$ for basis function $r = 1, \ldots, k$. Cressie and Johannesson (2008) suggest using basis functions set at multiple spatial resolutions, to better capture a range of scales of dependence. In this case, we let l denote the spatial resolution and assume $u_{r_l} \sim \mathcal{N}\left(0, \sigma^2_{\text{prior};l}\right)$ for basis function $r_k = 1, \ldots, k_l$, typically at two ($l = 1, 2$) or three ($l = 1, 2, 3$) different spatial resolutions.

A range of basis functions can be used in FRK, with Cressie and Johannesson (2008) noting they need not be orthogonal. Further, the specific form of the basis functions only really effects computation time rather than the approximation itself — provided, of course, that they are sufficient in number and coverage to reflect the smoothness (or irregularity) of the surface they approximate. Nychka et al. (2002) look at wavelet functions while Tzeng and Huang (2018) examine thin plate splines, which are both dense and orthogonal, as a means of automatic selection of FRK basis functions. For simplicity and computational speed we use local bi-square functions of the form

$$\mathscr{Z}(d) = \begin{cases} \left[1 - \left(\frac{d}{\varphi}\right)^2\right]^2 & |d| \leq \varphi \\ 0 & |d| > \varphi \end{cases} \tag{2.8}$$

where φ is the function radius and $d = d(\boldsymbol{s}, \boldsymbol{s'})$ is the distance between $\boldsymbol{s}$ and the function location (sometimes called a node or knot) $\boldsymbol{s'} \in \mathcal{D}$. The choice of φ enforces zero values in the basis functions for all points, $\boldsymbol{s}$, at distances beyond φ and is hence a boon to decreasing the computational burden since the resulting $(n+q) \times k$ basis function matrix will be sparse. Implicit in this formulation is that the Gaussian process is isotropic which we believe is a reasonable assumption for the latent field. An anisotropic correlation structure could be induced by non-spherical basis functions but this is not explored here. Hence we need only choose both k and φ. While φ could be chosen via the data we take a more practical approach, setting the radius of effect to ensure a regular grid of basis functions forms a complete cover of the domain, following the defaults of Zammit-Mangion and Cressie (2017) per single resolution of basis functions. This means we need to choose only k which can be done quickly and easily — this is illustrated in Section 2.5. Multiple resolutions can be implemented by choosing k_l with corresponding radius φ_l for each level l. A toy, two-dimensional example of the basis functions described here can be found in Figure 2.1.

We examined the use of various basis functions for the reduced rank approximation, before preferring the local bi-square functions in Equation (2.8). Particularly appealing were those that use the data to select their form (or parameters that control

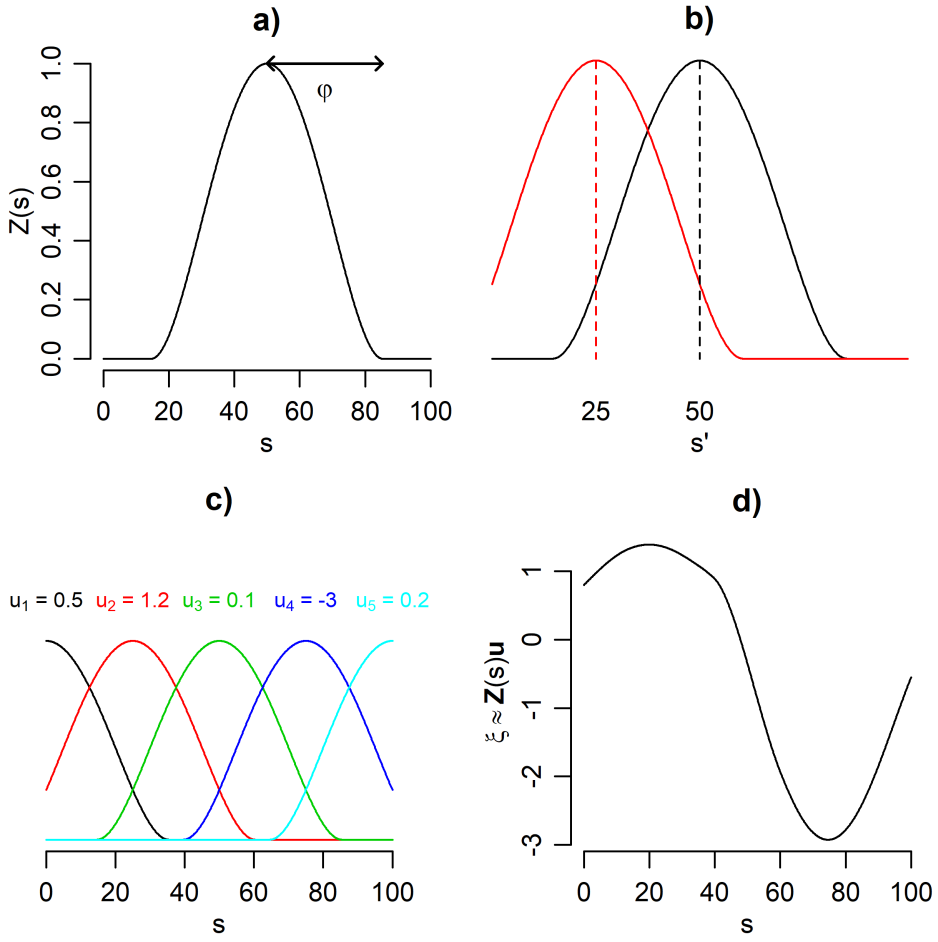

Figure 2.1: Local bi-square basis functions on a one-dimensional domain, $[0, 100]$, used to approximate the latent Gaussian effect. a) A single function with radius φ. b) Locations of basis functions, $\boldsymbol{s'}$ (also called knots or nodes). c) Each $r = 1, \ldots, k$ basis function has a corresponding random coefficient, u_r (which we will estimate from the data). d) Combining the basis functions and random coefficients provides the approximation to the latent Gaussian effect.

their form) as this removes ambiguity around these choices. Predictive processes (Banerjee et al., 2008) fit this category and have some optimal properties in regards to interpolating large, directly observed spatial processes. In fact, these can be thought of as a special case of FRK. However, the basis functions are themselves a function of the covariance parameters we are trying to estimate, which complicates optimisation such that it can be slow and quite unstable. Alternatively, using thin plate splines (as in Tzeng and Huang, 2018) requires no estimable parameters but the denseness of these basis functions (*i.e.* the large number of non-zero values these take throughout domain) also push computation beyond anything of practical use for this application. We find that computational speed depends mostly on whether

basis functions have only local support or span the domain of interest, as well as the number of parameters (or more loosely features) that must be chosen — either arbitrarily or via the data. Hence the appeal of the functions in Equation (2.8) — these have only local support (and so form a sparse matrix) determined by their radius which, on a regular grid, can be fixed by the choice of k.

Rank Reduced Variational Approximation to the log-Likelihood

Putting both variational and rank-reduced approximations together in the context of a LGCP we arrive at a simplified model formulation. The rank-reduced approximation of the GRF means the linear predictor of the LGCP is similar to a generalised linear mixed model, called a spatial random effects model by Cressie and Johannesson (2008). While we developed purpose-written code for the LGCP context, the `FRK` package (Zammit-Mangion and Cressie, 2017) (at time of writing) has in-built functions to apply this technique to spatially correlated responses that are binomial, Poisson, negative binomial or inverse Gaussian, as well as Gaussian responses.

In place of Equation (2.1) we now have

$$\ln \lambda(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u} \tag{2.9}$$

where we are interested in modelling the fixed effects, $\boldsymbol{\beta}$, while the random effects $\boldsymbol{u}$ (together with the basis functions, $\boldsymbol{Z}$), capture additional spatial clustering. Here the stochasticity in the point process intensity is entirely inherited by the $r = 1, \ldots, k$ random coefficients, $u_r \sim \mathcal{N}\left(0, \sigma^2_{\text{prior}}\right)$. When using VA, we approximate the "posterior" probability density for these coefficients (conditional on the observed point pattern, S_n) with the variational density, $\pi_{\text{VA}}(\boldsymbol{u})$, so that $u_r \mid S_n \overset{\text{VA}}{\sim} \mathcal{N}(\mu_r, \sigma^2_r) \implies \boldsymbol{u} \mid S_n \overset{\text{VA}}{\sim} \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We assume that $\boldsymbol{\Sigma}$ is diagonal, *i.e.* the basis functions are able to reflect all spatial correlation in the theoretical finite dimensional realisation of the Gaussian process they are approximating. This means $\boldsymbol{\Sigma} = \boldsymbol{I} \cdot \boldsymbol{\sigma}^2$, where $\boldsymbol{\sigma}^2 = \{\sigma^2_1, \ldots, \sigma^2_k\}^{\text{T}}$ and $\boldsymbol{I}$ is the identity matrix. This assumption greatly improves the speed by which we can compute parameter estimates but means there is no spatially structured covariance to the random coefficients —

again differing from the Laplace approach which permits this via the Hessian matrix, $H(\boldsymbol{\xi})$. The closed form (approximate, marginal) log-likelihood given by our VA method here is

$$\begin{aligned}\underline{\ell}_{\text{VA}}\left(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \sigma^{-2}_{\text{prior}}\right) = &\sum_{i=1}^{n} \boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{\mu} \\ &- \sum_{j=1}^{m} w_j \exp\left\{\boldsymbol{X}(\boldsymbol{s}_j)\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s}_j)\boldsymbol{\mu} + \frac{1}{2}\sum_{r=1}^{k}\sigma_r^2 Z_r(\boldsymbol{s}_j)^2\right\} \\ &- \frac{1}{2}\left[\sigma^{-2}_{\text{prior}}\left(\sum_{r=1}^{k}\mu_r^2 + \sigma_r^2\right) + k\ln\left(\sigma^{-2}_{\text{prior}}\right) + \left(\sum_{r=1}^{k}\ln\left(\sigma_r^2\right)\right) - k\right]\end{aligned} \tag{2.10}$$

Derivation of this can be found in Appendix A.1. Profiling the above with respect to the inverse of the prior variance, $\sigma^{-2}_{\text{prior}}$, can further simplify this objective function. This estimate depends only on the variational parameters at $\hat{\sigma}^2_{\text{prior}} = \frac{1}{k}\sum_r^k(\mu_r^2 + \sigma_r^2)$. A similar result is found when we decide to include multiple resolutions of basis functions, in this case the prior variances are profiled by $\hat{\sigma}^2_{\text{prior};l} = \frac{1}{k_l}\sum_r^{k_l}(\mu_r^2 + \sigma_r^2)$ within each resolution level l. In fact any form of prior variance-covariance (including unstructured) can be profiled under a Gaussian VA, meaning it has a closed form estimate depending only on the variational parameters.

This further highlights the similarity with VA GAMs as formulated by Hui et al. (2019). The inverse of the prior variance plays exactly the role of smoothing parameter since we can consider the Gaussian VA likelihood as a penalised likelihood — as we show in Appendix A.1. The approximate marginal log-likelihood is the expected log-likelihood (with respect to π_{VA}) of the point pattern, which is then penalised by how "far" our variational density diverges from its zero-mean multivariate normal prior. Hui et al. (2019) highlights it as a critical aspect of inference since the VA likelihood simultaneously provides estimates for, and controls the degree of penalisation to, the smoothing coefficients (in our case, the random $\boldsymbol{u}$).

2.3.3 Automatic Differentiation

The final component that permits our novel methodology for fitting LGCP models involves the technical advance of automatic differentiation (AD).

AD is the automatic calculation of the derivative of a programmed function. The technique was developed in the late 1980s (see Griewank, 1989). When programming a function in a low-level language, the accumulative nature of the elementary operations permits calculation of the chain rule with little additional computing cost. This can be exploited to give exact derivatives of the function quickly, irrespective of the number of parameters.

AD is particularly attractive for maximum likelihood frameworks since we can program complicated log-likelihoods involving large numbers of parameters and automatically obtain its gradient information. The benefit of this is threefold. First, including gradient information in our optimisation to fit parameters can speed up the numerical search of the likelihood surface (for example, Shanno, 1970). Next, we can automatically obtain the likelihood's second derivative so that we can estimate Fisher's information for standard errors to our point estimates. Finally, approximating intractable integrals using the Laplace approximation becomes trivial if we are able to program the integrand — as per the ingredients of Equation (2.6).

We program the approximation to the marginal log-likelihood as in Equation (2.10) for our variational model. For our Laplace-based model we program the integrand of Equation (2.4). Both are scripted in `C++` within the Template Model Builder (`TMB`) package in `R` (Kristensen et al., 2016). In addition, `TMB` provides built in Laplace approximations — again computationally efficient due to AD — that we use to arrive at Equation (2.6) for our Laplace-based model. Other examples of software providing a programmatic framework for AD include `AD Model Builder` (Fournier et al., 2012) and the `julia` programming language (Bezanson et al., 2017).

2.4 Simulation Study

We looked to simulations to answer the questions of how quickly and well our proposed methodology fits point patterns — particularly in comparison to `INLA`, the leading `R` package for fitting a spatial LGCP regression model.

Our data were simulated over a square domain with sides of length 100, $\mathcal{D} = \{[0, 100], [0, 100]\}$, as a Poisson point process. The intensity function was a log-linear function of a single, deterministic covariate X, and a zero-mean GRF (ξ) with an isotropic Gaussian covariance function:

$$\ln \lambda(\boldsymbol{s}) = \beta_0 + \beta_1 X(\boldsymbol{s}) + \xi(\boldsymbol{s}).$$

We treated ξ as an unmeasured/unobserved covariate, and hence this formed a LGCP.

Spatial statistics work differently at different scales and so we wanted this reflected in our simulation design. Hence we examined the interplay between the spatial scale of the covariate X and the latent random field ξ. We used a 2×2 simulation design, where each of X and ξ was either chosen to be *wiggly* (W) , with a correlation range of ≈ 5, or *smooth* (S), with a correlation range of ≈ 30 (Figure 2.2). We expect models to more accurately estimate the true data simulation process when the spatial scales of X and ξ do not coincide.

We used the `spatstat` package in `R` (Turner and Baddeley, 2005) to simulate 1000 point patterns from this LGCP within each scenario — hence, we replicated the procedure of fitting the competing models to a point pattern a total of 4000 times. We controlled the size of point patterns simulated through β_0 to examine our simulation scenarios with expected number of points $\mathbb{E}[N(\mathcal{D})] = 200,\ 500,\ 1000$. We standardised the covariate and set the marginal variance of the latent field to 1 so that the magnitudes of model components were roughly equal, to assist convergence in optimisation. We fixed the single covariate effect ($\beta_1 = 1.25$) as preliminary investigation revealed no change in relative performance of the competing models for a varied fixed effect size. To ensure an adequately fine quadrature approximation

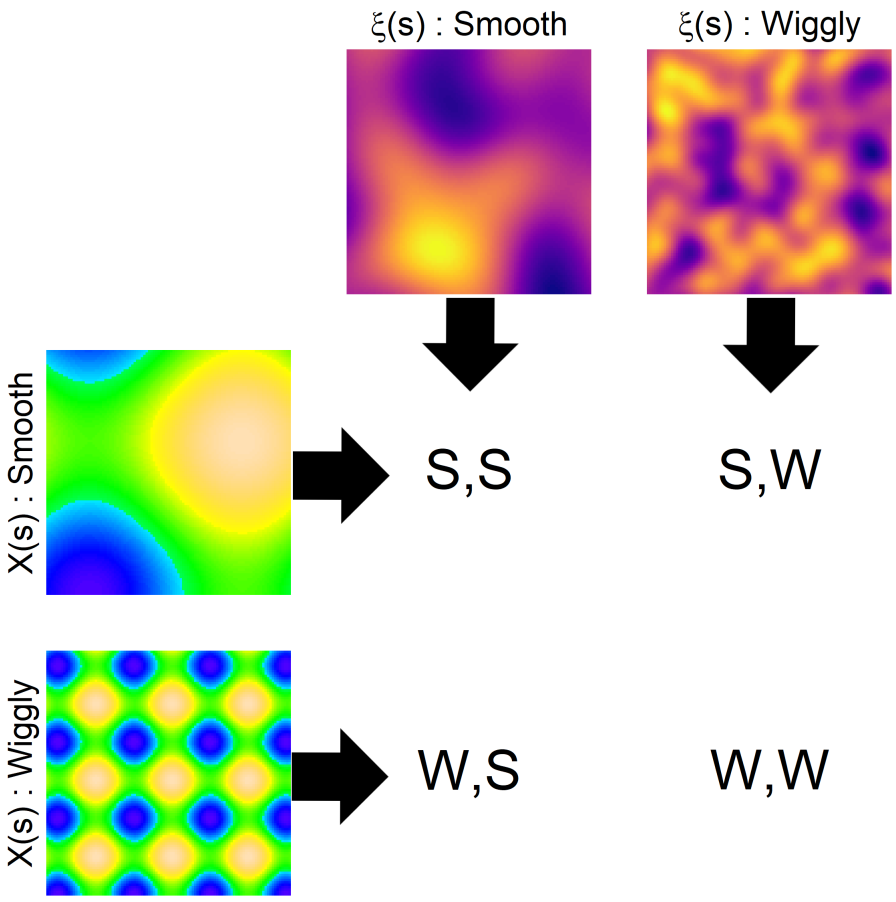

Figure 2.2: The 2×2 simulation design showing the scenarios examined. Deterministic functions are used for the covariate ($X(\boldsymbol{s})$) while particular examples of the latent field ($\xi(\boldsymbol{s})$) are shown here. "Smooth" means a range of effect ≈ 30 while "wiggly" means a range of effect ≈ 5.

(Equation 2.3) we used a regular 101×101 grid of quadrature points.

We fitted point process regression models to the data, assuming intensity was a log-linear function of X and unobserved ξ, using four different procedures:

INLA via the package `INLA`, as the most common approach currently used to fit a LGCP regression model. The GRF was approximated using a stochastic partial differential equation approach (Simpson et al., 2016). This uses a GMRF comprising piecewise linear basis functions (Lindgren et al., 2011) generated via the function `INLA::inla.mesh.2d()`. We use this under default settings given the observed point pattern, which typically produces hundreds of basis functions

VA the proposed methodology of Section 2.3, using a variational approximation to

the likelihood as in Equation (2.10), using FRK with either a sparse grid of basis functions (7×7) or a dense grid (14×14) as in Equation (2.8).

Lp the proposed methodology of Section 2.3, using a Laplace approximation to the likelihood as in Equation (2.6), using FRK with either a sparse grid of basis functions (7×7) or a dense grid (14×14) as in Equation (2.8).

IPP an inhomogeneous Poisson process, *i.e.* with ξ omitted, to study the implications of failing to account for the missing covariate when fitting the model

We expect Lp and VA to perform better when the spatial scale of the basis functions matches the scale of ξ, that is, using a set of basis functions set along a regular 7×7 grid should perform better when ξ is smooth (W,S or S,S in Figure 2.2), whereas using basis functions set along a 14×14 grid should perform better when ξ is wiggly (S,W or W,W). As mentioned previously, we also anticipate difficulties teasing apart effects of X and ξ when they operate at similar spatial scales (S,S or W,W).

We defined how well a method fits by: root mean square error in point estimates, (RMSE $\hat{\beta}_1$); if inference about β_1 is accurate (coverage probability and width of Wald confidence intervals on $\hat{\beta}_1$); and finally, how accurately it can approximate the process's underlying intensity for the purpose of prediction (Kullback-Leibler divergence between the fitted and true $\lambda(\mathcal{D})$). This has the form:

$$\int_{\mathcal{D}} \lambda(\boldsymbol{s}) \ln\left[\frac{\lambda(\boldsymbol{s})}{\hat{\lambda}(\boldsymbol{s})}\right] d\boldsymbol{s} \approx \sum_{i=1}^{q} w_j \lambda(\boldsymbol{s}_j) \ln\left[\frac{\lambda(\boldsymbol{s}_j)}{\hat{\lambda}(\boldsymbol{s}_j)}\right] - \sum_{i=1}^{q} w_j \left[\lambda(\boldsymbol{s}_j) - \hat{\lambda}(\boldsymbol{s}_j)\right] \quad (2.11)$$

Results summarising the various scenarios under investigation can be found in Figures 2.3-2.5. The full and detailed results tables can be found in Appendix A.2.

2.4.1 Simulation Results

In simulations, our methods of fitting a LGCP model were up to 1500 times faster than INLA, converting computation times in some cases from almost an hour to a few seconds (Figure 2.3). IPP was faster again, found to fit nearly instantly, but does not incorporate spatially correlated errors, which comes at considerable costs to performance, as seen below. The expected number of points did not seem to affect

speed of the LGCP models (VA and Lp) when fitting a sparse grid of basis functions (7×7). When the dense grid (14×14) was used in our proposed methods, on average, model fitting takes slightly longer for small point patterns ($\mathbb{E}[N(\mathcal{D})] = 200$) than for the larger ones ($\mathbb{E}[N(\mathcal{D})] = 500, 1000$) — this is a consequence of trying to fit more basis functions than there are data. In the remaining results we restrict our focus on large point patterns, *i.e.* $\mathbb{E}[N(\mathcal{D})] = 1000$.

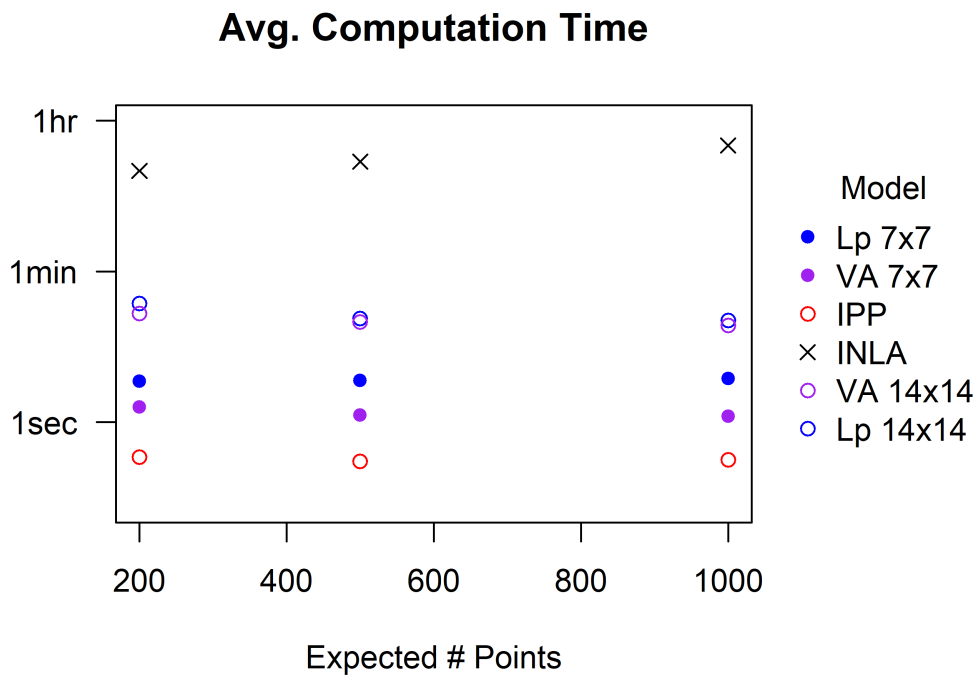

Figure 2.3: Average computation times (including calculating fitted values for the entire domain) over small, medium and large point patterns ($\mathbb{E}[N(\mathcal{D})] = 200,\ 500,\ 1000$). Models include IPP, INLA and our proposed method using variational (VA) and Laplace (Lp) approximations. The latter two models were fitted both using a coarse regular grid of basis functions, 7×7 (filled symbol) and a fine regular grid, 14×14. For small point patterns our methods were at least 36 times faster than INLA on average. For large point patterns they were as much as 1565 times faster than INLA on average.

Figure 2.4 compares point estimates across fitting techniques. It was typically easier to estimate the parameters when the covariate and latent fields were more distinct, *i.e.* acting at different spatial scales, as we found RMSE to be larger in W,W compared to W,S and S,S compared to S,W. Likewise with confidence intervals (Figure 2.5), all models had poor coverage when the spatial scales of the covariate and latent fields were similar.

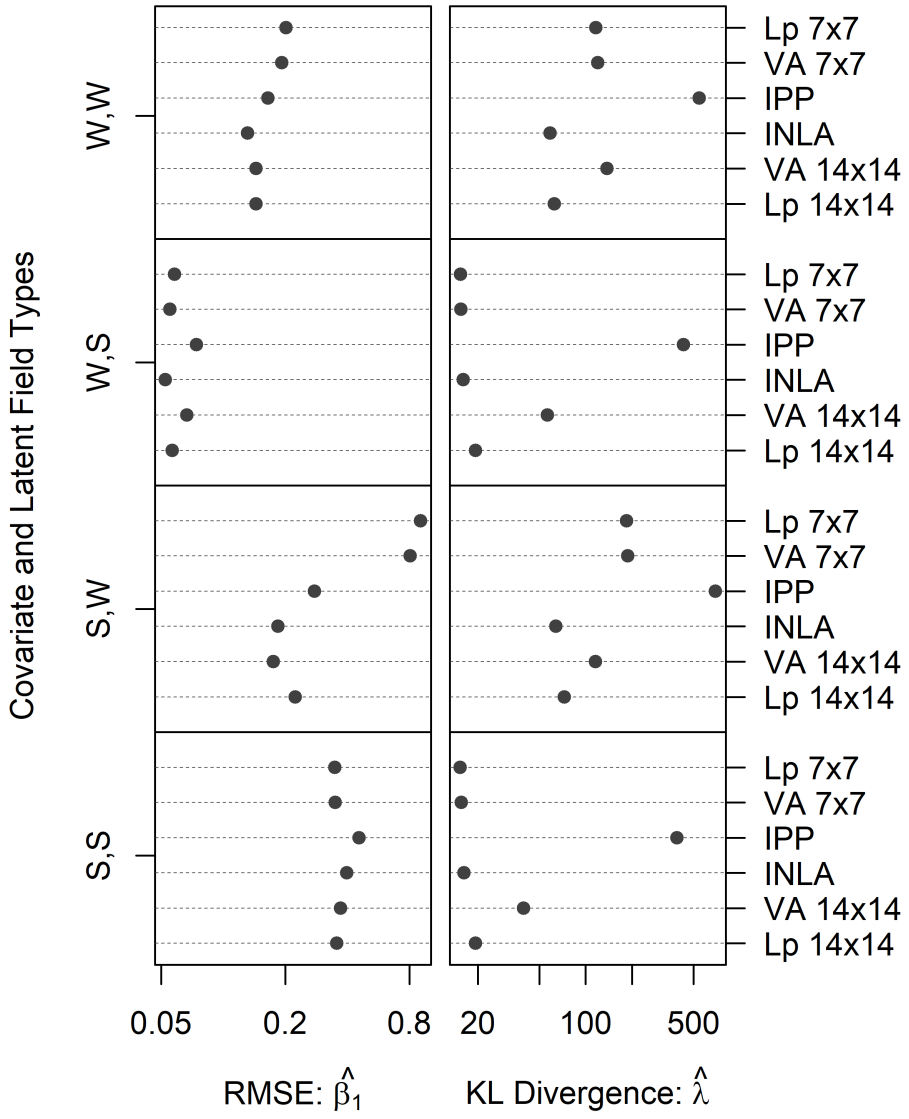

Figure 2.4: Performance of point estimators of the covariate effect (β_1) and intensity (λ), for simulation scenarios using either a wiggly (W) or a smooth (S) covariate and latent field (labelled with the covariate first, *e.g.* "W,S" means the covariate is wiggly and the latent field smooth). The first column shows root mean squared error estimating the slope coefficient β_1. The second column shows the Kullback-Leibler divergence from the true intensity field $\lambda(\mathcal{D})$ to that fitted by the model. We found that our proposed methods performed comparably to INLA in point estimation of β_1, and for Lp, also in estimation of intensity, provided we choose the most appropriate basis function configuration for the scenario. But note that even at the appropriate basis function configuration, Kullback-Leibler divergence was noticeably larger for VA than for INLA in all simulation scenarios.

Table 2.1: Model selection results for choosing the basis function configuration in simulations. The table reports the number of times, out of 1000 simulated datasets, where each basis function configuration (either 7×7 or 14×14) had the higher (approximate) marginal log-likelihood, for each of Laplace and Variational model fits, and for each simulation scenario. Note also that occasionally a Laplace fit failed to converge (Fit Fail) or the `spatstat` simulation failed to produced a point pattern (Sim. Fail).

	Laplace			Variational			
Scenario	14×14	7×7	Fit Fail	14×14	7×7	Fit Fail	Sim. Fail
S,S	19	927	51	0	997	0	3
S,W	954	0	46	789	211	0	0
W,S	16	974	9	0	999	0	1
W,W	1000	0	0	219	781	0	0

It does appear important that we select an appropriate number of basis functions for the type of latent field we are approximating — we needed many basis functions (14×14) when the latent field was wiggly (S,W or W,W) and fewer (7×7) when the latent field was smooth (S,S or W,S). Further, the cost of using too few basis functions (7×7 for S,W or W,W), is more than the cost of using more than are needed (14×14 for S,S or W,S), as expected, since bias tends to be more costly than overfitting. Our methods did as well at point estimation of β_1 as INLA's maximum *a posteriori* estimate, and the VA approach seemed slightly more accurate in this regard than the Lp approach. The opposite is true when comparing confidence intervals, where VA did more poorly than Lp on coverage — seemingly because of narrower confidence intervals (Figure 2.5, all panels). INLA was good at fitting the true intensity field but was matched by Lp, again, provided the basis function choice was appropriate. The VA approach seemed to generally do poorer than the corresponding Lp version at fitting λ, this was particularly the case when these methods were using many basis functions (14×14).

While choice of the number of basis functions (k) is key, the data could be used to guide this decision, as is typically done using standard information criteria like AIC (Akaike, 1973). But here we can simply select k to maximise the likelihood, since k is the dimension of a random effect and it doesn't affect the penalty term in common information criteria. We looked at how often this strategy would choose the more appropriate basis function configuration, in each simulation scenario.

Table 2.1 shows that, for the most part, we found a higher likelihood for the ba-

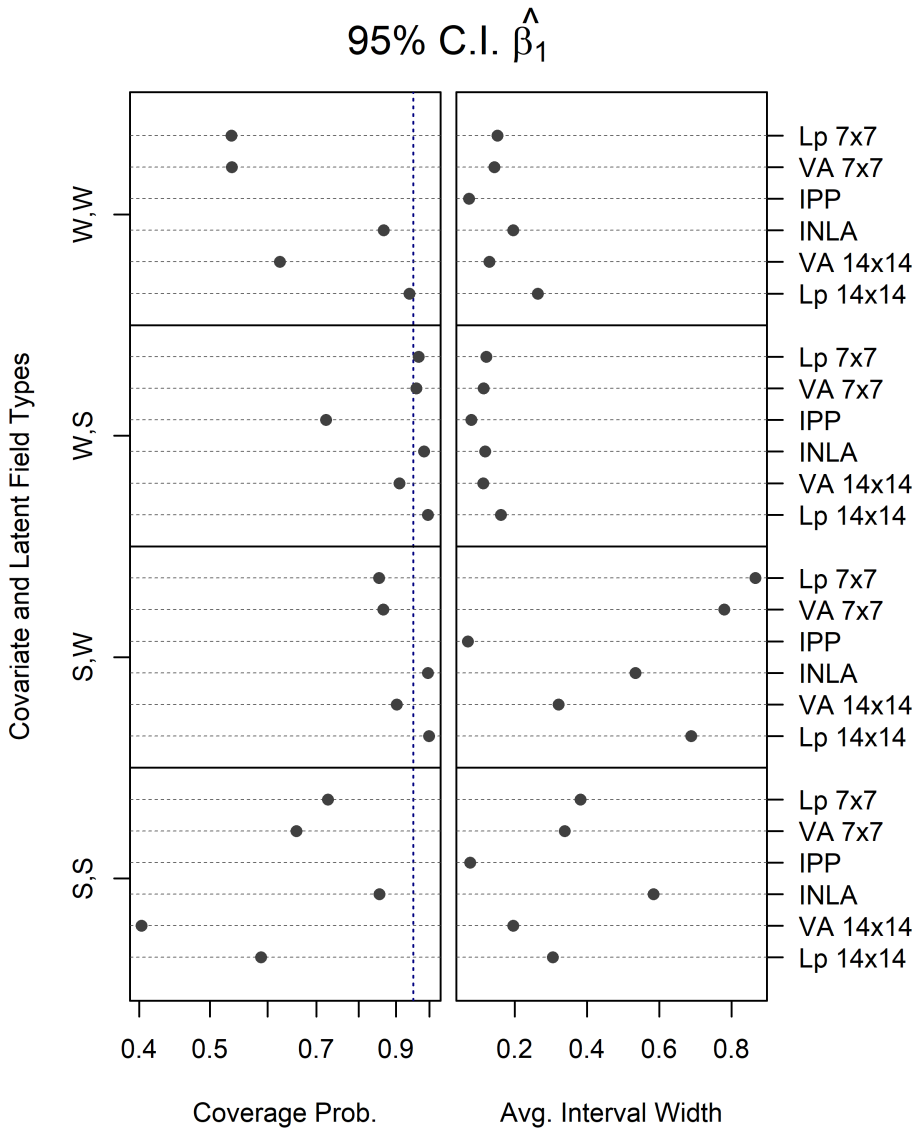

Figure 2.5: Performance of interval estimators for β_1, for simulation scenarios using either a wiggly (W) or a smooth (S) covariate and latent field (labelled with the covariate first, *e.g.* "W,S" means the covariate is wiggly and the latent field smooth). The first column shows coverage probabilities of 95% Wald intervals for the slope estimate, $\hat{\beta}_1$. The dashed blue line indicates a coverage probability of 95%. The second column shows the average widths of these confidence intervals. We found that, provided we choose the most appropriate basis function configuration for the scenario, the Lp version of our proposed method to be comparable to INLA in coverage, although its intervals tended to be wider. The VA version of our method tended to produce more narrow confidence intervals.

sis configuration that better suited the scenario, *i.e.* there tended to be a higher likelihood for a model with more basis functions when the latent field was wiggly (S,W and W,W). Conversely, there tended to be a higher likelihood when using less basis functions when the latent field was smooth (S,S and W,S). The exception to this is for the VA approach, where we had both a wiggly covariate and latent field. Note however that VA 7×7 seems to better estimate the intensity than VA 14×14 in this scenario anyway (Figure 2.4, right column, top panel) so there seems to be little cost to this behaviour. We also note that the Lp approach is far more sensitive to non-convergence as we found approximately a 5% fit failure rate for large point patterns — we discuss this further in Section 2.6.

2.5 Application: Gorilla Nesting Locations

We illustrate our proposed method by analysing the gorilla nesting dataset described in Section 1.1.1. Recall that we wish to model intensity of gorilla nests as a log-linear function of elevation above sea level (X_1); distance to nearest water source (X_2); and average temperature (X_3, as a three level ordinal factor). These are found in Figure 1.5b-d and are parameterised with β_1, β_2 and β_3 respectively. The model also includes an intercept term, β_0.

2.5.1 Methods

Here we aim to perform a model assessment exercise - namely a hold-one-out, four-fold cross-validation (CV) that predicts the likelihood in held out test areas of the domain. Figure 2.6a shows the spatially blocked CV folds used — call each fold $\mathcal{D}_h$ for $h = 1, 2, 3, 4$ so that $\mathcal{D} = \cup_{h=1}^{4} \mathcal{D}_h$. We compared model fits using predicted conditional log-likelihood, summed over each fold since $\mathcal{D}_h \cap \mathcal{D}_{h'} = \emptyset$ for any $h \neq h'$. That is

$$\sum_{h=1}^{4} \ln \pi \left(S^{(h)} | \hat{\boldsymbol{\beta}}^{\setminus h}, \hat{\xi}^{\setminus h} \right) = \sum_{h=1}^{4} \left[\sum_{i=1}^{n_h} \boldsymbol{X} \left(\boldsymbol{s}_i^{(h)} \right) \hat{\boldsymbol{\beta}}^{\setminus h} + \hat{\xi}^{\setminus h} \left(\boldsymbol{s}_i^{(h)} \right) - \sum_{i=n_h+1}^{n_h+q_h} w_i^{(h)} \exp \left\{ \boldsymbol{X} \left(\boldsymbol{s}_i^{(h)} \right) \hat{\boldsymbol{\beta}}^{\setminus h} + \hat{\xi}^{\setminus h} \left(\boldsymbol{s}_i^{(h)} \right) \right\} \right] \quad (2.12)$$

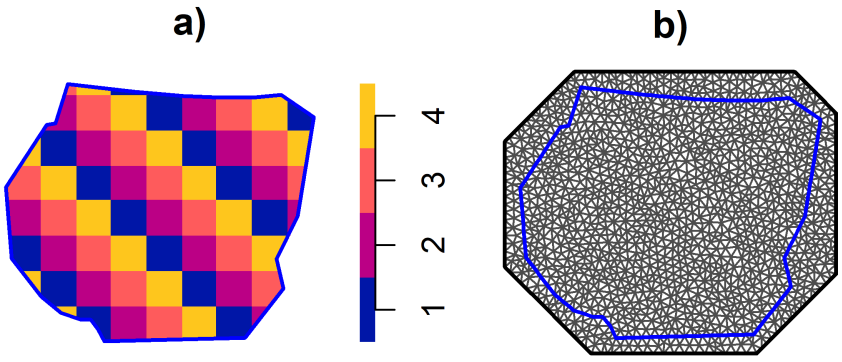

Figure 2.6: a) The partitioning of the domain for a four-fold cross validation. b) Mesh used for INLA is a fine scale Dirichlet tessellation of 1479 vertices.

where $S^{(h)}$ denotes the sets of presence points, $\left\{s_i^{(h)}\right\}_{i=1}^{n_h}$, and quadrature points, $\left\{s_i^{(h)}\right\}_{i=n_h+1}^{n_h+q_h}$, in fold h. $w_i^{(h)}$ denotes the size/weight at the i^{th} point in h, so that for $i = n_h + 1, \ldots, n_h + q_h$ these are quadrat sizes. $\hat{\boldsymbol{\beta}}^{\setminus h}$ are the fixed effect parameters estimated from the data excluding fold h. $\hat{\xi}^{\setminus h}$ is likewise the estimated latent field from the training data without fold h — the form this takes depends on how it is estimated within each fitting method we compare. For our proposed methods $\hat{\xi}(\boldsymbol{s}) = \boldsymbol{Z}(\boldsymbol{s})\,\hat{\boldsymbol{\mu}}$, for INLA this is the maximum *a posteriori* estimate on the log-linear scale.

As in the previous section, we modelled the gorilla nesting point pattern using four methods for comparison — an IPP using standard software, as well as an LGCP using the `INLA` package (again we subsequently refer to this as INLA) and two versions of our proposed methodology, *i.e.* using the variational (VA) and Laplace (Lp) approximation for the marginalised likelihood. Each method used all available predictors (Figure 1.1b-d) for fixed effects in addition to an intercept term. We saw in Figure 1.2 (top panel) that there was strong evidence that an IPP is not a valid assumption for this point pattern, so did not expect it to be competitive with the corresponding LGCPs. The more appropriate LGCP model was fitted with INLA, VA, and Lp, each of which include a latent GRF approximated using different means — INLA used the stochastic partial differential equation approach of Simpson et al. (2016), while our methods used FRK (Cressie and Johannesson, 2008).

The LGCP models examined here required us to choose an appropriate structure

for the GRF approximation. For INLA, decisions needed to be made concerning the mesh to evaluate the intensity at, and distributions to put on priors of parameters. This dataset has previously been analysed using the INLA framework in Bachl et al. (2019), with detailed accompanying code that includes a mesh built off a Dirichlet tessellation of the sanctuary shown in Figure 2.6b. We used this mesh for our analyses.

To implement our proposed reduced rank method, we needed to choose a basis functions configuration, which was informed by the data. We used a regular grid of k local bi-square basis functions (as in Section 2.3.2) and chose k to maximise either the likelihood (as in Table 2.1) or the predictive conditional likelihood (Equation 2.12). As the provided INLA mesh comprises 1479 vertices (effectively basis functions), we examined FRK basis function configurations ranging from $k = 0$ (*i.e.* an IPP) to $k < 1500$.

2.5.2 Results

Our main result is that our proposed method reduced the computation time required for a single model fit from minutes to seconds, with negligible loss of predictive performance. Specifically, our four-fold cross-validation procedure took over eight hours to complete on `INLA` using vanilla code following Illian et al. (2012) and mesh from Bachl et al. (2019), whereas the same procedure, using a variational approximation with the appropriate number of basis functions, took just 10 seconds (Table 2.2) and had slightly higher predictive conditional likelihood. To give a sense for the extent to which this speed up is due to the number of basis functions, we refitted our method using as many basis functions as INLA, and it slowed down somewhat (six minutes) but was still clearly much faster than INLA. Our method became inaccurate with this number of basis functions because of overfitting — recall that there are only 640 presence points in this dataset, and we would expect a sensible choice of the number of basis functions to be less than this number. Finally, as expected, IPP was much faster to fit than other methods, but clearly sub-optimal, with a very low predictive conditional likelihood.

We arrived at the decision to use $k = 63$ basis functions using Figure 2.7. The

Table 2.2: Four-fold cross-validated predicted conditional log-likelihoods of the various models and corresponding computation times. We see that our proposed methods (LP and VA) achieved higher predicted conditional log-likelihoods in under a minute compared to INLA that took over 8 hours. With a similar number of basis functions to INLA our method did poorly at prediction but computed the four-fold CV in ≈ 6 minutes.

	IPP	INLA	LP ($k = 63$)	VA ($k = 63$)	VA ($k = 1470$)
Predicted $\ell(\boldsymbol{\beta}\|\boldsymbol{\xi})$	1727.5	2243	2299.8	2301.4	1653.4
Comp. Time	0.72 sec	8.56 hrs	50.82 sec	10.67 sec	5.98 mins

variational likelihood was maximised at $k = 63$, whereas the Laplace likelihood was

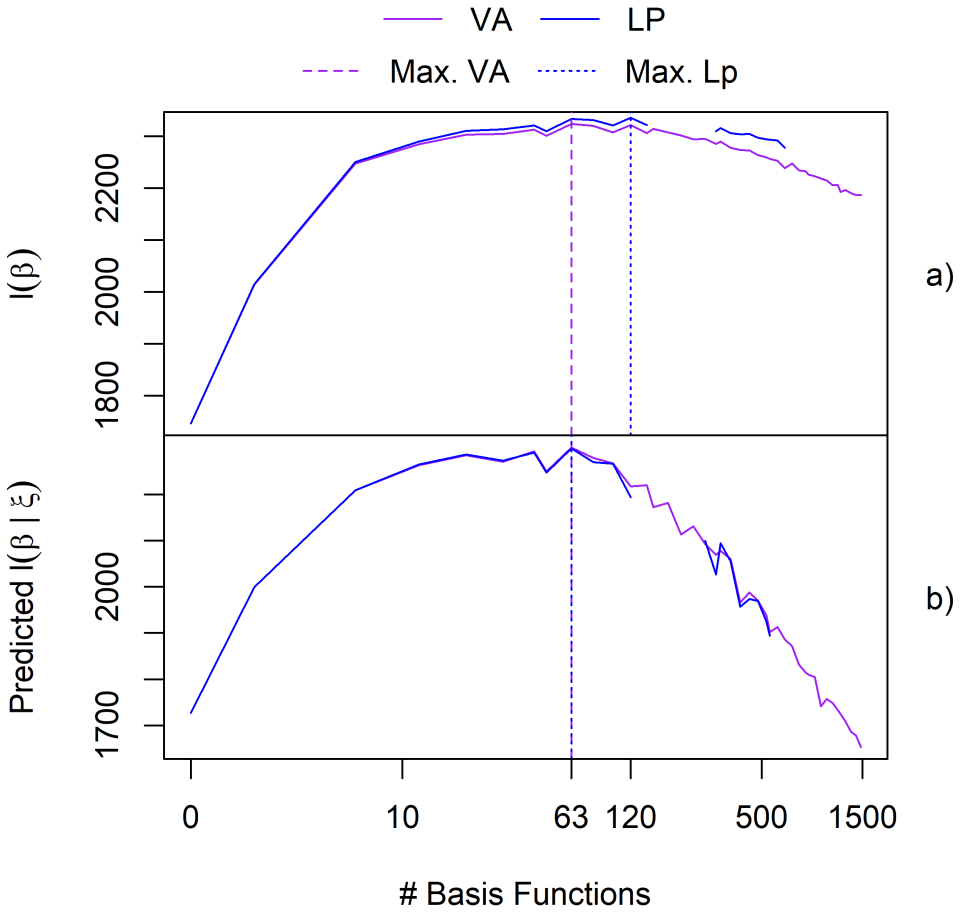

Figure 2.7: a) The approximate marginal log-likelihood of the proposed methodology as a function of increasingly dense regular grids of local bi-square basis functions. b) The predicted log-likelihood (conditional on the latent field, estimated using four-fold CV), as a function of increasingly dense regular grids of local bi-square basis functions. Vertical lines denote the basis functions at which the maximum is found for the variational (purple, dashed) and Laplace (blue, dotted) methods. From the fitted likelihood signal we would choose a configuration of 63 basis functions for the VA method and nearly double, 120, for the Laplace method. However, for predictive performance we found both methods did best with the 63 basis function configuration.

maximised at a slightly larger number of basis functions. For predictive conditional likelihood (Figure 2.7b), both approximation methods suggested $k = 63$.

While fitting a single LGCP can be done very quickly using proposed methods, computation time will be considerably longer if we need to fit multiple models in order to find the desired basis function configuration. For Figure 2.7, we fitted many LGCP models with different numbers of basis functions, and additionally in Figure 2.7b we used cross-validation as a tool for model selection. The total computation time for VA results was 20 minutes in Figure 2.7a and 80 minutes in Figure 2.7b. These times were exaggerated by the decision to try to fit models with approximately as many basis functions as INLA (for which $k = 1479$), when more than 200 clearly led to overfitting. The Laplace method scaled poorly with number of basis functions, with total computation times of 7.5 hours and 28 hours respectively for Figure 2.7a and Figure 2.7b. No attempt was made to explore different basis function configurations for INLA, given that a single fit took over 8 hours.

Further examination of the differences in these LGCP fits can be found in their fitted (log-)intensities and fixed effect estimates. Figure 2.8 shows the intensities and reveals that INLA produced a far more detailed intensity surface (Figure 2.8b) to that produced by our methods (Figure 2.8c-d) in regions to the central-south of the domain where very few data are present. This is probably because a much larger number of basis functions were used in estimating the intensity surface in the INLA fit (1479 compared to 63). We wonder if this large number led to overfitting in this region. Conversely, in the upper left corner of the domain INLA appears to estimate high nest density along the boundary where presence points are lacking and where our methods predict a decreasing intensity accordingly. As the upper left corner is the only domain boundary with nearby point events, we suspect this is due to the wide boundary-extension given to the mesh (Figure 2.6b), which is smoothing the intensity surface into the exterior of the observation window. We note that INLA did not have better predictive performance.

Results generally appear similar to our previous simulations involving a long correlation range on the latent field (*i.e.* a smooth latent field as in S,S and W,S scenarios in Section 2.4) where a coarse basis function configuration on our methods was able

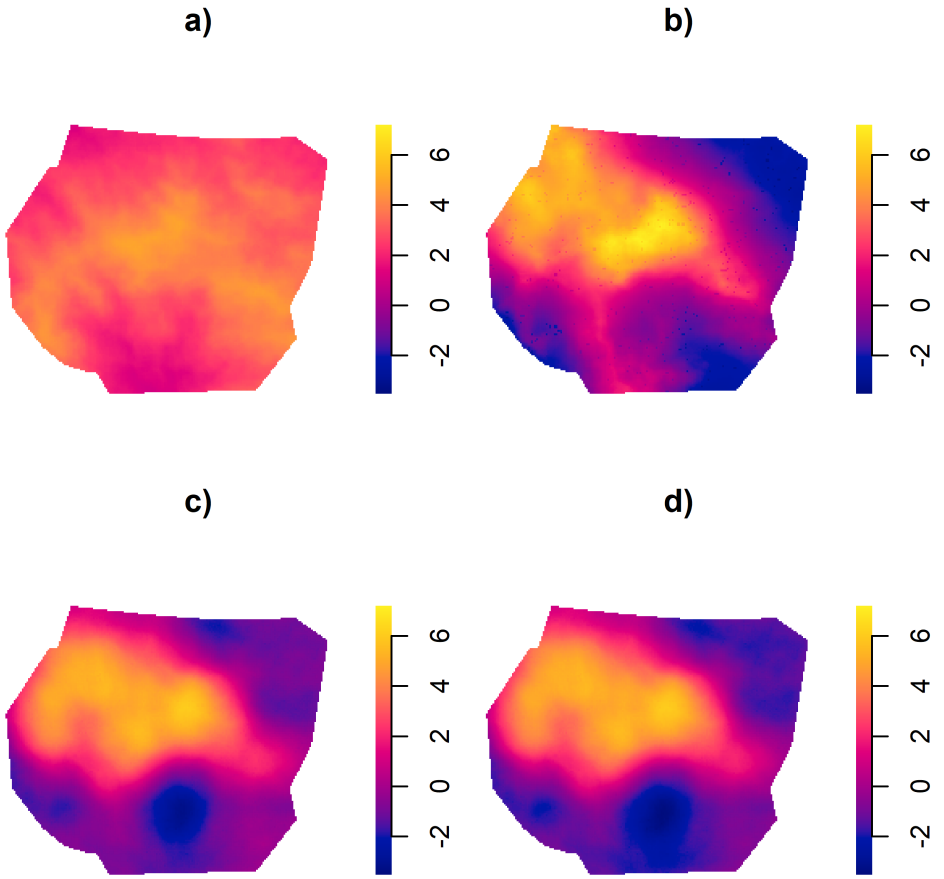

Figure 2.8: Fitted log-intensities for each of the models being compared: a) IPP b) INLA c) Lp d) VA. We see that INLA produced a far more detailed intensity surface than VA and Lp. This also shows the comparatively flat intensity surface estimated by the IPP model.

to better fit the true intensity of the LGCP than INLA. Indeed, the maximum *a posteriori* estimate for the correlation range parameter for the GRF as estimated by `INLA` is $\approx$ 1.6km, likewise, our optimal basis configuration in Figure 2.7b has radii of $\approx$ 1km. The length of the spatial domain is about 5.5km, so the estimated latent effect is very smooth.

We also constructed point and interval estimates of the fixed effects parameters in all models (Figure 2.9). The most conspicuous difference was that the effect of elevation (β_1) was estimated to be larger for IPP, with an unrealistically small standard error, which can be attributed to undue confidence in parameter estimates due to lack of a random field to account for spatial clustering. Amongst the LGCP fits, parameter estimates were generally similar, although confidence limits were longer for INLA

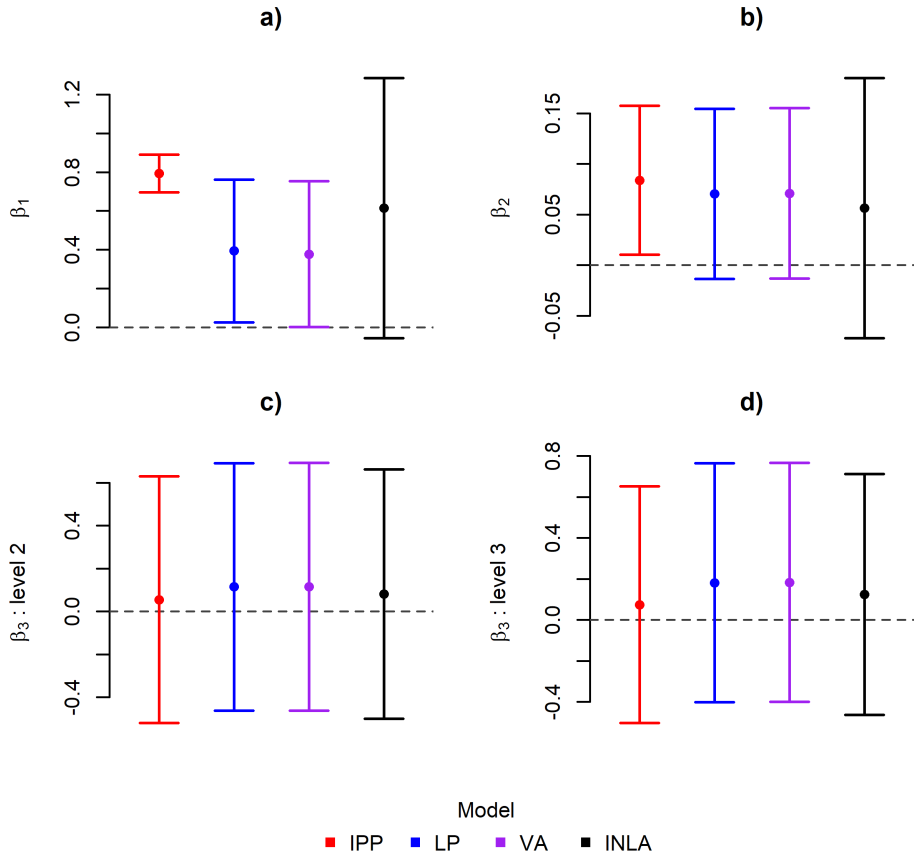

Figure 2.9: Estimated fixed effects 95% Wald confidence intervals for the various models being compared. a) Elevation effect. b) Distance to water effect. c) Heat category — contrast effect of Moderate to Coolest. d) Heat category — contrast effect of Warmest to Coolest. We see little difference in the estimated fixed effects for the LGCP models (INLA, VA, Lp), so much of the differences in predictive performance came down to the way the latent effect was modelled (or was not modelled for the IPP).

than Lp or VA, especially for elevation. Note that elevation was the covariate that was smoothest spatially (Figure 1.1b), varying at a comparable spatial scale to our basis functions, so spatial confounding appears to be an issue here (analogous to S,S in simulations). This was our worst case scenario from simulations (Figure 2.5) and we should not put a lot of weight in inferences about the elevation effect using any of the methods considered. As in Figure 2.4, even under this scenario we found both INLA and our method (using an appropraite number of basis functions) to most reliably estimate the fitted intensity. That is, we do not accurately untangle the contributions of the fixed effect and latent field but can nonetheless quite accurately model the resulting intensity.

Table 2.3: The computation time and likelihood for a single model fit for our proposed model for which we toggled on/off key components — a variational approximation for the marginalisation (VA, *vs* Laplace approximation, Lp); automatic differentiation (AD); sparse basis approximation for the latent field (Sparse / Dense $\boldsymbol{Z}$); and large number of basis functions (Large k). INLA is also included for comparison.

Method	Sparse/Dense $\boldsymbol{Z}$	AD	Large k	Comp. Time (sec)	$\ell(\boldsymbol{\beta})$
VA	Sparse	✓		4.30	2323.78
Lp	Sparse	✓		16.10	2333.62
VA	Dense	✓		41.80	2271.96
VA	Sparse			93.60	2323.78
VA	Sparse	✓	✓	96.19	2187.01
INLA			✓	718.87	2336.70

Finally, we want to get a sense of which components of our novel methodology contributed most to the substantial improvements in the speed with which models can be fit. We compared the fit times of the gorilla nesting model used in this section when using our method and leaving out each component — *i.e.* variational approximation for the marginalisation; sparse basis functions for the latent approximation (as described in Section 2.3.2); and automatic differentiation for gradient information in optimisation. Table 2.3 shows the computation times involved in fitting a single model when systematically excluding each component. Additionally, for comparison we include information for our model using the same number of basis functions as INLA. When fitting a "Dense" $\boldsymbol{Z}$ we used a thin plate spline basis (as in Tzeng and Huang, 2018) instead of the "Sparse" local bi-square basis (Equation 2.8). As an alternative to AD, we programmed and fitted in `R` using an optimiser without gradient information. When we examine the various component models we see that each contributes substantial relative gains in computation speed. The largest speed gains appear to be due to use of AD and use of a small number of basis functions, relative to INLA. The variational approximation yielded the most modest speed gain here, but it still offered an almost four-fold improvement in computation time compared to using a Laplace approximation.

2.6 Discussion

Our motivation for this chapter was to provide a fast method for fitting LGCPs to point patterns, to make the practice more accessible to researchers. In ecology, researchers currently use only an IPP in most instances — *i.e.* assuming there is no clustering in the pattern beyond that accounted for by predictors — despite this often being an unreasonable assumption. Our proposed methodology for fitting LGCP to point patterns proved to be orders-of-magnitude faster than existing methods in the literature.

We were able to achieve this considerable computational speed-up using three key components — as seen in Table 2.3 it is in combination that these permit such fast fitting. The larger speed advantages seem to come from AD and coding key steps in `C++`, as well as being able to use a small number of basis functions (for little loss in terms of fitted likelihood). We also found considerable computational savings from using a sparse FRK (Cressie and Johannesson, 2008) to approximate the latent GRF. Finally, VA provides an approximate closed form solution to the marginal likelihood that can be fit more quickly than using the more common Laplace approximation. The VA approach scales far better with number of basis functions, k, than the Laplace approach (Section 2.5). Using these advances and working in a maximum likelihood framework allows researchers to fit LGCP models in ways that have previously been computationally prohibitive and perform inference using likelihood-based statistical tools.

There are some key choices to be made about the number and nature of basis functions used for the rank reduction when implementing our method. Cressie and Johannesson (2008) suggest that the specification of basis functions and their locations is not particularly important. In the current context we would agree — considering the process we intend to approximate is not directly observed at all. Since our method fits very quickly we are able to fit and compare different basis configurations easily as we showed in Section 2.5. Similarly, we can combine the choice of number and range of basis functions as described in Section 2.3.2 and then use the signal within the changing likelihood to select an appropriate configuration,

as in Table 2.1 and Figure 2.7. Exploring different basis function configurations erodes some of the computational gains of our method, with an exhaustive basis search to choose k, along the lines of Figures 2.7b, taking over an hour on VA. To do a basis search efficiently we would recommend: that VA be used as it scales better with large k, and that Lp (if desired) only be applied once a k has been chosen; that the total number of basis functions considered be kept less than the number of presence points (*e.g.* $k \leq \frac{n}{2}$). Finally, we note that choice of basis function configuration might be important to performance of INLA as well, but exploring this issue would be computationally prohibitive on INLA in most applications.

Further investigation could be made into using multiple resolutions of basis functions, as recommended in Cressie and Johannesson (2008). In Section 2.4 our simulated truth only ever had a latent influence coming from a single field with constant range of effect, so multiple resolutions added nothing to the model performance except potential problems with over-parameterisation. In real data cases where a researcher may be missing various influential covariates/phenomena multiple resolutions may capture this. In our simulations we found point estimation and inference to be difficult when the scale of the single covariate and latent field coincided (scenarios S,S and W,W) and so multiple resolutions of basis functions may only increase the chance of such spatial confounding (Hodges and Reich, 2010). The merits of including more latent spatial effects in the model seems to depend whether the main aim of the research is to accurately predict the density of point events at certain locations, or to understand their relationship to particular measured phenomena.

VA provides a fast alternative to the Laplace approximation for the marginal likelihood we seek to fit. However, we saw in Section 2.4 that the VA version of our proposed method tended to underestimate error in point estimates, even in scenarios where the latent field was varying over a different spatial range to the spatial covariate. Underestimation of errors was also noticed in Hui et al. (2019), who suggest using the method of Louis (1982) to obtain a more robust estimator of the observed information matrix. The Louis (1982) method was designed for missing data analysis, so by using this method in this context, we are arguing that the variational parameters represent missing data. Given that the variational parameters are used

to estimate an unobserved random effect, this is somewhat defensible. The Laplace version of our method tended to do better at estimating standard errors but is not without issue. As we saw in Table 2.1 and Figure 2.7, it is prone to convergence failures in optimisation, often due to computational singularity in the Hessian. Similar problems have been found in Brooks et al. (2017) where scaling of predictors is a recommended solution. This is no help here, however we find providing warm starts within the parameter space can reduce failure rates. This includes providing a warm start for the prior variance(s) on the random effects, which could come from our VA version. A good practice might be to start with an exploratory VA version, as this is quicker and more robust, then use warm starts for parameters in the Laplace fit, and examine changes in Wald confidence intervals between the two.

Our proposed approach to fitting LGCPs offers speed gains so large that we can consider using LGCPs in different ways to how they have been used previously. We have already exploited this speed gain to perform model selection to choose the number of basis functions, by refitting our model on different basis function configurations (Figure 2.7). This would not be computationally feasible using INLA, where our cross-validation analysis took over eight hours to complete for a single basis function configuration. In addition, it would seem feasible to extend our model to handle more complex data structures. In ecology, there is interest in combining data sources (for example, Dorazio, 2014; Miller et al., 2019), including presence records of species which form spatial point patterns. Similarly, joint species distribution models are frequently constructed (for example, Pollock et al., 2014; Wilkinson et al., 2019; Hogg et al., 2021) in ecology and it would seem possible to fit these using presence records for multiple species. In both of these cases, some previous work has been done in the presence-only context, but under the assumption that the point patterns arise from an IPP (Fithian et al., 2015; Koshkina et al., 2017; Fletcher Jr et al., 2019). The main idea behind fitting models that integrate different data types, or multiple species, is to share information about measured predictors to improve model accuracy. An LGCP framework could be used to also share information about *unmeasured predictors* via shared latent fields. In Chapter 3 we will extend the methods presented here to combine presence/absence records with

presence-only records.

Chapter 3

Extending Data Integration Methods in Ecology

3.1 Introduction

Information on where a species has been found can come in a number of different forms. Previously we have discussed presence-only data, presence locations forming a point pattern, a data source becoming increasingly accessible with the rise of citizen science programmes and naturalist community platforms (Botella et al., 2021). Such data often are biased due to the uncontrolled nature of its collection, *e.g.* due to location accessibility. Another commonly available data type is presence/absence data, sometimes called site occupancy or survey data, the result of systematically sampling pre-determined areas and recording whether a species is present or absent. The binary response is modelled using a regression that is conditional on the sites that were sampled, which removes bias due to site selection (provided that sites were selected based on characteristics not directly related to the response variable, species presence/absence). However, presence/absence data may generally be less available as they come at a greater cost (in both time and funding). Further, transects are often selected to be representative of different environments and geographies, rather than specifically in search of a particular species, so presences may be very infrequently recorded, especially for rare species. One way to overcome the potential shortcomings of presence-only data and presence/absence data is by analysing both datasets using a single model. This is an example of what is often referred to as *data integration* or fusion. Miller et al. (2019) provide a review of some of the many forms this takes in the ecological literature.

A large subset of data integration methods are based on joint likelihood approaches which allow parameters to be shared across the data sources being combined. A key paper in this area was written by Fithian et al. (2015), who combine presence-only and presence/absence data and show that doing so can improve predictive performance when also pooling information across species. Dorazio (2014) likewise provided an early framework for integrating data from both planned and opportunistic surveys, with a focus on incorporating detection probability directly into the model. More recent examples include Koshkina et al. (2017) and Fletcher Jr et al. (2019). All these examples, however, implicitly assume both spatial independence

between observations within each dataset and independence between datasets, after conditioning on model predictors. We saw in Chapter 2 that the first assumption is often unrealistic. Measuring every covariate that drives the spatial distribution of a species is near impossible, let alone including them correctly in the model. Importantly, this form of model misspecification will affect both presence-only and presence/absence models in the same way, so it will induce dependence between data sources. Until recently there has been limited research that addresses dependence in integrated models, and it tends to only address dependence within rather than across datasets. Renner et al. (2019) addresses the issue of additional clustering in the presence-only data by using an area interaction model when combining data describing the distribution of Eurasian lynx. However, independence was assumed between the presence-only and presence/absence data sources.

Pacifici et al. (2017) combined presence/absence survey data and count data arising from a large scale citizen-science project with (measured) variable-effort. This data integration approach included spatially correlated random effects across data sources in a multivariate conditional auto-regressive model. This research is analogous to what we propose here except that our focus is on presence-only data arising from point processes over a continuous spatial domain, while Pacifici et al. (2017) conditioned on spatial locations where data where collected. A paper in pre-print (Watson et al., 2019) examines integrating various presence/absence datasets describing sightings of southern resident killer whales. Separate sets of covariates are proposed to model detection probability and observer effort that, along with a LGCP model, fully describe the true intensity process of the observed presences. All the presence/absence data are considered a point pattern and absences are only used to inform parameters that control the detection probability and observer bias. While the LGCP component permits shared clustering due to potentially missing covariates, as well as the sharing of effort and detectability parameters, Watson et al. (2019) acknowledge that properly estimating these wide range of sources of dependence can become problematic. Their model is implemented in INLA which we saw in Chapter 2 can be computationally burdensome.

In this chapter we propose a new methodology that extends the work of Fithian

et al. (2015), as applied to a single species, by assuming each data source shares a latent field (along with a common response to environmental influences) that can account for dependencies across data sources that may be the result of missing environmental predictors. Our method assumes that clustering and biases unique to the presence-only data can be adequately modelled with the inclusion of biasing covariates. The framework and methodology is outlined in Section 3.3 and builds on the developments of Chapter 2 for LGCP regression models. We assess performance of the method (relative to separately modelling the data sources) through simulations in Section 3.4. In Section 3.5 we analyse real data examples of flora in the Greater Blue Mountain World Heritage Area. The chapter is concluded with discussion in Section 3.6.

During write-up for the current work, we became aware of the work of Simmonds et al. (2020) that proposes extending the data integration method of Fithian et al. (2015) in a very similar way to that proposed here. Both bodies of work use a LGCP framework to share a latent field between data sources with the goal of accounting for dependence structures otherwise ignored. Both examine simulations in a variety of scenarios to determine when data integration leads to improved model performance. A key difference is that in simulations, we permit the same realised latent field across both datasets — playing the role of constant and missing environmental drivers of the species distribution. Simmonds et al. (2020), on the other hand, allow each dataset to have a separate realisation of the latent field — while having similar ranges of effect — to account for the data sources being collected at different times. However, if the latent field represents missing environmental variables, we would argue that the same environmental variables are missing from the model for each data source, hence the same latent field realisation is needed in the integrated data model. In this sense, the role of latent field is somewhat different to that we assume here. Other key differences are that we approximate the latent field using spatial random effects ("FRK", as in Chapter 2) while Simmonds et al. (2020) use the SPDE approximation within INLA, leading to substantial differences in computation time, and constraining their capacity to perform extensive simulations. Whereas Simmonds et al. (2020) evaluate their methodology using simulation only,

we apply the proposed methodology to real datasets as well as studying properties of our method via simulation. Another relevant paper is Conn et al. (2017), where abundance in aerial transects was modelled jointly with a point process model for where transects were established, in order to correct for bias that arises from "preferential sampling", *i.e.* sampling in locations where you expect to find the target species.

3.2 Existing Method

To integrate presence-only and presence/absence datasets, we follow the formulation of Fithian et al. (2015), for the case of a single species. The datasets comprise the presence-only data that forms a point pattern, S_n, in addition to a binary vector, $\boldsymbol{y} = y_1, \ldots, y_{n_{\text{survey}}}$, that represents whether a species was recorded as present or absent at each site. Strictly speaking these sites are regions rather than points, and for simplicity we will assume they all have constant area c and will identify them via the point location of their centre, denoted $\boldsymbol{s}_i^{\text{PA}} \in \mathcal{D}$ for $i = 1, \ldots, n_{\text{survey}}$. Let $A(\boldsymbol{s})$ be the observed species abundance of a site of size c centered at $\boldsymbol{s}$, indexed continuously over $\mathcal{D}$. Then $A\left(\boldsymbol{s}_i^{\text{PA}}\right)$ describes the number of species observed at each survey site, and $\boldsymbol{y}$ arises from a collection of Bernoulli random variables:

$$Y\left(\boldsymbol{s}_i^{\text{PA}}\right) = \begin{cases} 1 & A\left(\boldsymbol{s}_i^{\text{PA}}\right) \geq 1 \\ 0 & A\left(\boldsymbol{s}_i^{\text{PA}}\right) = 0. \end{cases}$$

Note that in the current work we focus on presence/absence data $Y\left(\boldsymbol{s}_i^{\text{PA}}\right)$, and assume the site abundance $A\left(\boldsymbol{s}_i^{\text{PA}}\right)$ is unobserved. We will also ignore the issue of imperfect detection, an issue which would only qualitatively affect our models if detection rates are different in different environments — see Guillera-Arroita (2017).

Presence/absence data are commonly modelled using binary regression (McCullagh and Nelder, 2019). Fithian et al. (2015) assume presences follow an inhomogeneous Poisson process with intensity proportional to $\exp\{\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}\}$, hence abundances $A(\boldsymbol{s})$ come from a Poisson distribution with mean $\mu_A(\boldsymbol{s}) = c\exp\{\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}\}$, where as previously $\boldsymbol{X}$ is a set of environmental variables that can be measured throughout

$\mathcal{D}$. Without loss of generality, we set $c = 1$. We can then derive a relation between probability of presence, $\mu_Y(\boldsymbol{s})$, and the linear predictor $\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}$:

$$\begin{aligned}\mu_Y(\boldsymbol{s}) = \Pr(Y(\boldsymbol{s}) = 1) &= 1 - \Pr(A(\boldsymbol{s}) = 0) \\ &= 1 - \exp\{-\mu_A(\boldsymbol{s})\} \\ &= 1 - \exp\{-\exp\{\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}\}\} \end{aligned} \tag{3.1}$$

This well-known classical result (Fisher, 1922) allows presence/absence data to be modelled using a generalised linear model with a complementary log-log link function. Most importantly, this permits the $\boldsymbol{\beta}$ to be incorporated into the model in a format that can be also be interpreted in the context of the presence-only data. Fithian et al. (2015) link the presence/absence and presence-only datasets by assuming that the intensity $\lambda(\boldsymbol{s})$ of the presence-only data is a thinned form of the mean abundance rate, where the thinning process $T(\boldsymbol{s})$ is assumed to be a linear combination of measured bias covariates $\boldsymbol{B}(\boldsymbol{s})$:

$$\begin{aligned}\lambda(\boldsymbol{s}) &= \exp\{T(\boldsymbol{s})\}\mu_A(\boldsymbol{s}) \\ &= \exp\{\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{B}(\boldsymbol{s})\boldsymbol{\tau}\}. \end{aligned} \tag{3.2}$$

The bias covariates $\boldsymbol{B}(\boldsymbol{s})$ aim to describe site availability and visitation rates, *e.g.* distance from a sealed road, as proposed in Warton et al. (2013) and Fithian and Hastie (2013).

Fithian et al. (2015) further assume that presence/absence data $Y(\boldsymbol{s}_i^{PA})$ are independent of presence-only data S_n, conditional on the environmental predictors $\boldsymbol{X}$. So in summary the model is:

$$\begin{aligned} Y(\boldsymbol{s}_i^{PA}) &\sim \text{Bernoulli}\left(\mu_Y(\boldsymbol{s}_i^{PA})\right) \\ S_n &\sim \text{IPP}\left(\lambda(\boldsymbol{s})\right) \\ Y(\boldsymbol{s}_i^{PA}) \mid \boldsymbol{X}(\boldsymbol{s}_i^{PA}) &\perp\!\!\!\perp S_n \mid \boldsymbol{X}(s) \end{aligned}$$

where $\mu_Y(\boldsymbol{s}_i^{PA})$ and $\lambda(\boldsymbol{s})$ are given by Equations (3.1) and (3.2), respectively.

The integrated data model is fitted by estimating parameters ($\boldsymbol{\beta}$ and $\boldsymbol{\tau}$) from a separable joint log-likelihood:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\tau}|\boldsymbol{y}, S_n, \boldsymbol{X}, \boldsymbol{B}) = \ln \pi_{\text{IPP}}(S_n|\boldsymbol{X}, \boldsymbol{B}) + \ln \pi_{\text{binomial}}(\boldsymbol{y}|\boldsymbol{X})$$

The benefit of integrating the data sources is that a researcher can reduce sampling uncertainty in estimation of $\boldsymbol{\beta}$ by drawing inference from a larger pool of information. Fithian et al. (2015) go further and propose an extension to simultaneously model multiple (similar) species jointly, assuming that the bias coefficients are shared between species to likewise assist estimation. The validity of bias coefficients being constant across species depends on both how similar the species are, and on the exact nature of how the presence-only data was collected. For example, the accessibility of a location (a potential bias covariate) may have a greater influence on whether a shrub is observed than that for a larger, more easily observed tree. In this chapter we will consider the single-species case only, and leave a multiple-species case for future work.

Implicit in the Fithian et al. (2015) formulation are two independence assumptions: independence within datasets, meaning all spatial clustering in presence-only and spatial autocorrelation in presence/absence data has been accounted for by $\boldsymbol{X}$ and $\boldsymbol{B}$; and independence across datasets, meaning that all coincident spatial patterns between presence-only and presence/absence data are captured by $\boldsymbol{X}$. The within-dataset independence assumption was discussed in the previous chapter, and as discussed there, it is prone to the criticism of being unrealistic in an ecological setting (Pacifici et al., 2017). But also note that if a key, shared environmental driver of the species is missing from the model, its effects will still be present in the form of unaccounted-for dependence between the datasets.

3.3 Proposed Extension

In the previous chapter we relaxed the conditional independence assumption by adding a latent Gaussian random field to the linear predictor. Here we assume

a latent Gaussian random field that is *shared* between data sources, to relax the conditional independence assumptions made within *and* between data sources.

To ease computational burden we again approximate the latent field as a linear combination of k spatial basis functions $\boldsymbol{Z}(\boldsymbol{s})$, with corresponding random coefficients $\boldsymbol{u} \sim \mathcal{N}_k(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ ("fixed rank kriging", Cressie and Johannesson, 2008, see Section 2.3.2). This avoids the high dimensional realisation of the latent field in both datasets. Hence we can make the intensity and presence probability stochastic by adding the approximate Gaussian field, $\boldsymbol{Z}\boldsymbol{u}$, *i.e.*

$$\begin{aligned} \ln \lambda(\boldsymbol{s}) &= \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u} + \boldsymbol{B}(\boldsymbol{s})\boldsymbol{\tau} \\ \ln(-\ln[1-\mu_Y(\boldsymbol{s})]) &= \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u} \end{aligned} \tag{3.3}$$

We can then assume that the data are independent conditional on both $\boldsymbol{X}$ and $\boldsymbol{u}$ and fit the model by maximising the marginalised joint likelihood

$$\ell\left(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2\right) = \int \ln \pi(\boldsymbol{y} \mid \boldsymbol{u}) + \ln \pi(S_n \mid \boldsymbol{u}) + \ln \pi(\boldsymbol{u})\, d\boldsymbol{u}$$

where $\pi(\boldsymbol{u})$ is the product of the k Gaussian probability densities, since we are again assuming no correlation between the $\boldsymbol{u}$, and conditional on $\mathbf{u}$, $\ln \pi(\boldsymbol{y} \mid \boldsymbol{u})$ is the Bernoulli log-likelihood for $\mathbf{y}$ and $\ln \pi(S_n \mid \boldsymbol{u})$ is the Poisson process log-likelihood for S_n:

$$\begin{aligned} \ln \pi(\boldsymbol{y} \mid \boldsymbol{u}) = &\sum_{i=1}^{n_{\text{survey}}} y_i \ln\left(1 - \exp\left\{-\exp\left(\boldsymbol{X}\left(\boldsymbol{s}_i^{\text{PA}}\right)\boldsymbol{\beta} + \boldsymbol{Z}\left(\boldsymbol{s}_i^{\text{PA}}\right)\boldsymbol{u}\right)\right\}\right) \\ &- (1 - y_i)\exp\left\{\boldsymbol{X}\left(\boldsymbol{s}_i^{\text{PA}}\right)\boldsymbol{\beta} + \boldsymbol{Z}\left(\boldsymbol{s}_i^{\text{PA}}\right)\boldsymbol{u}\right\} \\ \ln \pi(S_n \mid \boldsymbol{u}) = &|\mathcal{D}| - \int_{\mathcal{D}} \exp\left\{\boldsymbol{X}(\boldsymbol{t})\boldsymbol{\beta} + \boldsymbol{B}(\boldsymbol{t})\boldsymbol{\tau} + \boldsymbol{Z}(\boldsymbol{t})\boldsymbol{u}\right\} d\boldsymbol{t} \\ &+ \sum_{i=1}^{n} \boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta} + \boldsymbol{B}(\boldsymbol{s}_i)\boldsymbol{\tau} + \boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{u} \end{aligned}$$

As in the previous chapter we have to approximate both a spatial integral and an integral over the random coefficients $\boldsymbol{u}$. The former we achieve with numerical quadrature as in the previous chapter (Equation 2.3), the latter we will approach using a Laplace approximation, facilitated by automatic differentiation software within

`TMB` (Kristensen et al., 2016). In Chapter 2 we also explored the use of a variational approximation but this is not a trivial extension for the presence/absence likelihood component — see the discussion in Section 3.6 for further detail.

3.4 Simulation Study

We turn to simulation to look at if, and in what scenarios, we are able to improve models by jointly modelling the presence-only and presence/absence datasets.

In a similar set-up to Section 2.4, datasets were simulated over a square domain with sides of length 100, $\mathcal{D} = \{[1, 100], [1, 100]\}$. The presence-only data were again simulated from a Poisson point process, conditional on an observed Gaussian random field ξ. The presence/absence data comprised Bernoulli random variables from 1000 randomly sampled locations (survey sites) throughout $\mathcal{D}$. The probability of each site recording a presence was calculated from the mean abundance rate as in Equation (3.1). The mean abundance rate, μ_A was a log-linear function of three randomly generated spatial covariates, two observed, X_1 and X_2, and one unobserved, ξ:

$$\ln \mu_A = \beta_0 + \beta_1 X_1(\boldsymbol{s}) + \beta_2 X_2(\boldsymbol{s}) + \xi(\boldsymbol{s}).$$

This was thinned by two randomly generated, spatial, biasing covariates, B_1 and B_2 to create the intensity function for the presence-only data:

$$\ln \lambda(\boldsymbol{s}) = \beta_0 + \beta_1 X_1(\boldsymbol{s}) + \beta_2 X_2(\boldsymbol{s}) + \tau_0 + \tau_1 B_1(\boldsymbol{s}) + \tau_2 B_2(\boldsymbol{s}) + \xi(\boldsymbol{s})$$

As in Chapter 2, presence-only data were simulated using the `spatstat` package in `R` (Turner and Baddeley, 2005). For the presence/absence data, standard `R` software was used to simulate Bernoulli variables at spatial locations sampled uniformly over $\mathcal{D}$. The spatial covariates were zero-mean GRFs with isotropic Gaussian covariance function with a variety of correlation ranges.

We saw in Section 2.4 that the relative spatial scales of predictors and the latent field played an important role in our ability to model presence-only data accurately

as a LGCP — we found that it is typically easier to estimate model parameters when these spatial scales are more distinct. Although spatial confounding is of practical concern, it is outside our current scope and so we chose correlation ranges of the covariates to reflect different spatial scales. We generated X_1, X_2, B_1, B_2 and ξ using the `RandomFields` package in `R` (Schlather et al., 2015), with correlation ranges set to 20, 10, 20, 10 and 30 respectively. The variance of each field was fixed at one so that all model components were effectively standardised. An example of each component of a single simulation is given in Figure 3.1. The environmental and bias parameters (including intercepts) were fixed at $\boldsymbol{\beta} = (1.75, -1.2, 0.75)$ and $\boldsymbol{\tau} = (-2, 1.3, -0.8)$, reflecting a range of positive and negative effects on species' prevalence/inhibition across the domain in each dataset. These values also yielded $\mathbb{E}\left[N_{\text{presence-only}}\right] = 2000$, *i.e.* the presence-only data were expected to be twice as numerous as the simulated survey data.

As in the previous chapter, we studied the ability of the latent field to act as a surrogate for missing predictors. We emphasise that it is not realistic to assume that species distribution depends only on predictors included by a researcher within their model, and missing predictors should be considered the norm. For these reasons we looked at four simulation scenarios that reflect a variety of model specifications. These were:

1. All predictors were correctly included in the model (Correctly Specified)
2. An environmental predictor (X_1) was missing (Missing Env. Covariate)
3. A biasing predictor (B_1) was missing (Missing Bias Covariate)
4. One of each of an environmental and biasing predictor (X_2 and B_1) were missing (Missing Env. + Bias)

The presence-only and presence/absence datasets were fitted jointly or separately, using a variety of models for comparison. These included:

POPA Our proposed integrated data model sharing a latent field across datasets

PA A spatial random effect GLM with binary response, *i.e.* we dropped the presence-only data from the model proposed in Section 3.3, but kept the latent field,

and trained it on the presence/absence data only

PO A LGCP regression model, trained only on the presence-only data, *i.e.* we dropped the presence/absence data but kept the latent field

IPP POPA The integrated data model of Fithian et al. (2015), *i.e.* the latent field was dropped altogether, but the model was trained on both datasets

IPP PA A binary GLM with complementary log-log link function, trained only on the presence/absence data, *i.e.* we dropped both the latent field and the presence-only data

IPP PO An IPP regression model, trained only on the presence-only data, *i.e.* we dropped both the latent field and the presence/absence data

We anticipate that models including a latent field (POPA, PA and PO) will perform better than the corresponding procedures without a latent field due to greater flexibly to account for the various simulation scenarios. All procedures involving a latent field were approximating it via a regular grid of 100 bi-square basis functions (see Equation (2.8) for details).

We assessed the performance of a model by its predictive ability, as well as its ability to recover true covariate coefficients. For predictive ability, we performed a four-fold, hold-one-out, cross validation using spatial blocks constructed similarly to those of Section 2.5. As part of the data generation process we had the true underlying rates of interest: $\mu_A(\mathcal{D}) \approx \{\mu_A(\boldsymbol{s}_j)\}_{j=1}^{10000}$ and $\lambda(\mathcal{D}) \approx \{\lambda(\boldsymbol{s}_j)\}_{j=1}^{10000}$. For each fitted model we predicted $\{\hat{\mu}_A(\boldsymbol{s}_j)\}_{j=1}^{10000}$ and $\left\{\hat{\lambda}(\boldsymbol{s}_j)\right\}_{j=1}^{10000}$ then compared these to the truth as measured by Kullback-Leibler divergence (KL Div.) between the fields (as in Equation 2.11). For coefficient recovery we computed root mean square error (RMSE) in parameter estimates ($\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\tau}}$).

3.4.1 Simulation Results

Throughout this section the main focus is on comparing the performance of our proposed model (POPA; Section 3.3) against models individually fitted to the datasets that include a latent field (PA and PO). Comparisons of our method with the model

of Fithian et al. (2015) (IPP POPA; Section 3.2) are also highlighted as this shows the advances our novel extension provides. Results for all models can be found in Figures 3.2 and 3.3.

Table 3.1: Average predictive accuracy (Kullback-Leibler divergence from the true field, $\mu_A(\mathcal{D})$) for the integrated data model (POPA) versus the presence/absence data model (PA). We see that only in scenarios in which we correctly specified the bias covariates (first two rows) did we find any utility in using an integrated data model. Otherwise, combining data could be very detrimental to predictive performance.

Sim. Scenario	Avg. KL Div. $\mu_A \| \hat{\mu}_A$	
	POPA	PA
Correctly Specified	417.27	1015.43
Missing Env. Covariate	648.83	2171.44
Missing Bias Covariate	4230.08	1014.59
Missing Env. + Bias	4941.73	2261.93

In all simulations scenarios where the bias terms in the presence-only model were correctly specified, our proposed integrated data model (POPA) tended to perform substantially better at predicting presence/absence data than if it were modelled on its own (Table 3.1, Figure 3.2 top). We found the field estimated by presence/absence data only (PA) to be over twice as far from the truth (in KL Div.) as that estimated by POPA, on average, when we were able to correctly specify the fixed effects involved in both models, and further when an environmental covariate was missing. Conversely, in the two scenarios in which we failed to specify one of the bias covariates, we found predictive performance to be substantially worse for POPA than if the presence/absence data were analysed on their own.

A different trend was seen when evaluating how well POPA predicted presence-only intensity, compared to analysing presence-only data on its own (Table 3.2, Figure 3.2 bottom). Differences in Kullback-Leibler divergence were less than approximately 15%, but in all cases the integrated data model performed better that analysing presence-only data alone.

Figure 3.2 shows the predictive accuracy for each simulation scenario, for each of the six models compared. We found that the top performing models for predicting the mean abundance rate ($\hat{\mu}_A(\boldsymbol{s})$; top row) were POPA and PA, and when predicting intensity ($\hat{\lambda}(\boldsymbol{s})$; bottom row), the top performing models were POPA and PO. The

Table 3.2: Average predictive accuracy (Kullback-Leibler divergence from the true presence-only intensity field, $\lambda(\mathcal{D})$) for the integrated data model (POPA) versus the presence-only data model (PO). We see only marginal improvement, on average, for the integrated data model in all scenarios.

Sim. Scenario	Avg. KL Div. $\lambda \| \hat{\lambda}$	
	POPA	PO
Correctly Specified	185.98	218.78
Missing Env. Covariate	194.77	229.75
Missing Bias Covariate	283.79	317.83
Missing Env. + Bias	443.84	475.41

equivalent models that did not include a latent GRF (*i.e.* IPP POPA, IPP PA and IPP PO) were all outperformed by their LGCP-based counterparts in estimating the appropriate rate. PA tended to achieve lower KL Div. from the truth than IPP PA, as seen in the top row of panels (estimating $\hat{\mu}_A$). Likewise, PO tended to achieve lower KL Div. than IPP PO, as seen in the bottom row of panels (estimating $\hat{\lambda}$). Our proposed method achieved a smaller KL Div. than the single species version proposed by Fithian et al. (2015) (here labelled IPP POPA) in all scenarios except those in which we misspecified the bias model and were predicting μ_A (top row of panels, first and second from the right).

Similar trends were found when looking at the RMSE in estimating coefficients in the linear predictor (Figure 3.3). Specifically, POPA achieved the lowest RMSE for all parameters when the model was correctly specified (Figure 3.3, top row) or missing just an environmental covariate (Figure 3.3, second row), but separately modelling datasets (PA or PO) performed better when a bias covariate was missing (Figure 3.3, bottom two rows). The one exception to this rule was estimation of β_2 when missing a bias covariate, in which case POPA slightly outperformed PO.

The reason for the difference in performance across simulation scenarios is that missing bias predictors will bias predictions from the integrated model (POPA). Use of a shared latent field in the integrated data model gives robustness to missing *environmental* predictors, and integration of the two data types provides more information that can be used to estimate and account for a missing environmental predictor (as in Figure 3.2, second column). However, when a bias covariate is missing, the latent field tries to account for this but in so doing it seems to bias predictions in

the presence/absence dataset (Table 3.1). Presumably predictions of presence-only intensity ($\hat{\lambda}$) were also biased by this missing predictor, but seemingly less so for the integrated data model, which was informed by a second data source free of this bias.

3.5 Application: Flora in the Greater Blue Mountains

To illustrate our proposed integrated data model on some real data examples, we examined the spatial distributions of the four plant species outlined in Section 1.1.2. Having both detailed survey data and supplementary presence-only records allowed us to test our approach against models fitted to the two datasets separately.

3.5.1 Methods

We assessed whether the integrated data model improved upon the out-of-sample prediction of models fitted to the datasets individually. As in the previous section we measured this with a four fold, hold-one-out, cross validation (CV), set up in spatial blocks, such that $\mathcal{D} = \cup_{h=1}^{n_{\text{folds}}} \mathcal{D}_h$ and $\mathcal{D}_h \cap \mathcal{D}_{h'} = \emptyset$ for any $h \neq h'$ — see Figure 3.4 (left plot). We used this to calculate out-of-sample, predicted log-likelihoods on the presence-only and presence/absence data (both are conditional on the random effects). For the presence-only data component this is similar to that of Section 2.5, given by

$$\sum_{h=1}^{n_{\text{folds}}} \ln \pi \left(S^{(h)} \mid \hat{\boldsymbol{\beta}}^{\setminus h}, \hat{\boldsymbol{\tau}}^{\setminus h} \hat{\boldsymbol{u}}^{\setminus h} \right) = \sum_{h=1}^{n_{\text{folds}}} \left[\sum_{i=1}^{n_h} \boldsymbol{X}\left(\boldsymbol{s}_i^{(h)}\right) \hat{\boldsymbol{\beta}}^{\setminus h} + \boldsymbol{B}\left(\boldsymbol{s}_i^{(h)}\right) \hat{\boldsymbol{\tau}}^{\setminus h} + \boldsymbol{Z}\left(\boldsymbol{s}_i^{(h)}\right) \hat{\boldsymbol{u}}^{\setminus h} - \sum_{i=n_h+1}^{n_h+q_h} w_i^{(h)} e^{\boldsymbol{X}\left(\boldsymbol{s}_i^{(h)}\right)\hat{\boldsymbol{\beta}}^{\setminus h} + \boldsymbol{B}\left(\boldsymbol{s}_i^{(h)}\right)\hat{\boldsymbol{\tau}}^{\setminus h} + \boldsymbol{Z}\left(\boldsymbol{s}_i^{(h)}\right)\hat{\boldsymbol{u}}^{\setminus h}} \right] \quad (3.4)$$

where each of the $h = 1, \ldots, n_{\text{folds}}$ was held out in the training of $\hat{\boldsymbol{\beta}}^{\setminus h}$, $\hat{\boldsymbol{\tau}}^{\setminus h}$ and $\hat{\boldsymbol{u}}^{\setminus h}$. $S^{(h)} = \left\{ \boldsymbol{s}_i^{(h)} \right\}_{i=1}^{n_h+q_h}$ denotes the point pattern (and quadrature) within fold h and $w_i^{(h)}$ denotes the size/weight of the i^{th} quadrat within fold h (here equal to zero for $i = 1, \ldots, n_h$). The predicted conditional log-likelihood for the presence/absence

data component is given by

$$\sum_{h=1}^{n_{\text{folds}}} \ln \pi\left(\boldsymbol{y}^{(h)} | \hat{\boldsymbol{\beta}}^{\setminus h}, \hat{\boldsymbol{u}}^{\setminus h}\right) = \sum_{h=1}^{n_{\text{folds}}} \left[\sum_{i=1}^{n_{h;\text{survey}}} y_i^{(h)} \ln\left(1 - e^{-\exp\left\{\boldsymbol{X}\left(\boldsymbol{s}_i^{h;\text{PA}}\right)\hat{\boldsymbol{\beta}}^{\setminus h} + \boldsymbol{Z}\left(\boldsymbol{s}_i^{h;\text{PA}}\right)\hat{\boldsymbol{u}}^{\setminus h}\right\}}\right) - \left(1 - y_i^{(h)}\right) \exp\left\{\boldsymbol{X}\left(\boldsymbol{s}_i^{h;\text{PA}}\right)\hat{\boldsymbol{\beta}}^{\setminus h} + \boldsymbol{Z}\left(\boldsymbol{s}_i^{h;\text{PA}}\right)\hat{\boldsymbol{u}}^{\setminus h}\right\}\right] \tag{3.5}$$

where $\boldsymbol{y}^{(h)}$ denotes the binary response at survey sites $\left\{\boldsymbol{s}_i^{h;\text{PA}}\right\}$ for $i = 1, \ldots, n_{h;\text{survey}}$ within fold h. Again, each of the $h = 1, \ldots, n_{\text{folds}}$ was held out in the training of $\hat{\boldsymbol{\beta}}^{\setminus h}$ and $\hat{\boldsymbol{u}}^{\setminus h}$.

Each predictor variable was measured throughout the Greater Blue Mountain World Heritage Area (GBMWHA), and in a 100km buffer around this area, at a spatial resolution of 1km^2. This means that presence locations closer than this will be largely indistinguishable with respect to the predictors (though not so in the basis functions). Figure 3.5 displays the number of presence locations in either dataset within 1km of each other — as there are very few points within this distance, we assumed the 1km^2 resolution will be sufficiently fine for modelling all species. This also shows us that the two datasets are distinct, *i.e.* all distances between presence locations are non-zero.

We modelled the mean abundance rate, μ_A as a log-linear function of average annual minimum (MNT) and maximum (MXT) temperatures, and for all species (excluding *H. cernuus* which lacks the sufficient number of presences) we also included an interaction and quadratic terms for these. This was done because plant species will typically have an optimal range for temperature rather than responding to it linearly. We likewise modelled the presence-only intensity, additionally thinned by potential biasing predictors: distances to main road (D.Main) and urban areas (D.Urb) as in Equation (3.2). Due to the rarity and restricted distribution of *H. cernuus* presences, we performed a two-fold CV for this species — as in Figure 3.4 (right plot).

We performed the CV multiple times on increasingly dense, regular grids of locally compact bi-square functions in a similar way to that of Section 2.5. This allowed us to compare the integrated data model against the individual data models across

a range of basis functions — we were able to look at which approach achieves the highest predicted log-likelihood for any of the basis configurations trialled, including a model with no approximate latent field ($k = 0$, an IPP-based model). This approach would not be computationally feasible with other currently available software and was enabled by the fast model-fitting procedures developed in this thesis. We calculated the predicted log-likelihoods up to the largest number of basis functions that could be arranged on a regular grid and would not exceed the observed number of presences in either dataset, $\min\{n, n_{\text{survey}}\}$. This permitted valid comparison of the integrated and separate models without overfitting.

3.5.2 Results

The results for predictive performance can be found in Figure 3.6 for each of the four species, full details can be found in Appendix B.1. We found that the integrated model only improved predictive performance for one of the four species, *C. eximia* (Figure 3.6 top left). That is, there was an improvement for this species when comparing the joint predicted likelihood (Equations 3.4 and 3.5 summed) for the integrated data model (purple) and separate models (blue and red). The maximum predicted joint likelihood was -2065.4 when using 204 basis functions for the integrated data model, and -2201.5 when using just two broad basis functions when modelling the data separately. However, we saw large variability in the likelihood for the integrated data model, for basis configurations close to the number of presence-only data for this species (242) — Figure 3.6 top left panel, purple line beyond 150 basis functions. Hence there were a number of basis function configurations where the integrated data model did not have higher predictive likelihood. We can see similar instability in the integrated data model when the number of basis functions approaches $\min\{n, n_{\text{survey}}\}$ for *E. sparsifolia* and *H. cernuus* (Figure 3.6 bottom two panels). The remaining results were in line with those in Section 3.4 for scenarios where we have misspecified the bias predictors for the presence-only component of the model. We also found that for all species except *H. cernuus*, including a latent random field in the model improved upon assuming independence of data sources as in Fithian et al. (2015). This is seen in Figure 3.6, where the predicted likelihoods at

$k = 0$ were exceeded by at least one of the subsequent data points plotted ($k > 0$), except in the case of *H. cernuus* (bottom left). This too is consistent with results from Section 3.4.

3.6 Discussion

In this chapter, we extended the methodology of Fithian et al. (2015) to integrate presence-only and presence/absence data by jointly modelling the species response to environmental effects, in the case of a single species. Our contribution is to introduce a shared latent field that can account for additional spatial correlation in presences observed in either data source. Such additional correlation can arise from missing predictors that are important to the distribution of the target species — when omitted from the model, they induce dependence across datasets that is otherwise unaccounted for. Our simulation results in Section 3.4 showed that in certain scenarios, combining the data in this way can improve both predictive accuracy and coefficient estimation. However, when we applied the methodology to multiple species of flora in the GBMWHA in Section 3.5, we found little benefit in combining the datasets. We suspect that this is due to misspecification of the biasing predictors, a problem identified earlier in our simulation results.

There are a few possible explanations for the poor performance on the real data examples that cannot be ruled out, and serve to highlight the challenges involved with integrated data modelling. First, it is possible that in our example the presence/absence data are a sufficiently high quality dataset that it cannot be improved upon by presence-only data, which come with biases. There were 8223 survey sites, recording more presences than were found in the entire presence-only dataset (except for the rare species, *H. cernuus*). We found the only example of improvement in predictive performance was with the species with the largest number of presences (Figure 3.6). Perhaps a similarly large number of presence records is needed for a presence-only dataset to be informative in this sort of situation, we would need to repeat analyses on many more species however before making such a generalisation. Second, our models were far from perfect, and we cannot rule out missing predictors.

Environmental variables known to usually be important include soil type and rainfall (Hager and Benson, 2010). There is also the potential for measurement error in predictors (Stoklosa et al., 2015) and in identifying the spatial location of presence events (Hefley et al., 2014). We elected to fit a simple model for illustratory purposes, particularly important for our rarer species where there was little capacity to fit a more complex model anyway. But we notice that in simulations our model was robust to missing environmental predictors (Figure 3.2 two left-most columns and Figure 3.3 top two rows) so we are not sure that inclusion of additional predictors would have had much of an effect on results.

Our main finding, that misspecified bias in the presence-only data component can render integrated models somewhat useless, was also an outcome from the simulation study of Simmonds et al. (2020), who examined similar integrated data (LGCP-based) models. Specifically, the researcher's primary conclusion is: "... [integrated data models] outperformed single dataset models in some cases, but if bias in [presence-only] data was ignored then [integrated data models] did not provide any benefits over modelling [presence/absence] data alone.". Both the work of this chapter and Simmonds et al. (2020) attempted to account for bias in the presence-only component by incorporating covariates designed to quantify this bias (*e.g.* variables associated with accessibility). If the efficacy of integrated data models relies on getting this bias model right then this is a major limitation, given that as previously, our working assumption should always be that there are missing predictors.

What is needed is a term to account for missing predictors in the *bias model* — meaning that we need to include a second latent random field in the model, that applies to the presence-only component solely. That is, as previously we could model presence/absence data as:

$$\ln\left(-\ln\left[1-\mu_Y\left(\boldsymbol{s}\right)\right]\right)=\beta_0+\beta_1 X_1\left(\boldsymbol{s}\right)+\beta_2 X_2\left(\boldsymbol{s}\right)+\xi\left(\boldsymbol{s}\right)$$

but we could now model intensity of presence-only data as:

$$\ln\lambda\left(\boldsymbol{s}\right)=\beta_0+\beta_1 X_1\left(\boldsymbol{s}\right)+\beta_2 X_2\left(\boldsymbol{s}\right)+\tau_0+\tau_1 B_1\left(\boldsymbol{s}\right)+\tau_2 B_2\left(\boldsymbol{s}\right)+\xi\left(\boldsymbol{s}\right)+\psi\left(\boldsymbol{s}\right)$$

where $\psi(\boldsymbol{s})$ is a second Gaussian random field that applies to the presence-only data only. A model similar to this has considered previously (Simmonds et al., 2020), although they assumed different realisations of ξ for the different datasets. This model could be expected to have robustness to missing bias predictors. A challenge however is reliably fitting this model — we have noticed some instability when fitting an integrated data model with one latent field (*e.g.* Figure 3.6, all panels, except top right) and would expect the potential for considerable instability when estimating a second random field as well. Simmonds et al. (2020) seemed to address these stability issues by applying their model to simulated data only, where data were simulated with a rather unrealistically large number of presence points (usually thousands per dataset), and we did similarly in simulations with a single latent field, using 2000 expected presences to minimise convergence issues. One way forward, to improve computational stability, would be to model multiple species simultaneously, assuming a common latent bias field $\psi(\boldsymbol{s})$ across species, to be discussed further in Chapter 5. Bias covariates can often be assumed to affect multiple species in the same way, so one could expect a stronger signal from missing bias covariates when it is estimated jointly across multiple species (Fithian et al., 2015).

The instability in our integrated data model can arise due to the use of a Laplace approximation, when using a large number of basis functions (Figure 3.6 particularly top left and bottom right). The issue seems to be that the Hessian matrix, needed in the Laplace approximation (Equation 2.6), can be near singular when there are many basis functions — seen also in Table 2.1 and Figure 2.7. This problem only arose when the number of basis functions was close to the number of presence events, so it could be interpreted as a sign of overfitting. However efforts to improve stability could involve developing a variational approximation for this model, which was shown in the LGCP case to be more stable in this sort of situation (Figure 2.7). Developing a variational approximation for the presence/absence component of the model would be non-trivial, but Hui et al. (2019) suggest a way forward for Bernoulli models, in the context of variational GAMs.

The flora datasets (Section 1.1.2) were quite large and difficult to fit, as we had over $8,000$ presence/absence data points and we used $\approx 86,000$ quadrature points

in our presence-only fit. This large number of quadrature points was used due to the spatial roughness of our covariates hence our intensity surface (as in Renner et al., 2015). Fitting our integrated data model to a dataset of this size would have been computationally prohibitive without the advances in Chapter 2, yet we were able to fit multiple models, in order to do a basis search along the lines of Figure 2.7 in just a few hours. Simmonds et al. (2020) fitted similar models to simulated data using INLA, but used at most 500 presence/absence points, and chose parameters for simulation carefully, to ensure computational feasibility. There are many possible extensions of our model, such as the multiple species extension, which are now feasible due to our rapid algorithm.

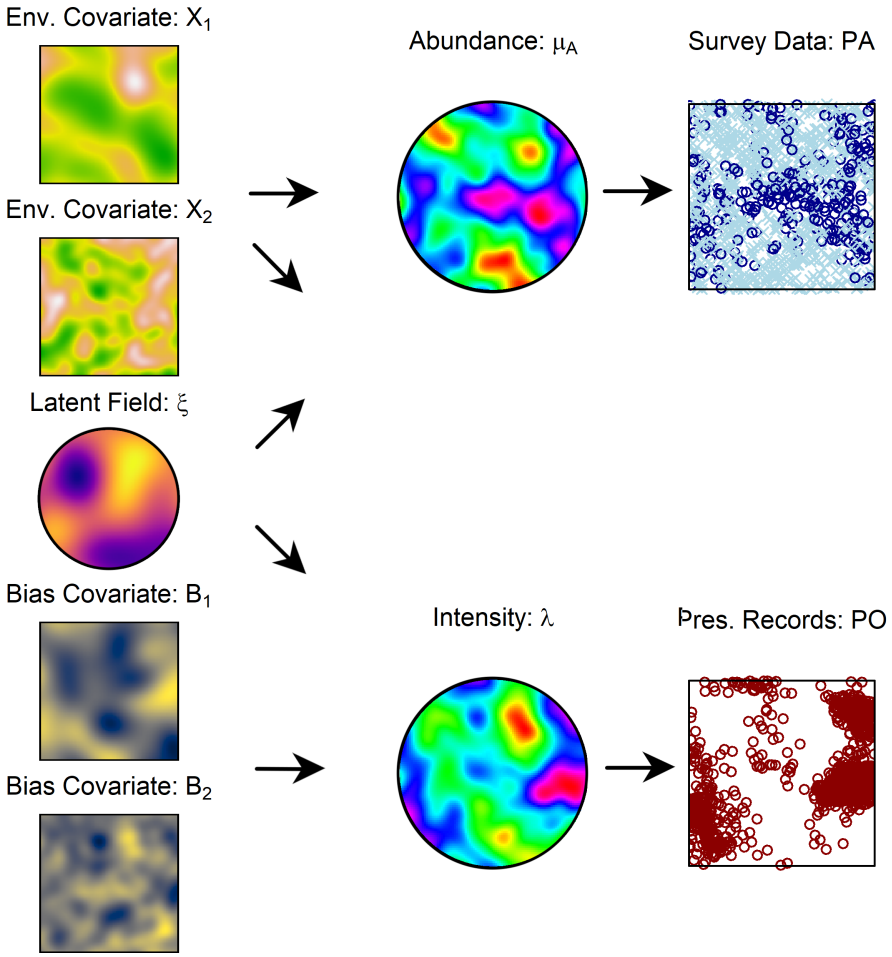

Figure 3.1: An example of the simulation process. Squares represent measured/observed variables or data, circles represent unmeasured/unobserved processes. We simulated a broad scale latent field, as well as two environmental covariates and two bias covariates — one of each with long and short correlation ranges

. We combined the latent field and environmental covariates to obtain the true species abundance rate (μ_A), but we additionally needed the bias covariates to obtain the presence-only intensity (λ). The presence/absence dataset was then simulated from μ and the presence-only dataset was simulated to have intensity λ.

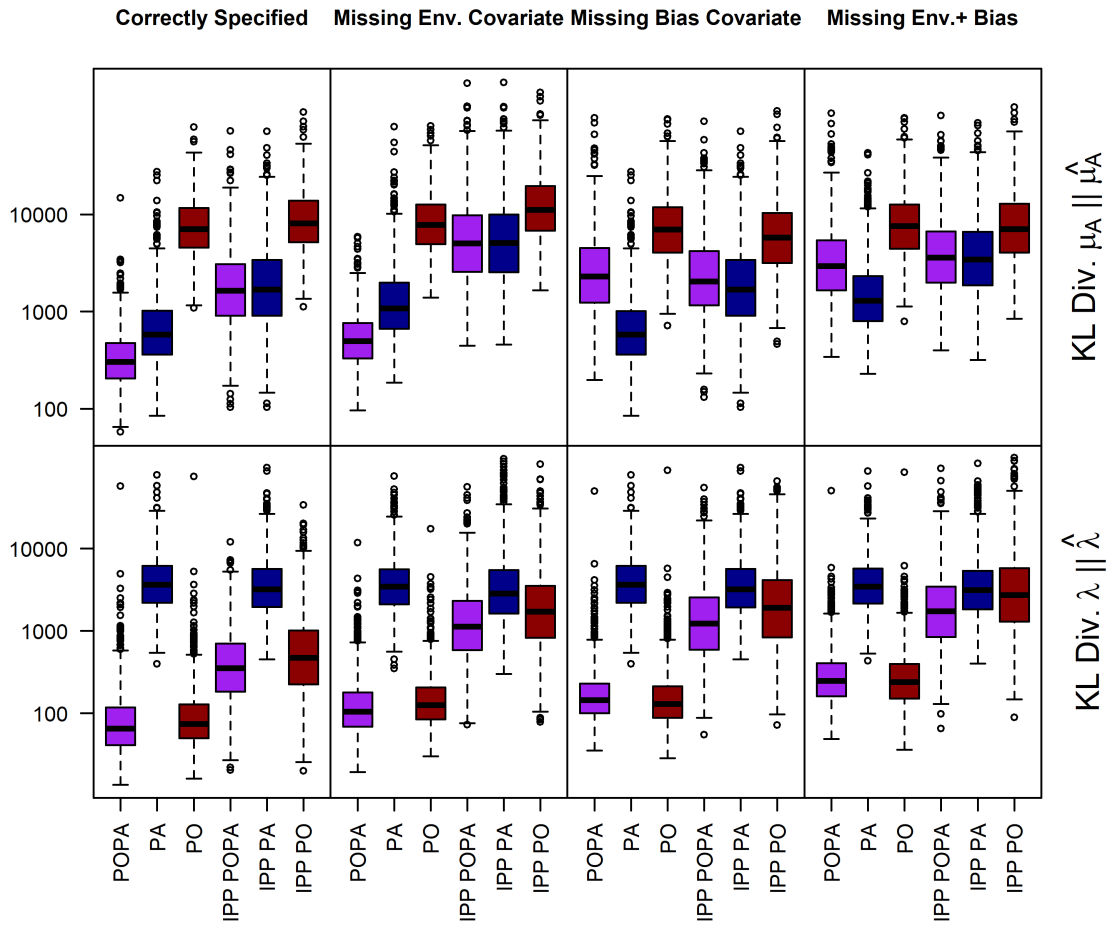

Figure 3.2: Accuracy of predictions measured by Kullback-Leibler divergence (KL Div.) from the true abundance rate (μ_A; top row) and true intensity rate (λ; bottom row) to that predicted by each model. Results of all 1000 simulations are shown as boxplots. Purple represents integrated data models, blue represents presence/absence data models and red represents presence-only data models. A model prefix of "IPP" indicates the model does not include a latent field. Only in the first two scenarios ("Correctly Specified" and "Missing Env. Covariate") did we find the integrated data model (POPA) more often achieved the lowest divergence.

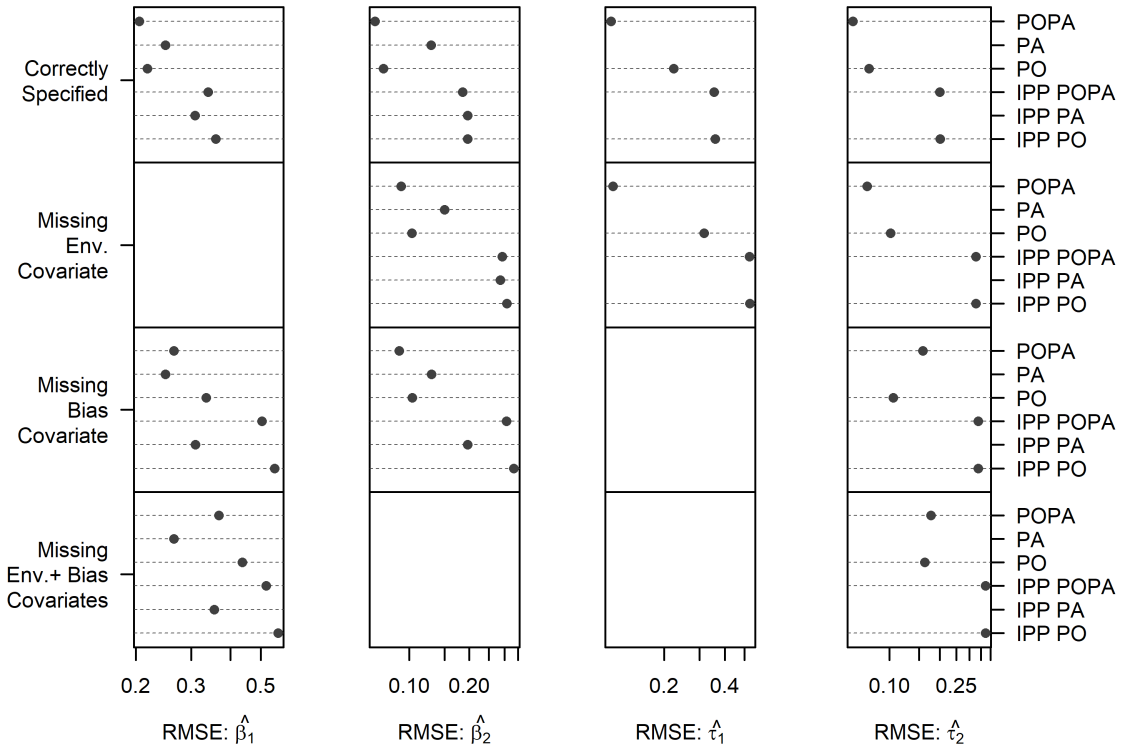

Figure 3.3: Accuracy in the ability of each model to recover the true parameter values as measured by root mean squared error (RMSE) across each scenario. Simulation scenarios are labelled on the left axis, models fitted are shown on the right axis. Each column represents one of four parameters being estimated. POPA indicates a model jointly fitted to the presence-only and presence/absence datasets. PA indicates a model fitted to the presence/absence data only. PO indicates a model fitted to the presence-only data solely. A model prefix of "IPP" indicates the model does not include a latent field. We see that our proposed model (POPA) achieved the smallest RMSE for each parameter in the first two rows (scenarios where biasing covariates are correctly specified). In the bottom two rows (scenarios where biasing covariates are misspecified), the PA model achieved the smallest RMSE for environmental parameters (two left-most columns) while the PO model achieved the smallest RMSE for bias parameters (two right-most columns). Panels missing information are scenarios in which the parameter corresponded to a covariate that is missing from the fitted model.

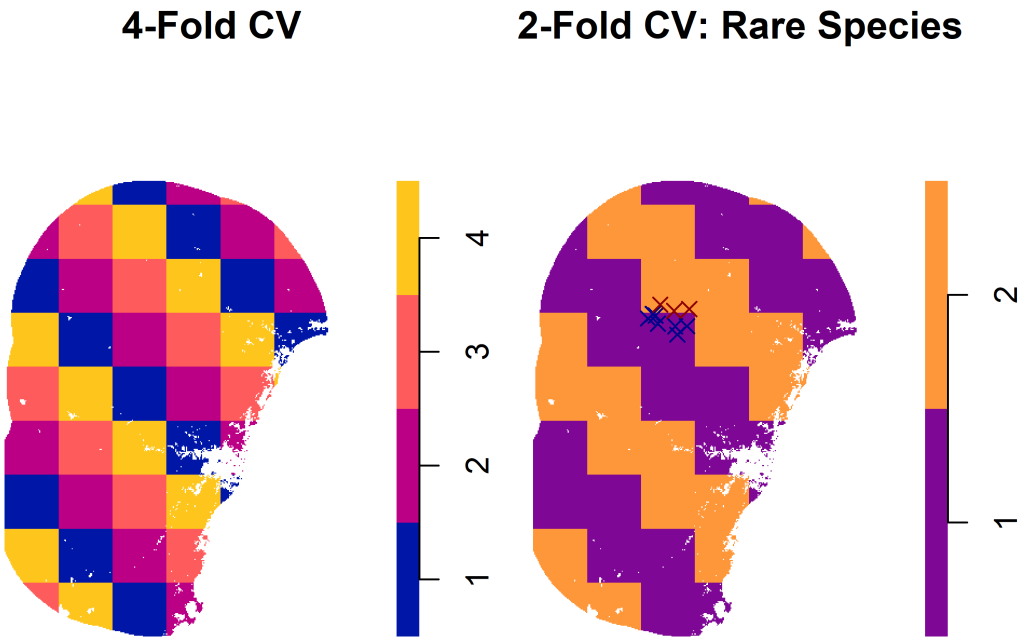

Figure 3.4: Graphical representation of the cross validation folds constructed through spatial blocks. Left panel: four-fold CV is used for *C. eximia*, *E. sparsifolia* and *E. canaliculata*. Right panel: two-fold CV is used for the rare species, *H. cernuus*.

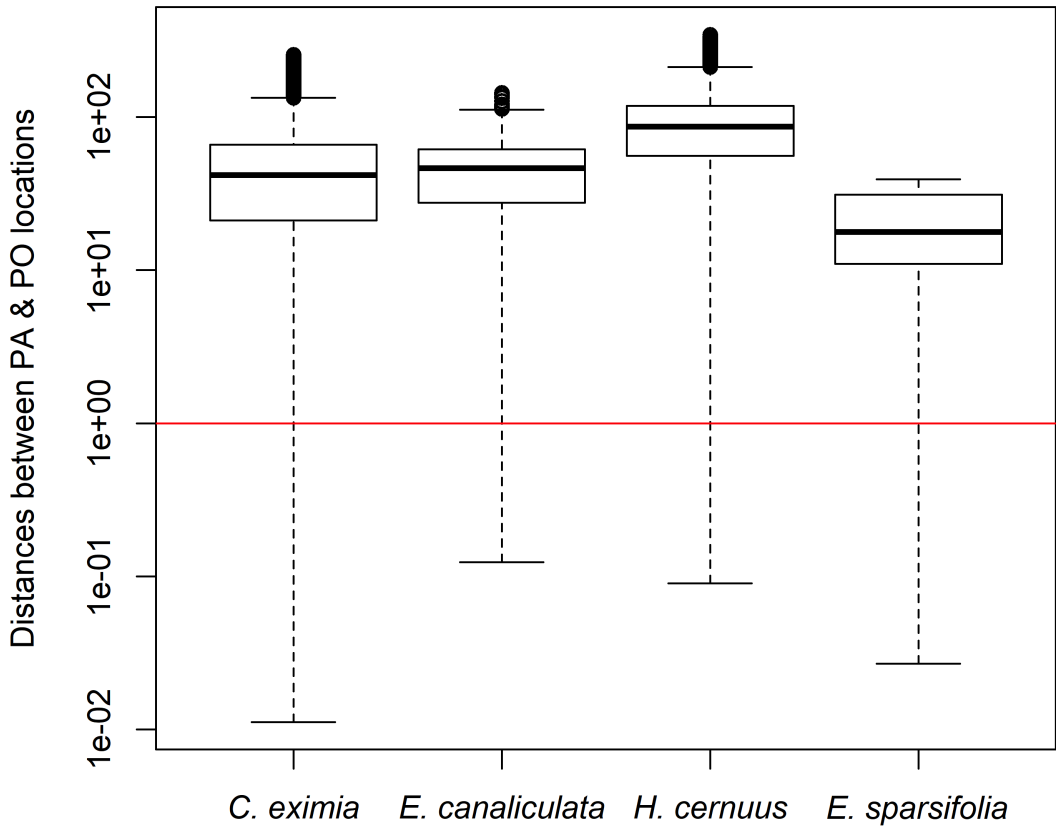

Figure 3.5: Box plots showing the range of the distances between presence locations in the presence/absence (PA) and presence-only (PO) data. Red line indicates a distance of 1km between point locations. Very few points were within 1km for each species.

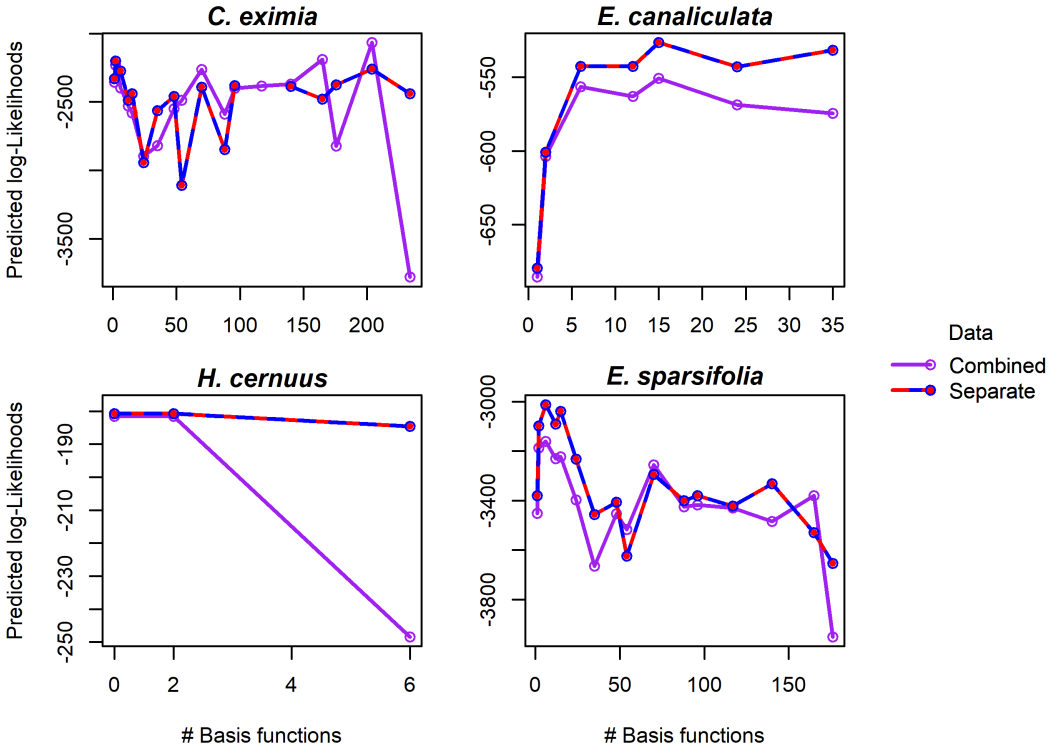

Figure 3.6: Out-of-sample, predicted, conditional combined log-likelihoods for both the presence-only and presence/absence data components (Equations 3.4 and 3.5 summed). Each panel shows the likelihoods over increasingly dense grids of basis functions, for each species. Red and blue points and lines represent results for when the data are modelled separately. Purple indicates the results of our proposed integrated data model. Integrating data sources improved predictive performance for only one of the four species (*C. eximia*).

Chapter 4

`scampr` R Package: Spatially Correlated, Approximate Modelling of Presences in R

4.1 Introduction

As discussed in Chapter 2 and elsewhere (Illian et al., 2012; Renner et al., 2015), it is important to account for missing covariates and potential bias in opportunistic collection of presence-only data, which can be achieved by fitting a log-Gaussian Cox process. Nevertheless, an inhomogeneous Poisson process (IPP) remains the more popular approach for modelling these data in ecology, even though is assumes no clustering due to missing predictors, an assumption that is often unrealistic. As previously mentioned, an IPP is most commonly fitted to data using MAXENT software (Phillips and Dudík, 2008), but also sometimes using the `spatstat` package (Turner and Baddeley, 2005) in `R` or user-written code (Renner et al., 2015).

R software packages are available to model presence-only data under a log-Gaussian Cox process (LGCP) framework but user uptake appears to be limited by long computation times; complicated user interfaces; or both. Taylor et al. (2013) developed one of the earliest tools, the `lgcp` package. This software uses MCMC samplers to estimate a LGCP model, which take a long time to fit to a single dataset — in the magnitude of hours. The `INLA` package (Rue et al., 2009) provides a general and faster framework for fitting latent Gaussian field models, including LGCP models. However, the generality of `INLA` means the interface was not designed specifically to fit point process models and it can be difficult to use. Bachl et al. (2019) attempts to make the INLA framework more accessible to those fitting a LGCP through the function `lgcp()` in the `inlabru` package. This provides a simpler user interface for INLA, but computation times are still prohibitive, especially when the analysis, such as cross-validation, requires many model fits.

In Chapter 2 we showed that presence-only data can be modelled with spatially correlated errors effectively and quickly using a combination of: closed form likelihood approximations; rank reduction; and automatic differentiation (AD). Using these advances we reduced computation time in some instances from hours to seconds, as compared to INLA, which is already known to be faster than other Bayesian competitors (Taylor and Diggle, 2014). In Chapter 3 we further showed the ability of this same framework to integrate presence-only data with survey data (pres-

ence/absence). This chapter describes `scampr`, an `R` package that implements these advances in modelling LGCP models and provides a simple interface for regression modelling of point event data. All of the common S3 functions familiar to users (such as `summary`, `plot`, `simulate`, `predict`,...) have also been written for `scampr` models. The package is built upon the advances of `TMB` (Kristensen et al., 2016) which enables us to code the likelihoods derived in Section 2.3.1 in `C++`, as well as providing automatic differentiation for easy access to gradient information — permitting fast optimisation, automated Laplace approximation, and automated estimation of the variance-covariance matrix of parameter estimates in `scampr` models. In this sense, the package is built upon similar foundations to the more widely-used `glmmTMB` package (Brooks et al., 2017).

The package name `scampr` stands for Spatially Correlated, Approximate Modelling of Presences in `R` however, the verb "scamper" — to run with quick, light steps — perfectly captures the motivation of this package: to give researchers access to complex spatial models that fit quickly and require only a light touch.

4.2 Fitting the LGCP Model

As in Chapter 1, the data comprises the n point events as $S_n = \{\boldsymbol{s}_i\}_{i=1}^n$, where $\boldsymbol{s}$ denotes the coordinates in some domain, $\mathcal{D}$. We assume these arise from a Poisson process with spatially varying intensity that is a log-linear function of predictors $\boldsymbol{X}$ and an unobserved Gaussian random field (GRF) $\xi(\boldsymbol{s})$:

$$\begin{aligned}\lambda(\boldsymbol{s}) &= \exp\{\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \xi(\boldsymbol{s})\} \\ &\approx \exp\{\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u}\}\end{aligned} \tag{4.1}$$

Because $\xi(\boldsymbol{s})$ is unobserved, this forms a LGCP. The latent field ξ is included to account for missing predictors that would otherwise be responsible for spatial clustering that is not captured by the Poisson assumption.

A LGCP model is fitted in `scampr` using the function `scampr()` — in fact the full suite of models offered by this package are fit using this function — using a

likelihood based approach to estimate parameters of the LGCP. The `scampr` model approximates the latent field using a linear combination of k basis functions ($\boldsymbol{Z}(\boldsymbol{s})$ above) and random normal coefficients ($\boldsymbol{u}$ above), hence the model has the form of a (spatial) random effects model (Cressie and Johannesson, 2008). If we were to drop the GRF, $\xi(\boldsymbol{s})$, from the model then we would have an inhomogeneous Poisson process (IPP). So in a way, a `scampr` is to an IPP what `lmer()` is to `lm()`.

We will use a simple example to illustrate fitting point process models within `scampr` — regressing the point pattern representing gorilla nesting locations (Figure 1.1a) against a single covariate, elevation (Figure 1.1b) that has been centered and scaled.

```
    # fit the LGCP model
lgcp_scampr <- scampr(pres ~ elev.std, data=gorillas.df)
```

In line with much of the regression modelling syntax in `R`, the model is described using formulae, written in compact symbolic form (see Chambers and Hastie, 1992) where the response (`pres` above) and fixed effects (`elev.std` above) are found within the data provided (`gorillas.df` above). Note that this syntax will automatically include an intercept term in the model. The `data` object should be a `data.frame` containing all terms in the formula, also spatial locations stored as `x` and `y`, and quadrat sizes stored as `quad.size`. For example:

```
head(gorillas.df)
```

	x	y	elevation	quad.size	pres
1	582.5184	676.8862	2008	0.0000000000	1
2	581.8230	677.4227	1699	0.0000000000	1
3	582.1310	676.9379	1872	0.0000000000	1
2275	582.2161	674.1754	1353	0.0007917668	0
2276	582.2437	674.1754	1353	0.0007917668	0
2277	582.2713	674.1754	1338	0.0007917668	0

While most of this should be familiar to `R` users, the model response and quadrat sizes (`quad.size` above) require some explanation. Unique to `scampr` is that the "response" variable must be a binary variable that identifies which rows are points

within the point pattern (`pres == 1` in this example) and those that are *quadrature points*, used to estimate a spatial integral over the study region using numerical quadrature. Specifically, we wish to maximise the likelihood:

$$\ell\left(S_n|\boldsymbol{\beta},\boldsymbol{\xi}\right)=\left\{\sum_{i=1}^{n}\boldsymbol{X}\left(\boldsymbol{s}_i\right)\boldsymbol{\beta}+\xi\left(\boldsymbol{s}_i\right)\right\}-\int_{\mathcal{D}}\lambda\left(t\right)\mathrm{d}t \tag{4.2}$$

where we approximate the spatial integral using an additional set of background or *quadrature* points $\{\boldsymbol{s}_i\}_{i=n+1}^{n+q}$, as:

$$\int_{D}\lambda\left(t\right)\mathrm{d}t\approx\sum_{i=n+1}^{n+q}w_i\exp\left\{\boldsymbol{X}\left(\boldsymbol{s}_i\right)\boldsymbol{\beta}+\xi\left(\boldsymbol{s}_i\right)\right\}. \tag{4.3}$$

All rows of data identified as the q quadrature points (*e.g.* by `pres == 0`) are used to approximate this spatial integral, with sizes w_j stored in `quad.size` (which additionally contains $w_i = 0$ at the presence locations by convention, but these are currently not used). Hence the `data.frame` provided to `scampr` point process models have $n+q$ rows. Quadrature points and their selection/setup are covered in more detail in Appendix C. The package `spatstat` contains the function `quad.scheme()` that can sample quadrature points and assign them sizes. Alternatively, the simplest way to create a quadrature scheme would be to randomly sample q points from the domain and set the sizes to $\frac{|\mathcal{D}|}{q}$ for quadrature points and to zero for presence points. This can be understood as a simple Monte Carlo approximation to the spatial integral.

In addition to `formula` and `data`, we can provide `scampr()` with additional arguments for further customisation. The names of the coordinate columns can be customised via the argument `coord.names`, as can the name of the quadrat size column, via the argument `quad.weights.name`. Default names are as given in `gorillas.df`. Other details, such as choice of the basis function configuration used to approximate the latent GRF, can be customised as seen later in Section 4.5.1. We can also fit an IPP model by setting the argument `model.type` to `"ipp"`.

Preliminary information about a `scampr` model fit can be obtained using `summary()` and `plot()`:

```
summary(lgcp_scampr)
```

```
Model Type:  LGCP with Variational Approx.
Formula:  pres ~ elev.std
AIC:  -4679.974    approx. marginal logLik:  2343.987
Basis functions per res.  12 99
Fixed Effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.63457    0.54456  1.1653   0.1219
elev.std     0.20246    0.23698  0.8543   0.1965
...
```

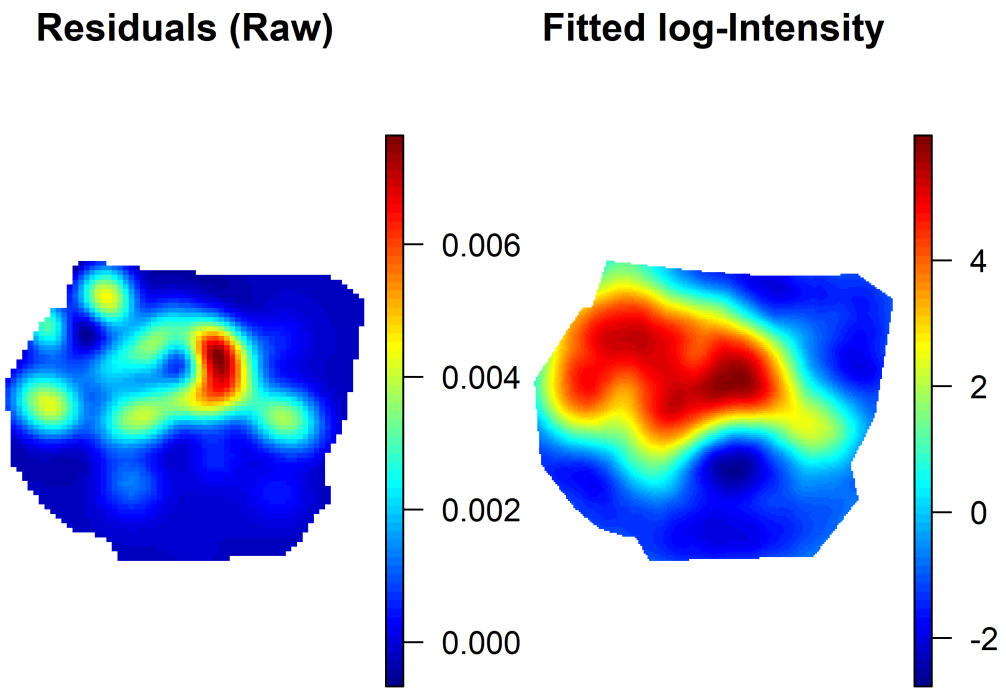

Figure 4.1: Plots available to `scampr` model fits are equivalent to those provided by `spatstat`. These display the residuals and fitted log-intensity across the quadrature of the domain for our example model fit: `scampr(pres~elev.std, gorillas.df)`. Plotting the model will produce both the residuals and fitted intensity. Individual plots may be selected using the argument `which`. Residuals are discussed in Section 4.4.

We find a small, positive fixed effect of elevation on gorilla nesting location — perhaps unsurprisingly, we might infer that gorillas tend to prefer to nest at higher elevations (though the effect is not significant, $p = 0.2$). Figure 4.1 by default presents a smoothed plot of residuals over the model quadrature points (as quadrats of $\approx 800m^2$

in the gorilla data example) along the lines of that produced in the `spatstat` package (Turner and Baddeley, 2005), and a map of predicted (log-)intensity. Residuals are computed in each quadrat by comparing the observed number of presences to the number expected within it, then spatially smoothed, to look for regions where there are substantially more or less presence points than expected. In Figure 4.1 (left panel) we can see a clustering of positive residuals just off center, toward the right of the region, suggesting we see slightly more presence points here than the fitted intensity would suggest. However, we note that the magnitude of these (indeed all residuals for this model) are very small indicating that the estimated intensity is closely fitted to the data.

The `scampr` function interfaces with the likelihood-based toolkit with S3 generic functions `logLik()`, `AIC()` and `confint()`, as well as, `residuals()`, `coef()` and `vcov()`. Given it only takes seconds to fit, there is also scope for bootstrap methods via `simulate()`. S3 is the most commonly used object oriented system in `R` (Wickham, 2019), and is used in many common regression-style functions and packages. This aims to make it easier for users to employ similar methods to different models, because the same function can be called irrespective of the type of model object it is being applied to.

Fitting the IPP to the gorilla dataset, we find that AIC is substantially larger than for the LGCP model:

```
    # additionally fit an IPP to the data
ipp_scampr <- scampr(pres ~ elev.std, data=gorillas.df,
    model.type="ipp")


    # compare the AIC for each model
AIC(ipp_scampr, lgcp_scampr)
        model       AIC
1  ipp_scampr -3535.194
2 lgcp_scampr -4679.974
```

This suggests lack-of-fit for the Poisson model, which was expected given other

diagnostics checked previously (Figure 1.2).

4.3 Integrated Data Models

In ecology, when constructing a species distribution model, sometimes data are available from multiple sources and it is of interest to combine them to construct a single joint model, in order to better estimate the environmental response of the target species. The `scampr` package can fit such models in the case where we have presence/absence data as well as a point pattern, and the two data sources are assumed independent conditional on a shared latent field, as in Chapter 3.

As well as observing the point process S_n as previously, we now additionally observe binary responses $y_1, \ldots, y_{n_{\text{survey}}}$, where each y_i is observed at spatial location $\boldsymbol{s}_i^{PA}$ and comes from a Bernoulli random variables with mean $\mu(\boldsymbol{s}_i^{PA})$.

As in Chapter 3, the joint model being fitted is

$$\begin{aligned} \log \lambda(\boldsymbol{s}) &= \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u} + \boldsymbol{B}(\boldsymbol{s})\boldsymbol{\tau} \\ \log\left[-\log\left(1-\mu(\boldsymbol{s})\right)\right] &= \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u} \end{aligned} \tag{4.4}$$

where λ is the intensity that generates the presence-only data and $\mu(\boldsymbol{s})$ is the probability of presence if we were to sample presence/absence data at location $\boldsymbol{s}$. That is, we model both data sources as the same function of environmental variables $\boldsymbol{X}(\boldsymbol{s})$ and missing covariates $\boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u}$, but we assume bias in the presence-only pattern that is a log-linear function of measured bias predictors $\boldsymbol{B}(\boldsymbol{s})$. This model extends the work of Fithian et al. (2015) to include a shared latent field to account for additional spatial dependence between data sources that is induced by, say, missing environmental covariates. See also, Simmonds et al. (2020) for a similarly derived model.

To fit the integrated data model using presence-only and presence/absence data, we can use `scampr()` but need to specify two additional arguments: the presence/absence dataset (`pa.data`) and a formula for the presence/absence model (`pa.formula`). We illustrate using one of several species of flora found in the Greater

Blue Mountains World Heritage Area, available in the `scampr` package as a list object labelled `flora`.

```
    # get the presence-only data for species "sp2"
po_data <- rbind(flora$po$sp2, flora$quad)

    # get the presence/absence data
pa_data <- flora$pa[ , c("sp2","x","y","MNT","MXT")]

head(pa_data)
     sp2       x        y        MNT        MXT
176    0 294.225 6472.975 -0.1581680  1.1590720
3652   0 244.898 6261.929 -0.4829140 -0.6964201
3653   0 245.485 6249.000 -0.3205410  0.4374917
3654   0 246.440 6254.260 -0.5370383 -0.9541274
8199   0 246.510 6129.448  0.8701941 -0.4902544
8200   0 246.820 6129.407  0.8701941 -0.4902544

    # set up a simple basis to approx. the GRF,
    # with 11 nodes along longest edge of domain
bfs <- simple_basis(11, po_data)

    # fit the combined data model
combined_mod <- scampr(pres ~ MNT*MXT + I(MXT^2) + I(MNT^2) +
    D.Urb + D.Main, po_data,
    pa.formula = sp2 ~ MNT*MXT + I(MXT^2) + I(MNT^2),
    pa.data = pa_data, basis.functions = bfs)
```

In the above, we model the distribution of a particular species (`sp2`) against a second order polynomial of average annual minimum (`MNT`) and maximum (`MXT`) temperatures — these are the shared $\boldsymbol{X}$ in Equation (4.4). For the presence-only biasing terms ($\boldsymbol{B}$ in Equation 4.4) we include distances to main road (`D.Main`) and

urban areas (D.Urb). The presence/quadrature identifier is again pres, and it is stored in the presence-only dataset, po_data. The binary response is given by sp2, and it is stored in the presence/absence dataset pa_data. All of the predictors found in pa.formula must also be found in the original presence-only formula. This is because the integrated data model assumes that the datasets arise from the same process, except the presence-only data is thinned by $\boldsymbol{B}(\boldsymbol{s})\boldsymbol{\tau}$. The models fitted and compared above do not use defaults for the basis functions, instead these are supplied by the object bfs — this will be covered in Section 4.5.1.

Below we compare the output from this integrated data model to what we would get from modelling the presence-only data only:

```
    # fit the presence-only data LGCP
po_mod <- scampr(pres ~ MNT*MXT + I(MXT^2) + I(MNT^2) +D.Urb + D.Main,
    po_data, basis.functions=bfs)
summary(po_mod)

Model Type:  Log-Gaussian Cox process - Variational Approx.
Formula:  pres ~ MNT * MXT + I(MXT^2) + I(MNT^2) + D.Urb + D.Main
AIC:  816.2181    approx. marginal logLik:  -399.109
Basis functions per res. 88
Fixed Effects:

            Estimate Std. Error z value  Pr(>|z|)
(Intercept) 14.62357    8.61428  1.6976 0.0447920 *
MNT         13.02200    6.24574  2.0849 0.0185373 *
MXT         -4.25134    5.96853 -0.7123 0.2381421
I(MXT^2)    -6.10353    2.22273 -2.7460 0.0030167 **
I(MNT^2)    -8.89611    2.76458 -3.2179 0.0006457 ***
D.Urb        1.22839    0.79160  1.5518 0.0603567 .
D.Main       0.03878    0.75167  0.0516 0.4794272
MNT:MXT      9.55775    4.71357  2.0277 0.0212950 *
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
---
Spatial Random Effects:
Posterior Means per Spatial Resolution(s):
  Min. 1st Qu. Median Mean 3rd Qu. Max.
1 -290    -216   -153 -130     -35   66
Prior Variance(s):
  res. 1
 3.1e+04

    #compare to integrated data model:
summary(combined_mod)
Model Type:  Combined data model
 w. spatially correlated errors - Laplace approx.
Formula:  pres ~ MNT + MXT + D.Urb + D.Main
|&| sp2 ~ MNT + MXT

AIC:  977.8951    approx. marginal logLik:  -477.9476
Basis functions per res. 88
Fixed Effects:

                Estimate Std. Error  z value  Pr(>|z|)
(Intercept)     -6.11457    1.49327  -4.0947 2.113e-05 ***
MNT              4.46402    1.68201   2.6540  0.003977 **
MXT             -2.17235    1.31521  -1.6517  0.049297 *
I(MXT^2)        -2.27112    0.55188  -4.1152 1.934e-05 ***
I(MNT^2)        -2.98853    0.87492  -3.4158  0.000318 ***
MNT:MXT          3.39893    1.17089   2.9029  0.001849 **
(Bias Intercept) -3.82164    0.31067 -12.3013 < 2.2e-16 ***
D.Urb           -0.67781    0.45309  -1.4960  0.067330 .
D.Main           0.46430    0.33126   1.4016  0.080517 .
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
---
Spatial Random Effects:
Posterior Means per Spatial Resolution(s):
  Min. 1st Qu.  Median   Mean  3rd Qu. Max.
1 -7.3   -0.51 -0.0098 -0.088 -8.5e-07  7.5
Prior Variance(s):
 res. 1
     19
```

In the integrated data model, all coefficients for temperature are estimated jointly across the presence/absence data and the presence-only data, which has led to smaller standard errors on these terms as compared to fitting a model to the presence-only data alone (`po.mod`). But even in the integrated data model, the bias predictors (`D.Urb` and `D.Main`) are still estimated from just the presence-only data, so for these coefficients there is less change in estimates of uncertainty across model fits.

The marginal likelihood of the integrated data model is estimated using a Laplace approximation (`model.type="laplace"`). At this stage, `"variational"` model types are not permitted for models involving presence/absence data (see Section 3.6). We could also fit an integrated data model without latent effects (here equivalent to the model of Fithian et al., 2015, for a single species) by setting `model.type = "ipp"`.

Users can also fit a model to just the presence/absence data, in the form of a binary regression with a complementary log-log link function (with or without spatial latent effects) by omitting the original `formula` and `data` arguments from the function. This is done in the example below to create `pa_mod`. Most functionality described earlier for `scampr` point process models are also available for integrated data models, with additional residuals and fitted values available for the presence/absence data component.

```
    # fit the presence/absence data binary regression
pa_mod <- scampr(pa.formula = sp2 ~ MNT*MXT + I(MXT^2) + I(MNT^2),
```

```
    pa.data=pa_data, basis.functions=bfs)


    # compare the likelihoods:
cbind(Separate = logLik(po_mod) + logLik(pa_mod),
    Combined = logLik(combined_mod))
      Separate  Combined
[1,] -633.7554 -477.9476
```

In the example above we find the combined data model achieves a higher fitted likelihood than that achieved modelling the datasets separately. However, this is not always the case. As we saw in Section 3.5, properly accounting for presence-only bias is challenging, but failure to adequately do so can mean integrated data models actually perform worse than individual data models.

4.4 Model Diagnostics and Inference

Formal model diagnostics for point process models are somewhat limited (see for example Baddeley et al., 2011). Models fitted in `spatstat` include residuals as described in Baddeley et al. (2005) and subsequently Baddeley et al. (2013). We have written code to also compute these residuals for point process models fitted using `scampr`, accessible via the generic function `residuals()`. Options include: `raw`, `inverse` and `pearson`. Raw residuals are unity at presence points and at quadrature points they are $-\hat{\lambda}(\boldsymbol{s}_i) w_i$ where as previously w_i is the quadrat size. Inverse and Pearson residuals then scale these according to the fitted intensity and square-root of the fitted intensity respectively. These definitions of residuals are derived from signal processing and survival analysis (*i.e.* one dimensional point processes in time) and exploit the idea of observed minus fitted from the raw residual process: $N(\mathcal{D}) - \int_{\mathcal{D}} \hat{\lambda}(t)\, dt$ (Baddeley et al., 2005). To plot residuals, we then compute a quadrat-level residual by summing all residuals that fall within it (for its quadrature point, plus any presence points in the quadrat). We then smooth the resulting residual surface using `image.smooth()` from the `fields` package (Nychka et al., 2017) — see, *e.g.* Figure 4.1 (left panel) for raw residuals. For the presence/absence component of an

integrated data model, the raw residuals are simply $y_i - \hat{\mu}_i$ at each of the $\boldsymbol{s}_i^{\text{PA}}$ survey sites for $i = 1, \ldots, n_{\text{survey}}$ and are accessed by setting the argument `data.type == "pa"` within `residuals()`.

Perhaps the most commonly used summary statistic for point process models is Ripley's K function (Ripley, 1977) as well as its inhomogeneous extension (Baddeley et al., 2000). The inhomogeneous K function can be used to diagnose clustering or inhibition between points (Baddeley et al., 2011). This is done by simulating point patterns from a model's fitted intensity and comparing the simulated and observed K functions. Naïve simulation envelopes, constructed by calculating critical quantiles in a pointwise fashion, are difficult to interpret because they only offer pointwise control of Type I error. A better solution is to construct global confidence bands that account for the functional nature of the data (Myllymäki et al., 2017), to ensure for example that if model assumptions are satisfied, then 95% of the time, sample K functions will fall entirely within their 95% confidence bands. In the `scampr` package, the function `kfunc_envelopes()` is used to construct global simulation envelopes around an inhomogeneous K function of a `scampr` model fit. Global envelopes are constructed using the `GET` package (Myllymäki and Mrkvička, 2019).

Examples of this, testing the IPP and LGCP model fits to the gorilla nesting data, are shown in Figure 4.2. An individual inhomogenous K function is calculated for a scampr model via the function `kfunc()` which acts as a wrapper for `spatstat::Kinhom()`. `kfunc_envelopes()` uses this and the generic `simulate()` which simulates point patterns from any fitted `scampr` point process model. Note however that simulation requires knowledge of the values of predictors all over the spatial domain $\mathcal{D}$. Hence quadrature points need to be sufficiently dense that they adequately characterise the domain to simulate from, or an additional argument `domain.data` is needed — a data frame containing a dense set of coordinates and formula terms over the entire domain of interest. The `simulate()` function in turn employs the generic `predict()` which predicts a `scampr` model's (log-)intensity at locations provided in the argument `newdata`.

As a `scampr` model is fitted using maximum (albeit approximate) likelihood, we can employ likelihood ratio testing and information criteria. Components for these

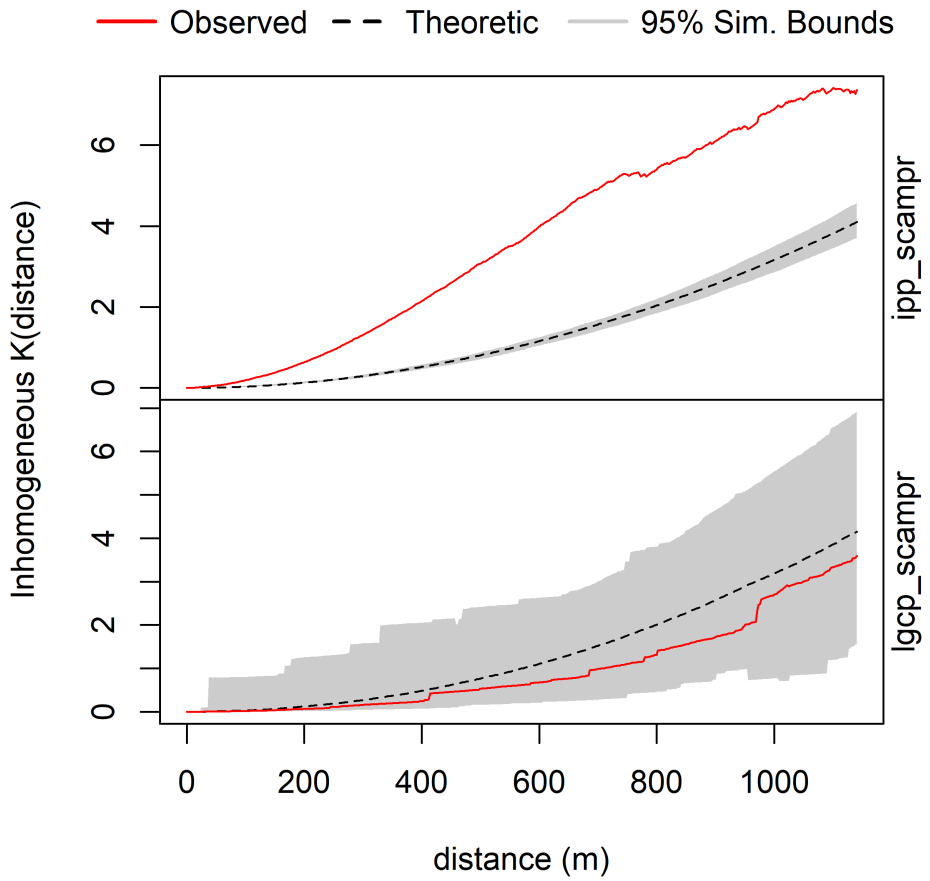

Figure 4.2: These are the resulting global envelope tests of inhomogeneous K functions from `kfunc_envelope(ipp_scampr)` (top panel) and `kfunc_envelope(lgcp_simple)` (bottom panel). There is strong evidence to suggest that there is additional clustering present beyond that which the IPP model can account for, since the observed K function exceeds the simulated bounds (top panel) by a considerable margin. There is no such evidence for the LGCP model (bottom panel).

are accessed with generic functions `logLik()` and `AIC()` respectively. We can use this to, for example, examine changes in likelihood across various configurations of basis functions to assess the adequacy of the approximated latent field structure — this is explored further in Section 4.5.1. Additionally, to estimate the variance-covariance matrix of the parameter estimators, `scampr` uses the Hessian matrix from the (approximate) marginal likelihood. With automatic differentiation providing exact derivatives, these are fast and easy to compute. Confidence intervals on parameters of a model can be computed using generic function `confint()`, which constructs Wald intervals.

```
confint(lgcp_scampr)
```

Table 4.1: The defaults for arguments in the `scampr` model function, including reasoning. For `basis.functions` the default is NULL, in which case `scampr` uses the default basis functions assigned by `FRK::auto_basis()` with `max.basis` equal to one quarter of the number of observations in the data (unless basis functions are not needed, *i.e.* when `model.type = "ipp"`).

Argument	Default	Reasoning
`formula`		
`data`		
`pa.formula`		
`pa.data`		
`coord.names`	`c("x", "y")`	NA
`quad.weights.name`	`"quad.size"`	NA
`basis.functions`	`NULL`	see caption
`model.type`	`"variational"`	fast/stable model for LGCP
`sparse`	`TRUE`	exploits sparse matrix operations
`se`	`TRUE`	calculates standard errors
`starting.pars`	`NULL`	parameters ($\boldsymbol{\theta}$) start at 0 (or 1, $\forall\theta > 0$)
`subset`	`NULL`	includes entire dataset

```
                                  2.5 %       97.5 %
(Intercept)                 -4.327365e-01  1.701884421
elev.std                    -2.620211e-01  0.666937194
VA Posterior Mean (bf 1.1)  -1.799269e-03  0.001793999
VA Posterior Mean (bf 1.2)  -1.778747e-03  0.001780693
VA Posterior Mean (bf 1.3)  -1.779502e-03  0.001781753
VA Posterior Mean (bf 1.4)  -1.776949e-03  0.001777337
VA Posterior Mean (bf 1.5)  -1.777244e-03  0.001776950
VA Posterior Mean (bf 1.6)  -1.795568e-03  0.001801060
VA Posterior Mean (bf 1.7)  -1.779544e-03  0.001781810
VA Posterior Mean (bf 1.8)  -1.799065e-03  0.001805023
...
```

4.5 Fine Tuning `scampr()`

In this section we go into the details on the key step of choosing the basis function configuration for the latent field, and advice on improving computation speed.

4.5.1 Basis Functions

To enable fast, likelihood-based fitting of LGCP, `scampr` models approximate the GRF with a linear combination of basis functions ($\boldsymbol{Z}(\boldsymbol{s})$ in Equations 4.1 and 4.4) within the framework of Cressie and Johannesson (2008), called fixed rank kriging (FRK). This framework is available in `R` as a package of the same name, `FRK` (Zammit-Mangion and Cressie, 2017). The `scampr` function is designed to interface with `FRK` by using the suite of basis functions available to `FRK::auto_basis()`, defaulting to a maximum number of basis functions (`max.basis`) equal to one quarter of the number of presence locations in the model. Rather than using this default configuration, any other set of basis functions could be constructed using the `FRK` package (of class `"Basis"`) and used in `scampr()` via the argument `basis.functions`. All available options for basis functions within the `FRK` package are shown in Figure 4.3. These functions can be calculated over different topologies including a line, a 2D plane or a spherical manifold, see Zammit-Mangion and Cressie (2017) for details.

In Chapter 2, we found that the best choices of basis function configuration had few or no estimable parameters, and only local support such that $\boldsymbol{Z}$ is sparse. Such basis function configurations were fast to fit with little detriment to statistical efficiency (see Sections 2.3.2 and 2.5). For this reason we recommend the use of locally supported, bi-square basis functions (Equation 2.8). Even though the `FRK` package offers this as an option, we have found some of the defaults to be less useful in the context of `scampr` LGCP models, perhaps because fitting a LGCP is not the primary intended use case for the `FRK` package. In particular, the defaults seem to choose more basis functions, at more spatial resolutions, than is optimal for our models. For example, when fitting the integrated data model to flora data in Section 4.3, the default `FRK` basis functions selected so many basis functions that the Hessian of the model fit was singular. The issue is that the number of basis functions chosen by `FRK` is informed by the number of quadrature points (which was large, $q \approx 86,000$, for the flora data) as well as based on the number of presence points. Instead, in Section 4.3 we used a purpose written *simple basis* option for basis function selection, which uses a regular grid of locally supported, bi-squared basis functions.

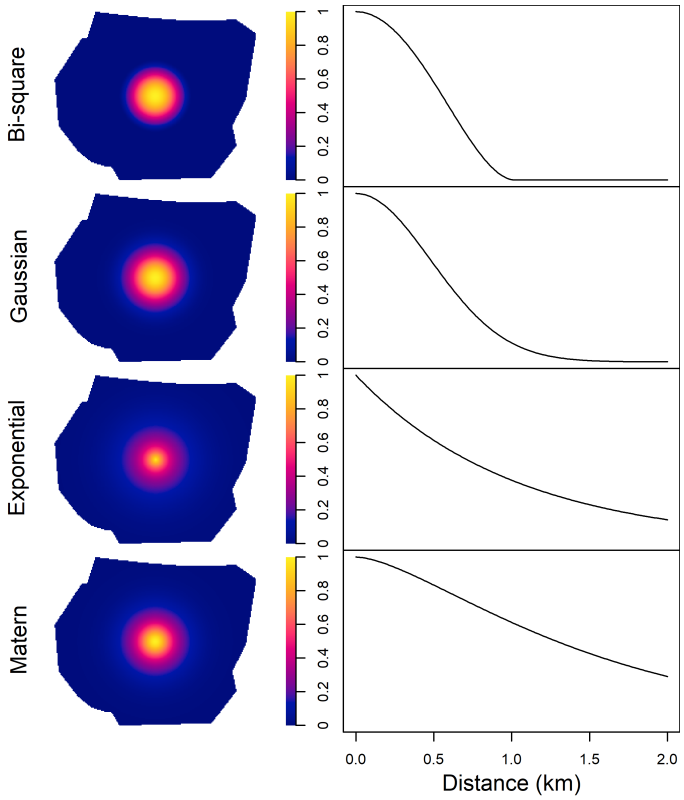

Figure 4.3: Basis functions available to `scampr` via the `FRK` package. These are (from top to bottom) local bi-square, Gaussian, Exponential and Matern (with smoothness parameter 3/2) — each using `FRK` default values for range parameters. Bi-square basis functions facilitate sparse matrices that lead to the fastest computation times.

Simple basis functions are supplied to the `scampr()` model function via the argument `basis.functions = simple_basis()`. This creates a `data.frame` (of class `"bf.df"`) that describes a single resolution of regularly spaced, local bi-square functions. At this stage these are only compatible within 2D Euclidean geometry. The argument `nodes.on.long.edge` gives the number of nodes to place along the widest axis of the data provided — compared to `FRK::auto_basis()` this offers greater control over the number of basis functions used. The default radius of the functions is set to the diagonal distance between nodes to ensure there are no gaps in coverage of the domain. This also ties the choice of the number of basis functions, k, to their radius and effectively means that choosing k is a proxy for choosing the range of effect of the latent field — see Section 2.3.2.

The `scampr` package does not permit users to supply a custom-built matrix describing the basis functions at both the presence points and quadrature points because

some of the functions we may want to apply to scampr objects in the package (such as simulate()) will require calculation of basis function values at new locations in the domain. Instead, users must provide details of the basis functions themselves — currently via either FRK::auto_basis() or simple_basis().

In the gorilla nesting data example, looking at the output we see that there may be an opportunity to improve upon the default method of basis function selection:

```
summary(lgcp_scampr)
...
Spatial Random Effects:
Posterior Means per Spatial Resolution(s):
     Min.  1st Qu.   Median     Mean 3rd Qu.    Max.
1 -3.5e-06  1.1e-07  1.1e-06  1.0e-06   3e-06 3.2e-06
2 -1.9e+00 -5.7e-01 -3.0e-01 -4.3e-02   3e-01 2.9e+00
Prior Variance(s):
 res. 1 res. 2
7.4e-07    1.6
```

The object lgcp_scampr, fitted using FRK::auto_basis() under default settings, has fitted basis functions at two resolutions, called res. 1 and res. 2 in the summary() output. Note however that there is evidence that including basis functions at the first spatial resolution are unnecessary, since its variance component is nearly computationally zero. Further inspection of the original output in Section 4.2 shows there were 12 basis functions in this first resolution, and 99 in the second. These are attempting to capture difference spatial scales of latent effect. The second resolution of basis functions has a variance component of 1.6, meaning that this finer scale configuration is making the bulk contribution to the estimated intensity surface. We can instead set up a *simple basis*, and do so by placing nine evenly spaced functions across the widest axis of the data as in the code below. We see an improved AIC when using this basis configuration:

```
# set up the basis functions
```

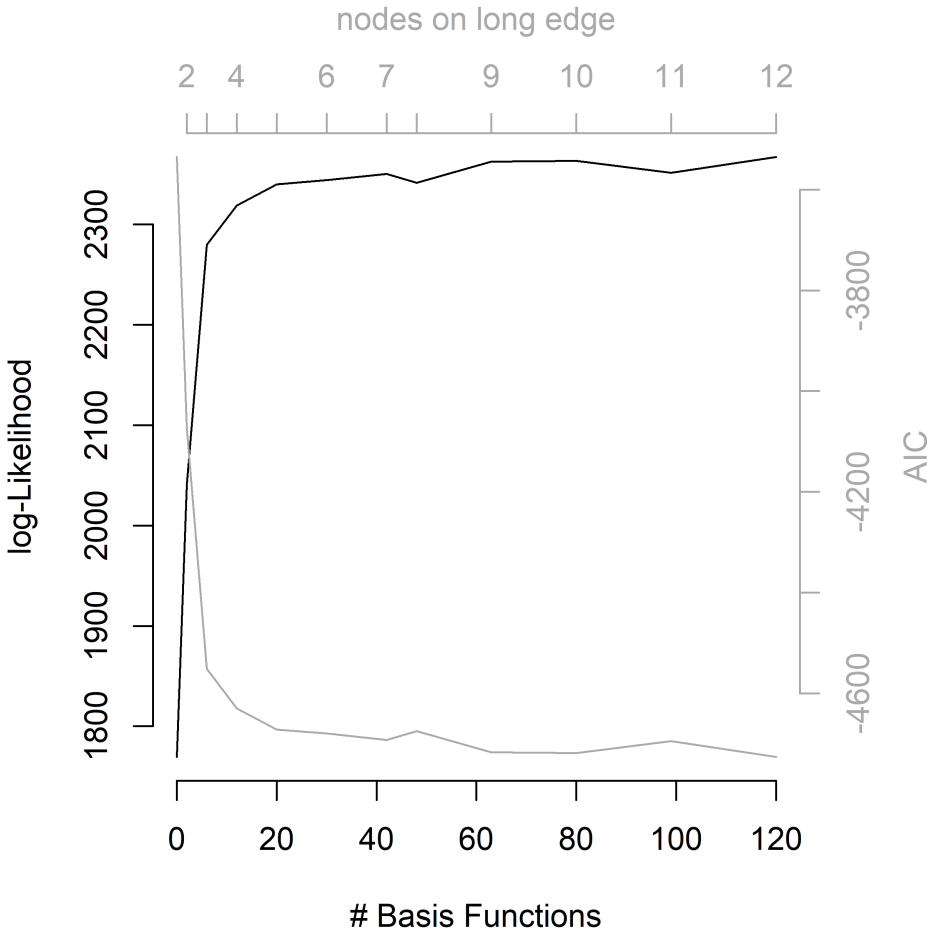

Figure 4.4: The results from `simple_basis_search(ipp_scampr)` for the gorilla nesting point process model example. These show the fitted log-likelihood and AIC over increasingly dense regular grids of basis functions, used to approximate the latent field in the LGCP. Over this range of basis function configurations we found the likelihood does not significantly change (Likelihood Ratio < 10, or similar for AIC) once we used $60-80$ basis functions.

```
bfs <- simple_basis(nodes.on.long.edge=9, data=gorillas.df)
lgcp_simple <- scampr(pres ~ elev.std, data=gorillas.df,
    basis.functions = bfs)
AIC(lgcp_scampr, lgcp_simple)
        model       AIC
1 lgcp_scampr -4679.974
2 lgcp_simple -4699.697
```

Choosing an appropriate number of basis functions to approximate the latent random field can be challenging, however the computational speed of `scampr` models that use `simple_basis()` allows us to fit many configurations and compare a

likelihood-based metric. The function simple_basis_search() allows the user to fit increasingly dense regular grids of local bi-square basis functions as created by simple_basis(). Fitting this many models would be computationally burdensome using previous software, but is feasible here given our rapid model fits.

As an example, we illustrated that the dual resolution of basis functions from the FRK default may not have been suitable for the gorilla nesting LGCP model. We arrived at the basis configuration used in lgcp_simple (*i.e.* nodes.on.long.edge=9) by performing such a simple basis search. The results are found in Figure 4.4 and suggest that far fewer basis functions (60-80) achieve the highest likelihood and lowest AIC.

```
    # model fits on increasing # basis fns
lls <- simple_basis_search(ipp_scampr, max.basis.functions=100)
[1] "Completed fit with 0 basis functions (IPP)"
[1] "Completed fit with 2 basis functions"
[1] "Completed fit with 6 basis functions"
[1] "Completed fit with 12 basis functions"
[1] "Completed fit with 20 basis functions"
[1] "Completed fit with 30 basis functions"
[1] "Completed fit with 42 basis functions"
[1] "Completed fit with 48 basis functions"
[1] "Completed fit with 63 basis functions"
[1] "Completed fit with 80 basis functions"
[1] "Completed fit with 99 basis functions"
lls
   nodes.on.long.edge  bf   loglik       aic
1                   1   0 1769.597 -3535.194
2                   2   2 2040.920 -4073.841
3                   3   6 2279.875 -4551.749
4                   4  12 2319.031 -4630.061
5                   5  20 2340.060 -4672.120
6                   6  30 2344.037 -4680.074
```

```
7                7  42 2350.533 -4693.067
8                8  48 2341.700 -4675.401
9                9  63 2362.718 -4717.437
10              10  80 2363.460 -4718.920
11              11  99 2351.468 -4694.937
```

Users can also perform spatial cross-validation, calculating out-of-sample predicted likelihoods (Equations 3.4 and 3.5), by adding to their `simple_basis_search` call the argument `po.fold.id` (and also `pa.fold.id` for integrated data models) — this could be used to do analyses along the lines of Figure 2.7b and Figure 3.6. These arguments are required as integer or factor vectors, and must be the same length as the data, describing the CV fold into which each location falls.

4.5.2 Speed Control

A key advantage of `scampr` is fast computation time, but there are a number of decisions that can be made in model-fitting that can have considerable implications for computational efficiency.

The first decision to make that has speed implications is choice of number and type of basis functions, discussed above. Computation speed slows down as the number of basis functions increases, and in cases of extreme overfitting, can become unstable. As previously, using choices of basis functions that encourage sparseness also encourages computational efficiency.

The second decision with speed implications is how to approximate the marginal likelihood, which can be controlled via the argument `model.type`. As previously demonstrated, when this is set to `"ipp"`, `scampr()` fits a model without latent effects. The IPP is by far the fastest model to fit, because the model involves no random effects and its likelihood does not involve an intractable marginalising integral, but it often fails to account for clustering in point patterns (as seen in Chapters 2 and 3). The other two options are `"laplace"` or `"variational"` — these are covered in detail in Section 2.3.1. The default (for LGCP models) is to use a variational approximation as this tends to be more stable and much faster to

calculate. Models using a Laplace approximation can fail due to overfitting (*e.g.* Table 2.1 and Figure 2.7), but in some instances can yield more accurate inferences (see Figures 2.4 and 2.5). As mentioned in Section 4.3, at this stage, the only option for implementing spatial random effects for an integrated data model is via the `"laplace"` option.

The third decision to make with speed implications is whether or not to store basis functions as a sparse matrix. This is specified via the argument `sparse`, which defaults to `TRUE`. Using sparse matrices is appropriate when using local bi-square basis functions, however in some other instances, *e.g.* Gaussian kernels with a long enough correlation range, storing these in sparse matrices actually slows down computation time.

The final decision that can have speed implications is choice of starting values for parameters. If good starting values are given, the journey to the maximum likelihood estimate will be much faster, and often, more likely to avoid problematic areas of the parameter space. This can be controlled via the argument `starting.pars` which will accept a named list of parameter values or another `scampr` model. We find the latter particularly useful, *e.g.* a user can quickly fit an IPP and then pass this model to the LGCP as starting values. Similarly, a user might initially fit a LGCP using VA but then use this solution as starting values for a fit using the Laplace approximation, to overcome potential Laplace issues with slowness and instability, for large models.

In the following we demonstrate some of the computational speed differences described in this section:

```
    # fit a scampr model using dense matrix operations
lgcp_simple_dense <- scampr(pres ~ elev.std, gorillas.df,
    basis.functions=bfs, sparse=F)
    # fit a Laplace version of the model
lgcp_simple_laplace <- scampr(pres ~ elev.std, gorillas.df,
    model.type="laplace", basis.functions=bfs)
    # fit a Laplace version of the model w. starting parameters
```

```
lgcp_simple_laplace_warm_start_pars <- scampr(pres ~ elev.std,
    gorillas.df, model.type="laplace", starting.pars=lgcp_simple,
    basis.functions=bfs)

    # compare the timing for all the models (in seconds)
ipp_scampr$cpu
   user  system elapsed
   0.04    0.00    0.06
lgcp_scampr$cpu
   user  system elapsed
  13.59    1.39   15.11
lgcp_simple$cpu
   user  system elapsed
   2.67    0.31    2.99
lgcp_simple_dense$cpu
   user  system elapsed
   3.04    0.44    3.49
lgcp_simple_laplace$cpu
   user  system elapsed
   6.90    0.42    7.34
lgcp_simple_laplace_warm_start_pars$cpu
   user  system elapsed
   4.58    0.34    4.92
```

4.6 Discussion

This chapter has introduced and illustrated the use of `scampr` — an `R` package that can be used to quickly fit a variety of latent effect models involving presence-only data (or more generally point patterns) based on LGCP. While other software exists to fit LGCP regression models, many involve long computation times and are not easy to use. Further, by using a maximum likelihood framework, we are able to access standard tools for likelihood-based inference and model selection. The

`scampr` package was also designed to maintain syntactic simplicity, similar in form to other popular regression modelling software. The speed advantages enable the use of diagnostic tools, exploratory fitting and cross validation procedures that were previously computationally prohibitive. In addition, it becomes computationally feasible to fit more complex models, such as incorporating presence/absence data into joint models capable of sharing latent effects.

There are plans to extend some of the functionality of `scampr` beyond that available at the time of writing. Models can accommodate data on a plane, line or sphere via the use of basis functions from `FRK` (Zammit-Mangion and Cressie, 2017). However, models using `simple_basis()` do not share this functionality and are currently limited to 2D Euclidean geometry. So too are the plotting functions: `plot()` and `image()`. Diagnosing model validity is also an area that could be improved for point process models (Baddeley et al., 2011). Beyond the residuals available to `scampr` models (as per Turner and Baddeley, 2005), randomised quantile residuals (Dunn and Smyth, 1996) would be an appropriate extension given the discrete nature of point patterns, and could be computed (conditionally on a point estimate of the latent field) directly from the Poisson cumulative distribution function. The `R` package `DHARMa` (Hartig, 2017, unpublished) uses simulation to estimate these residuals, and could be used if it were of interest to compute residuals that accounted for uncertainty in estimates of $\boldsymbol{\xi}(\boldsymbol{s})$. While these extensions would provide useful functionality to the `scampr` package, in its current form the package should assist researchers in fitting complex spatial models to their data quickly and easily. We hope this encourages greater uptake of spatial models that include latent fields to guard against model misspecification.

Chapter 5

Final Remarks

5.1 Summary

Log-Gaussian Cox processes are a useful framework for modelling point patterns, as the latent Gaussian random field provides a way to induce spatial correlation beyond that accounted for by predictor variables included in the model. This additional spatial correlation is important because presence-only data often exhibit further clustering (*e.g.* see Figures 1.2 and 1.6) to that explained by environmental predictors, largely due to model misspecification. In particular, some key phenomena driving the species' spatial distribution are not routinely included in models, such as fine-scale environmental variation, dispersal and biotic effects (Elith and Leathwick, 2009; McInerny and Purves, 2011; Wisz et al., 2013). LGCP models have sometimes been used in the ecological literature (*e.g.* using methods proposed by Taylor et al., 2013; Simpson et al., 2016; Bachl et al., 2019), but computation times have previously been lengthy and a barrier to more widespread use. In this thesis we have proposed a novel methodology for approximate fitting of LGCP, orders of magnitude faster than existing methods (specifically comparing to INLA, as in Rue et al., 2009).

The methodology proposed in this thesis enables the use of procedures important to the model-fitting process that were previously inaccessible, including diagnostics tools, model selection and validation. For example, because we have fast tools to fit a LGCP, simulation envelopes (as in Figures 1.2 and 1.6) can be used in model checking, which would be computationally prohibitive using other model-fitting algorithms. We were able to fit many models with different numbers (Figures 2.7 and 3.6) and types (Table 2.3 and Figure 4.3) of basis functions to guide decisions made in tuning our model. Additionally, the methodology proposed in this thesis uses a maximum likelihood framework, giving access to likelihood-based statistical tools such as information criteria (Section 4.5.1), or likelihood ratio testing for inference from models (Figure 4.4).

The fast computation times for our methodology open up the potential for LGCP to be used in more complex settings, such as for data integration in species distribution modelling. Data integration is a popular technique in the ecological literature for

combining sources of data (Dorazio, 2014; Fithian et al., 2015; Koshkina et al., 2017; Fletcher Jr et al., 2019). However, little attention has been given to spatial dependence between the data sources, again arising from missing or unmeasured predictors. We show that a latent field shared across data sources can account for this — in a LGCP-style extension of the work of Fithian et al. (2015). However, we found that if bias in presence-only data is not properly accounted for using measured predictors, data integration can actually be detrimental to SDMs — a similar finding to Simmonds et al. (2020). In our application to real data, we tended to find no improvement to SDMs when integrating presence-only data into a model for systematically collected data. We note that beyond the simulation setting (*e.g.* Dorazio, 2014; Fithian et al., 2015; Simmonds et al., 2020) it is not common to see evidence of improvement in the literature for real data applications (although see Koshkina et al., 2017). Conn et al. (2017), exploring the related problem of preferential sampling, had a similar experience. We suspect that the issue here is that models for the bias in presence-only data tend to be inadequate.

The final contribution of the thesis is the `scampr` package, freely available software in `R` (https://github.com/ElliotDovers/scampr), to fit LGCP regression modelling on point patterns and integrated data models that combine point patterns with presence/absence data. This builds upon the functionality of Turner and Baddeley (2005) in the `spatstat` package for point process models, permitting novel use of LGCP models due to fast computation times. This is facilitated by access to fast optimisation tools via the `TMB` package (Kristensen et al., 2016) and reduced rank estimation of spatial covariance via the `FRK` package (Zammit-Mangion and Cressie, 2017).

5.2 Future Research

We see several key, open questions that have been raised during the completion of this thesis which we highlight here.

We previously discussed the delicacy with which spatial confounding must be treated when adding spatial latent effects in Chapter 2. In particular, we found in simula-

tions that recovering the true model coefficients is most difficult when the latent field and covariates are acting at similar spatial scales. Likewise, when we set up basis functions to approximate the latent field, correlation between these and covariates in the model can also hinder coefficient recovery. Although our fast-fitting method makes exploratory analysis of models with different latent specifications less of a computational challenge, problematic spatial confounding can still arise. Hodges and Reich (2010) demonstrate this applies generally for linear models with spatially correlated errors, and propose so-called restricted spatial regression. This addresses collinearity by transforming variables relative to each other, such that the set of model predictors is orthogonal. A similar approach could be adopted in setting up basis functions to approximate the latent field of a LGCP model here. We found in Section 2.4 that while spatial confounding tended to make inference on, and estimation of, fixed environmental effects difficult, it had little effect on the predictive power of the model. Hence we suggest that the issue of how spatial confounding is handled should come down to the purpose of the analysis. If inference about the relationship of the point pattern to the environment is the primary aim then care must be taken. However, if the goal is prediction then a researcher need not worry because there is no need to tease apart contributions from fixed effects vs the latent field.

We have seen that a key consideration when modelling presence-only data is handling bias, and that when this is not appropriately accounted for, integrated data models can actually perform worse than if presence-only data was omitted (Section 3.5). One opportunity to improve our bias modelling is to estimate the bias in presence-only data simultaneously for multiple species using a joint species distribution model (JSDM).

Jointly modelling multiple species in the form of JSDMs allow researchers to estimate the distributions of species simultaneously and account for their co-occurrence, in addition to environmental response (Pollock et al., 2014) — a popular approach in the ecological literature. In our current context, if we are able to assume that biases found in presence-only data are constant across different taxa, it may be possible to untangle the two sources of additional clustering mentioned above and, in turn,

address the shortcomings of data integration identified here. Indeed Fithian et al. (2015) do this but in an inhomogeneous Poisson process setting, without a latent field to account for additional clustering or spatial dependence between data sources. A logical extension to the current body of work would be to fit a multi-species model with a common latent Gaussian field, shared across species, to capture missing predictors in the model for observer bias. Such a model assumes observer bias is not a function of species, which is reasonable in many instances – observer bias is primarily a property of the observer (*e.g.* accessibility) rather than being a property of the species being modelled. Consider the following extension of Equation (3.3) where for species $j = 1, \ldots, n_{\text{species}}$ the intensities for the presence-only data (λ) and the mean for the presence/absence data (μ) are given by

$$\begin{aligned} \log \lambda_j(\boldsymbol{s}) &= \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}_j + \boldsymbol{Z}_1(\boldsymbol{s})\boldsymbol{u}_j + \boldsymbol{Z}_2(\boldsymbol{s})\boldsymbol{\tau} \\ \log\left[-\log\left(1-\mu_j(\boldsymbol{s})\right)\right] &= \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}_j + \boldsymbol{Z}_1(\boldsymbol{s})\boldsymbol{u}_j. \end{aligned}$$

where $\boldsymbol{Z}_1(\boldsymbol{s})\boldsymbol{u}_j$ approximates the latent field that handles the spatial dependence between data sources for species j and $\boldsymbol{Z}_2(\boldsymbol{s})\boldsymbol{\tau}$ approximates the latent field handling the presence-only bias, assumed to be constant across species. Further, we could induce correlation across species by assuming $\{u_{r,j}\}_{j=1}^{n_{\text{species}}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{species}})$ for the rth spatial random effect. This model could be particularly useful in modelling rare species, for which there is little information with which to estimate observer bias effects, but borrowing strength from more abundant species (Ovaskainen and Soininen, 2011; Pollock et al., 2014) in a joint model could allow us to appropriately account for this bias.

A model along the lines of the above has not been fitted before – integrated species distribution models are already computationally challenging, a multi-species version very much more so. Introducing correlation across species via correlated $\boldsymbol{u}_j$ is the main difficulty. But the computational advances in LGCP modelling in this thesis, combined with recent innovations in fitting JSDMs (Niku et al., 2019) make a combined data JSDM computationally feasible. In the multi-species setting there are also questions to consider around how modelling aims inform performance metrics

(Wilkinson et al., 2021), and the issue of imperfect detection (Tobler et al., 2019) which could be quantified using multiple site visits.

Finally, we believe there are some functional extensions to the methodologies described in this thesis that may be important to future research.

Currently we have assumed all data used is static in time, however, spatio-temporal models would seem important to understanding species response in a changing environment, and are a relatively straight-forward extension. Climate change is a necessary consideration in monitoring and managing populations of species into the future (Parmesan, 2006) so treating data as temporally static when fitting a SDM may not yield answers to important research questions. Likewise, assuming that species' response to the environment is static in time may not be reasonable. In the current context, whether a temporal aspect is introduced into the model discretely or as a continuous process would depend on how the dataset is structured. For example, Zammit-Mangion et al. (2012) modelled conflict data using a discrete-time series of continuous-space LGCPs, because conflict data is often logged discretely, as the day of event rather than precise timing. In contrast, Shirota and Gelfand (2017b) analysed crime events by considering time-of-day as continuous and cyclical.

Another functional extension, particular to the `scampr` package introduced in this thesis, is to allow models to handle a spherical topology. This would permit use for global datasets and should be a straight-forward extension. The `FRK` package (Zammit-Mangion and Cressie, 2017) already permits this for the spatial basis functions used by `scampr`. A secondary challenge would arise in adopting spherical geometry for the elements of the package that are built upon the Euclidean functionality within `spatstat` (Turner and Baddeley, 2005). This requires adopting to the spherical domain procedures for interpolation, domain definition, and boundary handling.

In providing a novel approach to quickly fitting LGCPs to point patterns; demonstrating how this can be exploited for integrated data models in ecology; and supplying software for researchers to easily implement these advances; we have set a platform that can be built upon in numerous ways. We are excited to see where

this goes and what can now be achieved with presence-only data in future ecological research.

Appendix A

Additional results for Chapter 2

A.1 Variational approximation to the proposed rank-reduced, marginal, log-likelihood for a LGCP Model

Consider point pattern $S_n = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ driven by LGCP with $\ln \lambda(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u}(\boldsymbol{s})$, with the columns of $\boldsymbol{X}$ being the p predictors (along with corresponding fixed effect coefficients $\boldsymbol{\beta}$, for conciseness this may include an intercept, β_0 with $X_0(\boldsymbol{s}) = 1$). The linear combination of k basis functions and random coefficients, $\boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u}$, approximates a latent Gaussian process so that the intensity of the LGCP is stochastic, inheriting the randomness from $\boldsymbol{u} \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{prior}})$. To fit the LGCP in a frequentist setting we wish to optimise the marginalised joint density (*i.e.* marginal log-likelihood) given by

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \log \int \pi(S_n, \boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} \\ &= \log \int \pi(S_n|\boldsymbol{u}) \, \pi(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}\end{aligned}$$

with respect to the fixed effects, $\boldsymbol{\beta}$. Here we are using π as a transferable density function, individual instances of which are distinguished by the function arguments.

We can get around intractability of the integral within these equations via variational

approximation. The following is a derivation of the variational lower bound.

By definition

$$\ell(\boldsymbol{\beta}) \equiv \ln \pi(S_n)$$

Now we introduce the variational density over the k dimensional random vector $\boldsymbol{u}$. Call the density π_{VA} and so we have that $\int \pi_{\text{VA}}(\boldsymbol{u})\, d\boldsymbol{u} = 1$ and hence we can write the log-likelihood as

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \ln \pi(S_n) \cdot \int \pi_{\text{VA}}(\boldsymbol{u})\, d\boldsymbol{u} \\ &= \int \pi_{\text{VA}}(\boldsymbol{u}) \ln \pi(S_n)\, d\boldsymbol{u}\end{aligned}$$

Now consider the conditional density of the random vector $\boldsymbol{u}$, given the point pattern, S_n and note that $\frac{\pi(\boldsymbol{u}|S_n)}{\pi(\boldsymbol{u}|S_n)} = 1$. This allows us to write

$$\int \pi_{\text{VA}}(\boldsymbol{u}) \ln \pi(S_n)\, d\boldsymbol{u} = \int \pi_{\text{VA}}(\boldsymbol{u}) \ln \left[\frac{\pi(\boldsymbol{u}|S_n)}{\pi(\boldsymbol{u}|S_n)} \cdot \pi(S_n)\right] d\boldsymbol{u}$$

Next, since $\pi(\boldsymbol{u}|S_n) \cdot \pi(S_n) = \pi(S_n, \boldsymbol{u})$ and noting that $\frac{\pi_{\text{VA}}(\boldsymbol{u})}{\pi_{\text{VA}}(\boldsymbol{u})} = 1$, we can write

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \int \pi_{\text{VA}}(\boldsymbol{u}) \ln \left[\frac{\pi(S_n, \boldsymbol{u})}{\pi(\boldsymbol{u}|S_n)} \cdot \frac{\pi_{\text{VA}}(\boldsymbol{u})}{\pi_{\text{VA}}(\boldsymbol{u})}\right] d\boldsymbol{u} \\ &= \int \pi_{\text{VA}}(\boldsymbol{u}) \ln \left(\frac{\pi(S_n, \boldsymbol{u})}{\pi_{\text{VA}}(\boldsymbol{u})}\right) + \pi_{\text{VA}}(\boldsymbol{u}) \ln \left(\frac{\pi_{\text{VA}}(\boldsymbol{u})}{\pi(\boldsymbol{u}|S_n)}\right) d\boldsymbol{u} \\ &= \underline{\ell}_{\text{VA}}(\boldsymbol{\beta}) + D_{\text{KL}}\left[\pi_{\text{VA}}(\boldsymbol{u}) \,||\, \pi(\boldsymbol{u}|S_n)\right]\end{aligned}$$

$\underline{\ell}_{\text{VA}}(\boldsymbol{\beta})$ is the variational lower bound in Equation (2.5) as $D_{\text{KL}}\left[\pi_{\text{VA}}(\boldsymbol{u}) \,||\, \pi(\boldsymbol{u}|S_n)\right]$ is the non-negative Kullback-Leibler (KL) divergence from the true (and unknown) "posterior" distribution of $\boldsymbol{u}$ to the variational density $\pi_{\text{VA}}(\boldsymbol{u})$. This second term will be close to zero provided π_{VA} is close (in divergence) to the true posterior. The first term is used as an approximate likelihood and is a lower bound for the full likelihood. This we term the variational approximation to the (log-)likelihood or, more concisely the VA likelihood.

There are a variety of types of variational approximation, we use a parametric approach (Ormerod and Wand, 2010) and assume that $\boldsymbol{u}|S_n \overset{\pi_{\text{VA}}(\cdot)}{\sim} \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and choose parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that simultaneously (with $\boldsymbol{\beta}$) maximise $\underline{\ell}_{\text{VA}}$. Maximising $\underline{\ell}_{\text{VA}}$ with respect to variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ minimises the KL divergence between $\pi_{\text{VA}}(\boldsymbol{u})$ and $\pi(\boldsymbol{u}|S_n, \boldsymbol{\beta})$, that is, it finds parameters that make the variational approximation $\pi_{\text{VA}}(\boldsymbol{u})$ as close as possible to the true posterior density of $\boldsymbol{u}$ (in KL divergence).

The VA likelihood can be reexpressed as

$$\begin{aligned}\underline{\ell}_{\text{VA}}(\boldsymbol{\beta}) &= \int \pi_{\text{VA}}(\boldsymbol{u}) \ln\left(\frac{\pi(S_n, \boldsymbol{u})}{\pi_{\text{VA}}(\boldsymbol{u})}\right) d\boldsymbol{u} \\ &= \int \pi_{\text{VA}}(\boldsymbol{u}) \ln\left(\frac{\pi(S_n|\boldsymbol{u}) \cdot \pi(\boldsymbol{u})}{\pi_{\text{VA}}(\boldsymbol{u})}\right) d\boldsymbol{u} \end{aligned} \tag{A.1}$$

where $\pi(S_n|\boldsymbol{u})$ is probability density function of an IPP as in Equation (2.2) and $\pi(\boldsymbol{u})$ is the "prior" probability density function of $\boldsymbol{u}$ — we stated earlier that this follows a zero-mean multivariate normal distribution (i.e. $\boldsymbol{u} \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma}_{\text{prior}})$).

Next let $\mathbb{E}_{\pi_{\text{VA}}}[\cdot]$ denote the expectation with respect to the variational distribution $\pi_{\text{VA}}(\boldsymbol{u})$ and let $|\cdot|$ denote the matrix determinant. Also note that

$$\mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{v})^{\mathrm{T}} \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{v})\right] = \operatorname{Tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + (\boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{v})^{\mathrm{T}} \boldsymbol{A} (\boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{v})$$

for $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ and $\boldsymbol{v} \in \mathbb{R}^{\dim(\boldsymbol{x})}$, $\boldsymbol{A} \in \mathbb{R}^{\dim(\boldsymbol{x}) \times \dim(\boldsymbol{x})}$.

We now show derivation of the form solution to Equation A.1:

$$
\begin{aligned}
\underline{\ell}_{\mathrm{VA}}(\boldsymbol{\beta}) &= \mathbb{E}_{\pi_{\mathrm{VA}}}\left[\ln \pi\left(S_n|\boldsymbol{u}\right)\right] + \mathbb{E}_{\pi_{\mathrm{VA}}}\left[\ln \pi\left(\boldsymbol{u}\right)\right] - \mathbb{E}_{\pi_{\mathrm{VA}}}\left[\ln \pi_{\mathrm{VA}}\left(\boldsymbol{u}\right)\right] \\
&= \mathbb{E}_{\pi_{\mathrm{VA}}}\left[\ln \pi\left(S_n|\boldsymbol{u}\right)\right] \\
&\quad + \mathbb{E}_{\pi_{\mathrm{VA}}}\left[-\frac{k}{2}\ln\left(2\pi\right) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}| - \frac{1}{2}\left(\boldsymbol{u}-\mathbf{0}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\left(\boldsymbol{u}-\mathbf{0}\right)\right] \\
&\quad - \mathbb{E}_{\pi_{\mathrm{VA}}}\left[-\frac{k}{2}\ln\left(2\pi\right) + \frac{1}{2}\ln|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\left(\boldsymbol{u}-\boldsymbol{\mu}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{u}-\boldsymbol{\mu}\right)\right] \\
&= \mathbb{E}_{\pi_{\mathrm{VA}}}\left[\ln \pi\left(S_n|\boldsymbol{u}\right)\right] \\
&\quad \cancel{-\frac{k}{2}\ln\left(2\pi\right)} + \frac{1}{2}\ln|\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}| - \frac{1}{2}\mathbb{E}_{\pi_{\mathrm{VA}}}\left[\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{u}\right] \\
&\quad \cancel{+\frac{k}{2}\ln\left(2\pi\right)} - \frac{1}{2}\ln|\boldsymbol{\Sigma}^{-1}| + \frac{1}{2}\mathbb{E}_{\pi_{\mathrm{VA}}}\left[\left(\boldsymbol{u}-\boldsymbol{\mu}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{u}-\boldsymbol{\mu}\right)\right] \\
&= \mathbb{E}_{\pi_{\mathrm{VA}}}\left[\ln \pi\left(S_n|\boldsymbol{u}\right)\right] \\
&\quad + \frac{1}{2}\ln|\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\Sigma}\right) - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\mu} \\
&\quad - \frac{1}{2}\ln|\boldsymbol{\Sigma}^{-1}| + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) \\
&= \mathbb{E}_{\pi_{\mathrm{VA}}}\left[\ln \pi\left(S_n|\boldsymbol{u}\right)\right] \\
&\quad + \frac{1}{2}\ln|\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\Sigma}\right) - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\mu} \\
&\quad - \frac{1}{2}\ln|\boldsymbol{\Sigma}^{-1}| + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{I}\right) \\
&= \mathbb{E}_{\pi_{\mathrm{VA}}}\left[\ln \pi\left(S_n|\boldsymbol{u}\right)\right] \\
&\quad + \frac{1}{2}\ln|\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\Sigma}\right) - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\mu} - \frac{1}{2}\ln|\boldsymbol{\Sigma}^{-1}| + \frac{1}{2}k
\end{aligned}
$$

The first term in the above involves the expectation of an inhomogeneous Poisson process likelihood as in Equation (2.2) and is further approximated with numerical quadrature, as in Equation (2.3). *i.e.*

$$
\ln \pi\left(S_n|\boldsymbol{u}\right) \approx \left\{\sum_{i=1}^{n}\boldsymbol{X}\left(\boldsymbol{s}_i\right)\boldsymbol{\beta} + \boldsymbol{Z}\left(\boldsymbol{s}_i\right)\boldsymbol{u}\right\} - \left\{\sum_{i=n+1}^{n+q} w_i \exp\left\{\boldsymbol{X}\left(\boldsymbol{s}_i\right)\boldsymbol{\beta} + \boldsymbol{Z}\left(\boldsymbol{s}_i\right)\boldsymbol{u}\right\}\right\}
$$

Hence the VA likelihood becomes

$$\begin{aligned}
\underline{\ell}_{\mathrm{VA}}(\boldsymbol{\beta}) &= \mathbb{E}_{\pi_{\mathrm{VA}}}\left[\sum_{i=1}^{n}\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}+\boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{u}\right] \\
&\quad - \mathbb{E}_{\pi_{\mathrm{VA}}}\left[\sum_{i=n+1}^{n+q} w_i \exp\left\{\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}+\boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{u}\right\}\right] \\
&\quad + \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\right| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\Sigma}\right) - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\mu} - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}^{-1}\right| + \frac{1}{2}k \\
&= \left\{\sum_{i=1}^{n}\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}+\boldsymbol{Z}(\boldsymbol{s}_i)\,\mathbb{E}_{\pi_{\mathrm{VA}}}[\boldsymbol{u}]\right\} \\
&\quad - \left\{\sum_{i=n+1}^{n+q} w_i \exp\left\{\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}\right\}\mathbb{E}_{\pi_{\mathrm{VA}}}\left[\exp\left\{\boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{u}\right\}\right]\right\} \\
&\quad - \frac{1}{2}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\mu} - \ln\left|\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\Sigma}\right| - k\right]
\end{aligned}$$

Since $\boldsymbol{u}|S_n \overset{\pi_{\mathrm{VA}}}{\sim} \mathcal{N}_k(\boldsymbol{\mu},\boldsymbol{\Sigma})$, we have $\mathbb{E}_{\pi_{\mathrm{VA}}}[\boldsymbol{u}] = \boldsymbol{\mu}$ and, noting the moment generating function of a multivariate Gaussian vector, we have $\mathbb{E}_{\pi_{\mathrm{VA}}}\left[\exp\left\{\boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u}\right\}\right] = \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\Sigma}\boldsymbol{Z}(\boldsymbol{s})^{\mathrm{T}}$. So the VA likelihood (*i.e.* the variational approximation to the marginal log-likelihood) is given by

$$\begin{aligned}
\underline{\ell}_{\mathrm{VA}}(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}+\boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{\mu} \\
&\quad - \sum_{i=n+1}^{n+q} w_i \exp\left\{\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}+\boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{\Sigma}\boldsymbol{Z}(\boldsymbol{s}_i)^{\mathrm{T}}\right\} \\
&\quad - \frac{1}{2}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\mu} - \ln\left|\boldsymbol{\Sigma}_{\mathrm{prior}}^{-1}\boldsymbol{\Sigma}\right| - k\right] \qquad \text{(A.2)}
\end{aligned}$$

It can be noted here that the maximum (variational) likelihood estimate of the prior variance-covariance matrix, $\boldsymbol{\Sigma}_{\mathrm{prior}}$ (or parameters that comprise it, as explored in next paragraph), depends only on the variational parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Hence we can further simplify our likelihood by profiling. This will be true for all Gaussian VA for which the prior distribution on the marginalised random variables is zero-mean Gaussian.

A.1.1 Constrained VA likelihood

To further simplify the VA likelihood we can place constraints on variance-covariance matrices in Equation (A.2). During this thesis we variously tried unconstrained, spatially structured (*i.e.* we used a Gaussian kernel) and diagonal versions of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\text{prior}}$. We found little difference between model fits that permit correlation between the random coefficients and those that assume they are independent. Additionally, assuming independence (*i.e.* that $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\text{prior}}$ are diagonal) allows for great computational savings by avoiding the intense calculations involved in the log-determinants in Equation (A.2). So we further simplify our VA likelihood by constraining the prior distribution on the random coefficients to be $\boldsymbol{u} \sim \mathcal{N}_k\left(\mathbf{0}, \sigma^2_{\text{prior}}\boldsymbol{I}\right)$. As we are using the approximation of the latent Gaussian field, $\xi(\boldsymbol{s}) \approx \boldsymbol{Z}(\boldsymbol{s})\,\boldsymbol{u}$ this constraint is effectively assuming that all of the spatial correlation in ξ is derived from the basis functions, $\boldsymbol{Z}(\boldsymbol{s})$ (Cressie and Johannesson, 2008). Likewise, we assume that the random coefficients are independent within the variational density, but permit each a separate variance estimated from the data. *i.e.* $\boldsymbol{u}|S_n \overset{\pi_{\text{VA}}(\cdot)}{\sim} \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{I}\boldsymbol{\sigma}^2)$ where $\boldsymbol{\sigma}^2 = (\sigma^2_1, \ldots, \sigma^2_k)^{\mathrm{T}}$. Implementing these constraints, Equation (A.2) becomes

$$
\begin{aligned}
\underline{\ell}_{\text{VA}}(\boldsymbol{\beta}) = &\sum_{i=1}^{n} \boldsymbol{X}(\boldsymbol{s}_i)\,\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s}_i)\,\boldsymbol{\mu} \\
&- \sum_{i=n+1}^{n+q} w_i \exp\left\{\boldsymbol{X}(\boldsymbol{s}_i)\,\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s}_i)\,\boldsymbol{\mu} + \frac{1}{2}\sum_{r=1}^{k} Z_r(\boldsymbol{s}_i)^2\,\sigma^2_r\right\} \\
&- \frac{1}{2}\left[\sigma^{-2}_{\text{prior}}\sum_{r=1}^{k}\sigma^2_r + \sigma^{-2}_{\text{prior}}\sum_{r=1}^{k}\mu^2_r - \ln\left(\prod_{r=1}^{k}\sigma^{-2}_{\text{prior}}\sigma^2_r\right) - k\right] \\
= &\sum_{i=1}^{n} \boldsymbol{X}(\boldsymbol{s}_i)\,\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s}_i)\,\boldsymbol{\mu} \\
&- \sum_{i=n+1}^{n+q} w_i \exp\left\{\boldsymbol{X}(\boldsymbol{s}_i)\,\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s}_i)\,\boldsymbol{\mu} + \frac{1}{2}\sum_{r=1}^{k} Z_r(\boldsymbol{s}_i)^2\,\sigma^2_r\right\} \\
&- \frac{1}{2}\sigma^{-2}_{\text{prior}}\left[\sum_{r=1}^{k}\left(\sigma^2_r + \mu^2_r\right)\right] - \frac{k}{2}\ln\sigma^{-2}_{\text{prior}} - \frac{1}{2}\left[\sum_{r=1}^{k}\ln\sigma^2_r\right] + \frac{k}{2}
\end{aligned}
\tag{A.3}
$$

In Section 2.3.2 we describe the use of multiple resolutions of basis functions. The VA likelihood is similarly derived and simplified to include this type of basis function

configuration when we permit each resolution, l, to have a common variance $\sigma^2_{l;\text{prior}}$.

A.1.2 Profiled VA log-likelihood

The VA likelihood can be profiled with respect to the parameter(s) that describe the variance and correlation structure on the random coefficients in their "prior" distribution. The profiled maximum will be a function of the variational parameters, the structure of which will depend on how we have specified the prior and variational distributions. Here we derive the profile likelihood for the constrained version from the previous section, *i.e.* we profile Equation (A.3) with respect to the inverse of the "prior" variance, $\sigma^{-2}_{\text{prior}}$. Taking the derivative and looking for the maximum gives

$$\begin{aligned}\frac{\partial}{\partial\sigma^{-2}_{\text{prior}}}\underline{\ell}_{\text{VA}}(\boldsymbol{\beta}) &= \frac{\partial}{\partial\sigma^{-2}_{\text{prior}}}\left\{-\frac{1}{2}\sigma^{-2}_{\text{prior}}\left[\sum_{r=1}^{k}\left(\sigma_r^2+\mu_r^2\right)\right]+\frac{k}{2}\ln\left(\sigma^{-2}_{\text{prior}}\right)\right\}\\ &= -\frac{1}{2}\left[\sum_{r=1}^{k}\left(\sigma_r^2+\mu_r^2\right)\right]+\frac{1}{2}\frac{k}{\sigma^{-2}_{\text{prior}}}\end{aligned}$$

So that the maximum likelihood estimate is

$$\begin{aligned}\hat{\sigma}^{-2}_{\text{prior}} &= \left(\frac{1}{k}\sum_{r=1}^{k}\left(\sigma_r^2+\mu_r^2\right)\right)^{-1}\\ \hat{\sigma}^{2}_{\text{prior}} &= \frac{1}{k}\sum_{r=1}^{k}\left(\sigma_r^2+\mu_r^2\right)\end{aligned}$$

provided it exists ($\sigma_r^2 > 0, \forall r = 1,\ldots,k$ ensures this). We see it is a maximum by

$$\begin{aligned}\frac{\partial^2}{\partial\sigma^{-2}_{\text{prior}}\partial\sigma^{-2}_{\text{prior}}}\underline{\ell}_{\text{VA}}(\boldsymbol{\beta}) &= \frac{\partial}{\partial\sigma^{-2}_{\text{prior}}}\left\{\frac{k}{2}\left(\sigma^{-2}_{\text{prior}}\right)^{-1}\right\}\\ &= -k\left(\sigma^{-2}_{\text{prior}}\right)^{-2}\\ &< 0\end{aligned}$$

It can likewise be shown that if we have $l = 1,\ldots,n_{\text{res.}}$ resolutions of basis functions, each permitted a common prior variance $\sigma^2_{l;\text{prior}}$ then the maximum likelihood

estimates for each resolution are

$$\hat{\sigma}^2_{l;\text{prior}} = \frac{1}{k_l} \sum_{r_l=1}^{k_l} \left(\sigma^2_{r_l} + \mu^2_{r_l}\right)$$

where r_l takes values $r = k_{l-1} + 1, \ldots, k_l$.

Plugging the single resolution result back into Equation (A.3) gives us the simplified approximate log-likelihood

$$\begin{aligned}
\underline{\ell}_{\text{VA}}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{\mu} \\
&\quad - \sum_{i=n+1}^{n+q} w_i \exp\left\{\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{\mu} + \frac{1}{2}\sum_{r=1}^{k} Z_r(\boldsymbol{s}_i)^2 \sigma^2_r\right\} \\
&\quad - \frac{1}{2}\left[k \ln\left(\frac{1}{k}\sum_{r=1}^{k} \sigma^2_r + \mu^2_r\right) - \sum_{r=1}^{k} \ln \sigma^2_r\right] \\
&= \sum_{i=1}^{n} \boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{\mu} \\
&\quad - \sum_{i=n+1}^{n+q} w_i \exp\left\{\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{s}_i)\boldsymbol{\mu} + \frac{1}{2}\sum_{r=1}^{k} Z_r(\boldsymbol{s}_i)^2 \sigma^2_r\right\} \\
&\quad - \frac{1}{2}\left[\sum_{r=1}^{k} \ln \frac{\hat{\sigma}^2_{\text{prior}}}{\sigma^2_r}\right]
\end{aligned} \tag{A.4}$$

This is the one of the likelihood approximations we maximise in Chapter 2.

A.1.3 VA as a penalised likelihood

The variational approximation to the marginal log-likelihood (ℓ_{VA}) can be considered a penalised likelihood. We can re-write Equation (A.1) as

$$\begin{aligned}
\underline{\ell}_{\text{VA}}(\boldsymbol{\beta}) &= \int_{\boldsymbol{u}} \pi_{\text{VA}}(\boldsymbol{u}) \ln \pi(S_n|\boldsymbol{u}, \boldsymbol{\beta})\, d\boldsymbol{u} - \int_{\boldsymbol{u}} \pi_{\text{VA}}(\boldsymbol{u}) \ln\left[\frac{\pi_{\text{VA}}(\boldsymbol{u})}{\pi(\boldsymbol{u})}\right] d\boldsymbol{u} \\
&= \mathbb{E}_{\pi_{\text{VA}}}\left[\ln \pi(S_n|\boldsymbol{u}, \boldsymbol{\beta})\right] - D_{\text{KL}}\left[\pi_{\text{VA}}(\boldsymbol{u}) \,||\, \pi(\boldsymbol{u})\right]
\end{aligned} \tag{A.5}$$

The first term above is simply the conditional log-likelihood in Equation (2.2) evaluated taking the random effect expectation for the variational density. The second term in the above is the strictly non-negative, Kullback-Leibler divergence from the prior density ($\pi(\boldsymbol{u})$) to the variational density ($\pi_{\text{VA}}(\boldsymbol{u})$). So when we fit this objective function we are maximising a conditional mean likelihood (first term) penalised by how "far" the prior shifts to the variational density. Given the variational density is approximating the posterior density on the random effects this can be understood as a type of complexity penalty.

A.2 Complete Simulation Results

The following tables reveal the simulation result summaries for the various scenarios examined in Chapter 2. Models compared include an Inhomogeneous Poisson Process (IPP) and LGCP models fitted by Integrated Nested Laplace Approximations (INLA), as well as our proposed method using a variational approximation (VA) and Laplace approximation (Lp). The number of basis functions used in each model is given by k. For INLA, $k(n)$ means the number of mesh vertices provided as default from `INLA::inla.mesh.2d()` when a user supplies the point pattern of size n. For our methods (VA and Lp) Dual Res. means default number for two resolutions of k from `FRK::auto_basis()` was used. Other basis configurations include regular grids of 14×14, 10×10 or 7×7. Metrics include: root mean squared error in the estimated slope coefficient (RMSE β_1); average Kullback-Leibler divergence between the true and fitted log-intensity (KL Div. $\lambda||\hat{\lambda}$); average coverage probability for a 95% Wald confidence intervals for the slope parameter (95% Coverage Prob. $\hat{\beta}_1$); average width of 95% Wald confidence intervals for the slope parameter (CI β_1 Width); average marginal fitted log-likelihood ($\ell(\boldsymbol{\beta})$) and; average computation time in seconds (Comp. Time). Also included are the number of times the model failed to fit entirely (Fit Failure); the number of times the model failed to estimate standard errors ($\widehat{\text{SE}}$ Failure); and the number of times the optimiser converged to a nonsense result (Poor Convergence). Within each scenario we simulate 1000 small ($\mathbb{E}[N] = 200$), medium ($\mathbb{E}[N] = 500$) or large ($\mathbb{E}[N] = 1000$) point patterns.

Table A.1: Complete simulation result summaries for scenario S,S — *i.e.* both the covariate and latent field range of effect are approximately 30 units.

$\mathbb{E}[N]$	RMSE β_1	KL Div. $\lambda \Vert \hat{\lambda}$	95% Coverage Prob. $\hat{\beta}_1$	CI β_1 Width	$\ell(\boldsymbol{\beta})$	Comp. Time	Model	k	Fit Failure	SE Failure	Poor Convergence
200	0.44	11.26	0.87	0.68	-1138.31	954.52	INLA	$k(n)$	0	0	0
200	0.47	88.48	0.26	0.17	-1195.26	0.35	IPP	0	0	0	0
200	0.40	11.66	0.66	0.41	-1086.40	34.13	Lp	14x14	550	0	24
200	0.40	10.59	0.72	0.44	-1116.63	9.13	Lp	10x10	118	1	8
200	0.40	9.64	0.77	0.49	-1125.86	3.22	Lp	7x7	4	0	0
200	0.45	10.17	0.79	0.59	-1129.56	10.99	Lp	Dual Res.	5	1	1
200	0.95	11.71	0.62	0.38	-1138.28	5.22	VA	Dual Res.	0	76	0
200	3.07	21.01	0.52	0.39	-1143.55	16.58	VA	14x14	0	0	3
200	0.41	11.92	0.58	0.34	-1136.00	4.67	VA	10x10	0	0	0
200	0.40	10.26	0.69	0.42	-1131.47	1.55	VA	7x7	0	0	0
500	0.41	13.60	0.86	0.62	-2046.00	1302.15	INLA	$k(n)$	0	0	0
500	0.46	207.14	0.16	0.11	-2210.18	0.35	IPP	0	0	0	0
500	0.37	16.05	0.62	0.33	-2120.09	19.19	Lp	14x14	193	0	5
500	0.37	13.95	0.68	0.37	-2048.55	5.91	Lp	10x10	68	0	3
500	0.37	12.61	0.75	0.42	-2034.00	2.92	Lp	7x7	1	0	0
500	0.43	12.79	0.78	0.53	-2032.63	11.19	Lp	Dual Res.	1	0	0
500	0.40	14.39	0.58	0.32	-2065.78	4.96	VA	Dual Res.	0	62	0
500	0.39	26.10	0.44	0.23	-2064.73	15.02	VA	14x14	0	0	0
500	0.38	15.17	0.52	0.29	-2050.41	3.99	VA	10x10	0	0	0
500	0.37	13.02	0.67	0.37	-2040.94	1.29	VA	7x7	0	0	0
1000	0.39	16.21	0.86	0.59	-2907.89	2094.72	INLA	$k(n)$	0	0	0
1000	0.46	391.79	0.12	0.08	-3246.53	0.35	IPP	0	0	0	0
1000	0.35	19.29	0.59	0.31	-2925.06	15.47	Lp	14x14	50	0	6
1000	0.35	16.91	0.66	0.34	-2908.44	5.52	Lp	10x10	20	0	1
1000	0.35	15.32	0.73	0.38	-2894.79	2.79	Lp	7x7	0	0	0
1000	0.42	15.42	0.77	0.51	-2894.58	11.34	Lp	Dual Res.	2	2	0
1000	0.39	16.86	0.58	0.30	-2943.59	5.03	VA	Dual Res.	0	39	0
1000	0.37	38.70	0.40	0.20	-2946.26	14.74	VA	14x14	0	0	0
1000	0.36	17.79	0.51	0.26	-2919.17	3.90	VA	10x10	0	0	0
1000	0.35	15.60	0.66	0.34	-2904.49	1.25	VA	7x7	0	0	0

Table A.2: Complete simulation result summaries for scenario S,W — *i.e.* the range of effects is approximately 30 units for the covariate and approximately 5 units for the latent field.

$\mathbb{E}[N]$	RMSE β_1	KL Div. $\lambda \Vert \hat{\lambda}$	95% Coverage Prob. $\hat{\beta}_1$	CI β_1 Width	$\ell(\boldsymbol{\beta})$	Comp. Time	Model	k	Fit Failure	SE Failure	Poor Convergence
200	0.23	35.73	0.96	0.52	-1175.70	957.30	INLA	$k(n)$	0	0	0
200	0.29	143.38	0.41	0.15	-1248.64	0.38	IPP	0	0	0	0
200	0.28	36.20	0.97	0.61	-1153.33	27.13	Lp	14x14	577	1	20
200	0.27	41.65	0.99	0.66	-1171.95	8.09	Lp	10x10	68	2	6
200	0.33	53.83	0.98	0.71	-1178.31	2.92	Lp	7x7	1	0	0
200	0.29	44.71	0.99	0.71	-1170.79	11.05	Lp	Dual Res.	11	0	0
200	0.24	52.31	0.94	0.46	-1192.02	4.89	VA	Dual Res.	0	83	0
200	0.22	46.66	0.86	0.34	-1185.88	17.84	VA	14x14	0	0	0
200	0.23	49.41	0.93	0.42	-1185.74	4.75	VA	10x10	0	0	0
200	0.29	56.86	0.96	0.57	-1186.78	1.55	VA	7x7	0	0	0
500	0.20	50.78	0.98	0.52	-2122.03	1305.66	INLA	$k(n)$	0	0	0
500	0.28	356.94	0.27	0.10	-2370.41	0.33	IPP	0	0	0	0
500	0.22	55.30	1.00	0.61	-2124.84	13.47	Lp	14x14	241	0	10
500	0.29	67.87	0.99	0.77	-2128.19	5.54	Lp	10x10	24	0	2
500	0.50	107.30	0.96	0.79	-2158.62	2.76	Lp	7x7	1	0	0
500	0.43	75.16	0.99	0.88	-2134.32	10.07	Lp	Dual Res.	2	0	0
500	0.27	88.28	0.97	0.55	-2175.35	4.87	VA	Dual Res.	0	89	0
500	0.20	83.55	0.88	0.31	-2169.61	12.74	VA	14x14	0	0	0
500	0.22	81.77	0.96	0.46	-2162.58	3.16	VA	10x10	0	0	0
500	0.42	110.69	0.96	0.67	-2171.65	1.09	VA	7x7	0	0	0
1000	0.19	64.16	0.99	0.54	-3014.99	2253.57	INLA	$k(n)$	0	0	0
1000	0.28	697.32	0.18	0.07	-3565.48	0.42	IPP	0	0	0	0
1000	0.22	72.81	1.00	0.69	-3017.55	12.19	Lp	14x14	45	0	3
1000	0.36	96.39	1.00	0.90	-3039.98	5.19	Lp	10x10	2	0	1
1000	0.91	184.60	0.85	0.87	-3110.61	2.65	Lp	7x7	0	0	0
1000	0.69	110.38	0.97	1.13	-3053.04	9.77	Lp	Dual Res.	0	0	0
1000	0.42	124.71	0.97	0.72	-3103.20	4.31	VA	Dual Res.	0	6	0
1000	0.18	115.30	0.90	0.32	-3106.05	13.06	VA	14x14	0	0	0
1000	0.25	113.56	0.99	0.58	-3093.61	3.21	VA	10x10	0	0	0
1000	0.81	187.90	0.86	0.78	-3128.18	1.15	VA	7x7	0	0	0

Table A.3: Complete simulation result summaries for scenario W,S — *i.e.* the range of effects is approximately 5 units for the covariate and approximately 30 units for the latent field.

$\mathbb{E}[N]$	RMSE β_1	KL Div. $\lambda \lvert\rvert \hat{\lambda}$	95% Coverage Prob. $\hat{\beta}_1$	CI β_1 Width	$\ell(\boldsymbol{\beta})$	Comp. Time	Model	k	Fit Failure	SE Failure	Poor Convergence
200	0.23	35.73	0.96	0.52	-1175.70	957.30	INLA	$k(n)$	0	0	0
200	0.29	143.38	0.41	0.15	-1248.64	0.38	IPP	0	0	0	0
200	0.28	36.20	0.97	0.61	-1153.33	27.13	Lp	14x14	577	1	20
200	0.27	41.65	0.99	0.66	-1171.95	8.09	Lp	10x10	68	2	6
200	0.33	53.83	0.98	0.71	-1178.31	2.92	Lp	7x7	1	0	0
200	0.29	44.71	0.99	0.71	-1170.79	11.05	Lp	Dual Res.	11	0	0
200	0.24	52.31	0.94	0.46	-1192.02	4.89	VA	Dual Res.	0	83	0
200	0.22	46.66	0.86	0.34	-1185.88	17.84	VA	14x14	0	0	0
200	0.23	49.41	0.93	0.42	-1185.74	4.75	VA	10x10	0	0	0
200	0.29	56.86	0.96	0.57	-1186.78	1.55	VA	7x7	0	0	0
500	0.20	50.78	0.98	0.52	-2122.03	1305.66	INLA	$k(n)$	0	0	0
500	0.28	356.94	0.27	0.10	-2370.41	0.33	IPP	0	0	0	0
500	0.22	55.30	1.00	0.61	-2124.84	13.47	Lp	14x14	241	0	10
500	0.29	67.87	0.99	0.77	-2128.19	5.54	Lp	10x10	24	0	2
500	0.50	107.30	0.96	0.79	-2158.62	2.76	Lp	7x7	1	0	0
500	0.43	75.16	0.99	0.88	-2134.32	10.07	Lp	Dual Res.	2	0	0
500	0.27	88.28	0.97	0.55	-2175.35	4.87	VA	Dual Res.	0	89	0
500	0.20	83.55	0.88	0.31	-2169.61	12.74	VA	14x14	0	0	0
500	0.22	81.77	0.96	0.46	-2162.58	3.16	VA	10x10	0	0	0
500	0.42	110.69	0.96	0.67	-2171.65	1.09	VA	7x7	0	0	0
1000	0.19	64.16	0.99	0.54	-3014.99	2253.57	INLA	$k(n)$	0	0	0
1000	0.28	697.32	0.18	0.07	-3565.48	0.42	IPP	0	0	0	0
1000	0.22	72.81	1.00	0.69	-3017.55	12.19	Lp	14x14	45	0	3
1000	0.36	96.39	1.00	0.90	-3039.98	5.19	Lp	10x10	2	0	1
1000	0.91	184.60	0.85	0.87	-3110.61	2.65	Lp	7x7	0	0	0
1000	0.69	110.38	0.97	1.13	-3053.04	9.77	Lp	Dual Res.	0	0	0
1000	0.42	124.71	0.97	0.72	-3103.20	4.31	VA	Dual Res.	0	6	0
1000	0.18	115.30	0.90	0.32	-3106.05	13.06	VA	14x14	0	0	0
1000	0.25	113.56	0.99	0.58	-3093.61	3.21	VA	10x10	0	0	0
1000	0.81	187.90	0.86	0.78	-3128.18	1.15	VA	7x7	0	0	0

Table A.4: Complete simulation result summaries for scenario W,W — *i.e.* both the covariate and latent field range of effect are approximately 5 units.

$\mathbb{E}[N]$	RMSE β_1	KL Div. $\lambda \lvert\rvert \hat{\lambda}$	95% Coverage Prob. $\hat{\beta}_1$	CI β_1 Width	$\ell(\boldsymbol{\beta})$	Comp. Time	Model	k	Fit Failure	SE Failure	Poor Convergence
200	0.23	35.73	0.96	0.52	-1175.70	957.30	INLA	$k(n)$	0	0	0
200	0.29	143.38	0.41	0.15	-1248.64	0.38	IPP	0	0	0	0
200	0.28	36.20	0.97	0.61	-1153.33	27.13	Lp	14x14	577	1	20
200	0.27	41.65	0.99	0.66	-1171.95	8.09	Lp	10x10	68	2	6
200	0.33	53.83	0.98	0.71	-1178.31	2.92	Lp	7x7	1	0	0
200	0.29	44.71	0.99	0.71	-1170.79	11.05	Lp	Dual Res.	11	0	0
200	0.24	52.31	0.94	0.46	-1192.02	4.89	VA	Dual Res.	0	83	0
200	0.22	46.66	0.86	0.34	-1185.88	17.84	VA	14x14	0	0	0
200	0.23	49.41	0.93	0.42	-1185.74	4.75	VA	10x10	0	0	0
200	0.29	56.86	0.96	0.57	-1186.78	1.55	VA	7x7	0	0	0
500	0.20	50.78	0.98	0.52	-2122.03	1305.66	INLA	$k(n)$	0	0	0
500	0.28	356.94	0.27	0.10	-2370.41	0.33	IPP	0	0	0	0
500	0.22	55.30	1.00	0.61	-2124.84	13.47	Lp	14x14	241	0	10
500	0.29	67.87	0.99	0.77	-2128.19	5.54	Lp	10x10	24	0	2
500	0.50	107.30	0.96	0.79	-2158.62	2.76	Lp	7x7	1	0	0
500	0.43	75.16	0.99	0.88	-2134.32	10.07	Lp	Dual Res.	2	0	0
500	0.27	88.28	0.97	0.55	-2175.35	4.87	VA	Dual Res.	0	89	0
500	0.20	83.55	0.88	0.31	-2169.61	12.74	VA	14x14	0	0	0
500	0.22	81.77	0.96	0.46	-2162.58	3.16	VA	10x10	0	0	0
500	0.42	110.69	0.96	0.67	-2171.65	1.09	VA	7x7	0	0	0
1000	0.19	64.16	0.99	0.54	-3014.99	2253.57	INLA	$k(n)$	0	0	0
1000	0.28	697.32	0.18	0.07	-3565.48	0.42	IPP	0	0	0	0
1000	0.22	72.81	1.00	0.69	-3017.55	12.19	Lp	14x14	45	0	3
1000	0.36	96.39	1.00	0.90	-3039.98	5.19	Lp	10x10	2	0	1
1000	0.91	184.60	0.85	0.87	-3110.61	2.65	Lp	7x7	0	0	0
1000	0.69	110.38	0.97	1.13	-3053.04	9.77	Lp	Dual Res.	0	0	0
1000	0.42	124.71	0.97	0.72	-3103.20	4.31	VA	Dual Res.	0	6	0
1000	0.18	115.30	0.90	0.32	-3106.05	13.06	VA	14x14	0	0	0
1000	0.25	113.56	0.99	0.58	-3093.61	3.21	VA	10x10	0	0	0
1000	0.81	187.90	0.86	0.78	-3128.18	1.15	VA	7x7	0	0	0

Appendix B

Additional results for Chapter 3

B.1 Complete Flora Data Integration Results

Table B.1: Complete results for *Corymbia eximia* four-fold cross-validation using the integrated data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data; Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data; ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data.

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-2155.22	4328.44	-1306.65	-1048.36	0.88
2	-2028.18	4078.36	-1240.96	-991.13	0.90
6	-1960.60	3943.19	-1303.72	-1093.57	0.89
12	-1890.15	3802.30	-1352.83	-1177.04	0.89
15	-1924.14	3870.28	-1357.74	-1222.41	0.88
24	-1873.26	3768.51	-1587.35	-1308.37	0.90
35	-1846.83	3715.66	-1480.54	-1338.50	0.90
48	-1849.54	3721.08	-1340.71	-1208.93	0.89
54	-1840.98	3703.95	-1277.54	-1209.97	0.89
70	-1794.33	3610.67	-1223.38	-1039.22	0.91
88	-1805.44	3632.88	-1342.30	-1248.17	0.90
96	-1777.46	3576.93	-1226.73	-1173.01	0.90
117	-1768.56	3559.13	-1286.00	-1098.06	0.91
140	-1792.14	3606.28	-1208.79	-1161.89	0.89
165	-1747.60	3517.20	-1179.88	-1010.90	0.92
176	-1776.26	3574.53	-1402.86	-1420.52	0.90
204	-1742.25	3506.50	-1133.31	-932.10	0.93
234	-1758.19	3538.39	-2007.53	-1773.44	0.92

Table B.2: Complete results for *Corymbia eximia* four-fold cross-validation using the presence-only data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data; Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data (missing here as this is a presence-only data model); ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data (missing here as this is a presence-only data model). Other missing values indicate a failed model convergence.

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-1180.29	2376.59	-1294.34		
2	-1098.36	2216.71	-1228.23		
6	-1045.02	2110.05	-1266.24		
12	-1020.77	2061.54	-1400.43		
15	-1017.70	2055.41	-1319.50		
24	-1024.86	2069.72	-1896.53		
35	-1007.70	2035.40	-1416.89		
48	-993.39	2006.78	-1365.74		
54	-982.12	1984.24	-2038.20		
70	-967.12	1954.23	-1313.68		
88	-961.10	1942.21	-1736.50		
96	-965.82	1951.65	-1287.06		
117	-948.36	1916.73			
140	-957.72	1935.43	-1239.64		
165	-943.35	1906.70	-1430.98		
176	-954.96	1929.91	-1212.58		
204	-939.13	1898.26	-1292.13		
234	-951.60	1923.20	-1231.91		

Table B.3: Complete results for *Corymbia eximia* four-fold cross-validation using the presence/absence data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data (missing here as this is a presence/absence data model); Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data; ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data.

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-939.67	1891.34		-1034.51	0.88
2	-900.46	1816.91		-973.30	0.90
6	-862.63	1741.25		-1006.98	0.90
12	-843.08	1702.16		-1086.34	0.89
15	-851.29	1718.58		-1121.15	0.88
24	-824.63	1665.27		-1046.01	0.91
35	-823.17	1662.35		-1148.30	0.89
48	-826.32	1668.65		-1093.08	0.89
54	-831.29	1678.59		-1071.32	0.89
70	-803.97	1623.95		-1078.15	0.90
88	-814.08	1644.16		-1110.31	0.88
96	-797.98	1611.97		-1094.09	0.89
117	-792.17	1600.35		-1072.02	0.90
140	-805.78	1627.56		-1146.18	0.87
165	-785.05	1586.11		-1048.49	0.91
176	-801.54	1619.08		-1161.83	0.88
204	-785.11	1586.22		-966.68	0.92
234	-794.93	1605.85		-1207.22	0.89

Table B.4: Complete results for *Eucalyptus canaliculata* four-fold cross-validation using the integrated data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data; Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data; ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data.

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-613.40	1244.79	-286.19	-399.19	0.83
2	-513.61	1049.21	-286.36	-317.53	0.95
6	-486.88	995.76	-287.52	-268.98	0.97
12	-485.21	992.42	-294.07	-268.99	0.97
15	-477.54	977.07	-290.46	-260.33	0.97
24	-484.44	990.89	-294.81	-274.15	0.97
35	-480.53	983.06	-304.68	-270.00	0.96

Table B.5: Complete results for *Eucalyptus canaliculata* four-fold cross-validation using the presence-only data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data; Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data (missing here as this is a presence-only data model); ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data (missing here as this is a presence-only data model).

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-246.43	508.85	-284.14		
2	-244.93	509.87	-285.19		
6	-243.73	507.46	-285.46		
12	-242.54	505.09	-286.82		
15	-242.37	504.73	-281.62		
24	-242.82	505.65	-287.14		
35	-243.31	506.62	-287.02		

Table B.6: Complete results for *Eucalyptus canaliculata* four-fold cross-validation using the presence/absence data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data (missing here as this is a presence/absence data model); Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data; ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data.

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-350.36	712.71		-395.20	0.83
2	-264.88	545.75		-315.64	0.95
6	-238.45	492.90		-257.40	0.97
12	-236.00	488.01		-255.94	0.97
15	-231.00	478.01		-245.03	0.97
24	-235.73	487.47		-255.88	0.97
35	-232.65	481.30		-244.87	0.98

Table B.7: Complete results for *Homoranthus cernuus* four-fold cross-validation using the integrated data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data; Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data; ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data.

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-171.27	354.55	-114.24	-67.26	0.76
2	-166.32	348.65	-114.25	-67.26	0.76
6	-118.75	253.51	-109.49	-138.95	0.99

Table B.8: Complete results for *Homoranthus cernuus* four-fold cross-validation using the presence-only data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data; Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data (missing here as this is a presence-only data model); ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data (missing here as this is a presence-only data model).

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-105.47	220.93	-113.59		
2	-105.47	224.93	-113.59		
6	-81.95	177.89	-117.45		

Table B.9: Complete results for *Homoranthus cernuus* four-fold cross-validation using the presence/absence data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data (missing here as this is a presence/absence data model); Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data; ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data.

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-64.25	134.51		-67.17	0.79
2	-64.25	138.51		-67.17	0.79
6	-48.48	106.97		-67.17	0.79

Table B.10: Complete results for *Eucalyptus sparsifolia* four-fold cross-validation using the integrated data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data; Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data; ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data.

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-3287.19	6592.37	-1371.05	-2080.10	0.68
2	-3024.11	6070.23	-1300.78	-1885.59	0.78
6	-2924.95	5871.89	-1244.37	-1915.32	0.79
12	-2910.80	5843.60	-1261.20	-1969.05	0.77
15	-2923.98	5869.97	-1252.25	-1970.55	0.77
24	-2898.65	5819.31	-1288.16	-2109.32	0.75
35	-2865.31	5752.63	-1324.85	-2340.37	0.74
48	-2833.94	5689.89	-1247.91	-2205.59	0.75
54	-2848.23	5718.45	-1286.02	-2231.36	0.75
70	-2764.69	5551.37	-1238.60	-2017.35	0.79
88	-2773.32	5568.65	-1238.66	-2186.15	0.76
96	-2762.68	5547.37	-1209.31	-2208.84	0.77
117	-2747.88	5517.77	-1234.14	-2195.03	0.77
140	-2742.83	5507.66	-1182.50	-2301.67	0.77
165	-2744.59	5511.17	-1183.71	-2196.07	0.77
176	-2735.73	5493.46	-1207.38	-2744.07	0.76

Table B.11: Complete results for *Eucalyptus sparsifolia* four-fold cross-validation using the presence-only data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data; Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data (missing here as this is a presence-only data model); ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data (missing here as this is a presence-only data model).

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\text{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\text{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-1225.73	2467.47	-1325.60		
2	-1140.11	2300.23	-1255.17		
6	-1067.91	2155.82	-1162.17		
12	-1069.75	2159.50	-1226.07		
15	-1058.52	2137.05	-1160.93		
24	-1071.44	2162.89	-1226.07		
35	-1054.22	2128.44	-1247.65		
48	-1046.22	2112.43	-1261.42		
54	-1055.38	2130.76	-1301.34		
70	-1037.04	2094.08	-1346.35		
88	-1026.27	2072.53	-1245.69		
96	-1018.75	2057.49	-1317.13		
117	-1016.25	2052.50	-1291.51		
140	-1016.28	2052.56	-1310.76		
165	-1007.42	2034.85	-1328.25		
176	-1002.87	2025.74	-1314.25		

Table B.12: Complete results for *Eucalyptus sparsifolia* four-fold cross-validation using the presence/absence data model. k is the number of basis functions; $\underline{\ell}(\boldsymbol{\beta})$ is the fitted likelihood; AIC is Akaike's Information Criteria; Predicted $\ell_{\mathrm{PO}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence-only data (missing here as this is a presence/absence data model); Predicted $\ell_{\mathrm{PA}}(\boldsymbol{\beta}|\boldsymbol{\xi})$ is the predicted, conditional log-likelihood on the presence/absence data; ROC AUC is the area under the receiver operator characteristic curve for the presence/absence data.

k	$\underline{\ell}(\boldsymbol{\beta})$	AIC	Predicted $\ell_{\mathrm{PO}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	Predicted $\ell_{\mathrm{PA}}(\boldsymbol{\beta}\|\boldsymbol{\xi})$	ROC AUC
0	-1969.38	3950.76		-2054.08	0.70
2	-1774.37	3564.74		-1843.22	0.80
6	-1736.32	3488.63		-1849.90	0.81
12	-1725.79	3467.58		-1864.59	0.80
15	-1733.93	3483.86		-1877.86	0.80
24	-1718.88	3453.76		-2006.30	0.77
35	-1713.49	3442.98		-2209.16	0.74
48	-1695.83	3407.67		-2144.51	0.75
54	-1693.51	3403.02		-2322.53	0.75
70	-1641.77	3299.54		-1947.67	0.80
88	-1659.63	3335.26		-2154.08	0.77
96	-1656.60	3329.20		-2063.38	0.78
117	-1651.30	3318.59		-2131.38	0.77
140	-1646.34	3308.69		-2020.28	0.80
165	-1649.25	3314.50		-2201.26	0.77
176	-1649.15	3314.29		-2339.44	0.77

Appendix C

Additional Details for Chapter 4

C.1 Data Requirements

Data involved in modelling an IPP or LGCP against predictors often comprises the records of presence locations in the form of a list of coordinates within the domain of interest, as well as geo-referenced grids of the predictors covering the same domain. Here we illustrate how to convert this into a single data frame required by the `scampr` model function using the example of the `gorillas` dataset provided within the package `inlabru`.

C.1.1 Interpolation of Predictors

First, any predictors to be included in a model must be available at the presence locations. For this, interpolation from the geo-referenced grid of a predictor to the presence points is required. How this is best done will vary depending on the precise form of the predictors available, however in the gorilla nesting example, `over()` from the package `sp` (Pebesma and Bivand, 2005) provides a suitable interpolation function for this:

```
require(inlabru)
data(gorillas, package = "inlabru")
    # coordinate names req. for consistency
coord.names <- c('x', 'y')
```

```
    # interpolation function
library(sp)
f.elevation = function(x,y) {
    # turn coordinates into SpatialPoints object:
  spp = SpatialPoints(data.frame(x=x,y=y))
    # attach the appropriate coordinate reference system (CRS)
  proj4string(spp) = CRS(proj4string(gorillas$gcov$elevation))
    # extract values at spp coords, from SpatialGridDataFrame
  v = over(spp, gorillas$gcov$elevation)
  return(v$elevation)
}
pres.locs <- as.data.frame(gorillas$nests@coords)
colnames(pres.locs) <- coord.names
pres.locs$elevation <- f.elevation(pres.locs$x, pres.locs$y)
```

Another option for interpolation is the `interp_im()` function in the `spatstat` package (Turner and Baddeley, 2005). This function requires the predictor in raster format, which `spatstat` refers to as an `image`.

C.1.2 Quadrature (or Background) Points

The second data requirement is to specify quadrature points that serve to approximate the spatial integral within to point process models, *i.e.* Equation (2.3). As numerical quadrature is the most common approach to approximating the spatial integral, the data frame provided to `scampr` point process models must include quadrature points and their corresponding sizes (or areas), in addition to the point events themselves. Advice on how to choose the number and location of quadrature points is well covered by Renner et al. (2015, p. 370). Like any numerical approximation to an integral, the more quadrature points, the more accurate the integral approximation. So if computational constraints are not a limitation, it would be advisable to simply use as quadrature points the entire geo-referenced predictor grid at the finest available resolution. This however is often impractical, and Renner et al.

(2015) suggest two ways to reduce the number of quadrature points: randomly sampling as many as are needed to produce a stable approximate likelihood; or using quadrats that are not all equal in size, with larger quadrats in areas with less presence points (hence presumably lower predicted intensity). The use of rank-reduction in `scampr` (Section 2.3.2) mitigates much of the computational burden of selecting many quadrature points, because computational time then scales linearly with $n+q$. It should also be noted that in point process models, quadrature points are *NOT* assumed to be absences (despite sometimes being referred to as pseudo-absences) and having a quadrature point at a presence location is not a concern.

The `spatstat` package contains the function `quad.scheme()`, used to select quadrature points and assign them appropriate sizes. However, this introduces a range of uniquely classed objects which forces the package's coding structures upon the user — something our package is trying to minimise. Hence, we will manually set up a quadrature scheme for the `gorillas` dataset, and store it, together with relevant predictors, in an object called `gorillas.df`.

```
    # get the coordinates, as a 2 column data frame
quad <- data.frame(gorillas$gcov$elevation@coords)
colnames(quad) <- coord.names
    # add elevation data
quad$elevation <- gorillas$gcov$elevation@data$elevation
    # regular grid provides dist. between quad pts.
dx <- min(diff(unique(quad$x)))
dy <- min(diff(unique(quad$y)))

    # Calculate quadrature sizes #
    # rectangles centered at the quadrature
quad$quad.size <- dx * dy

    # The boundary of the domain is provided
bnd <- data.frame(
    gorillas$boundary@polygons[[1]]@Polygons[[1]]@coords)
```

```
colnames(bnd) <- coord.names

    # Use mgcv::in.out() to get interior quad
quad.in.bnd <- mgcv::in.out(bnd = as.matrix(bnd),
    x = as.matrix(quad[ , c("x", "y")]))
quad <- quad[quad.in.bnd, ]

    # Combine the presence locations and quad into a data frame
    # with a binary identifier 'pres'
gorillas.df <- rbind(
    cbind(pres.locs, quad.size = 0, pres = 1),
    cbind(quad[quad.in.bnd, ], pres = 0)
)
head(gorillas.df)
         x        y elevation quad.size pres
1 582.5184 676.8862      2008         0    1
2 581.8230 677.4227      1699         0    1
3 582.1310 676.9379      1872         0    1
4 582.1119 677.4200      1678         0    1
5 582.5851 677.5097      1658         0    1
6 582.3023 677.5216      1655         0    1
```

This forms the template for what is required of a data frame to be used in `scampr` point process models. Columns should include: coordinates, quadrature size, a binary identifier for presence/quadrature points, and any predictors to be included in the `formula` for the model. Note that quadrature points are denoted as `pres == 0`, but this does not mean that they should be interpreted as absences. Rather, they are quadrats with which we approximate the spatial integral (see, *e.g.* Berman and Turner, 1992). Hence these rows also require a non-zero size/area attached to them (*i.e.* `quad.size` in the above), otherwise these rows of the dataset would be redundant. Presence points (`pres == 1`) on the other hand can have a size of zero — this indicates that these points are not used to estimate the spatial integral, but they

will still contribute to other parts of the likelihood. Not described in the above are edge effects (or border corrections) for the quadrature points. In the gorilla nesting data example we are assuming that all quadrats have equal sizes/areas (`dx * dy`), however along the boundaries of the domain these may differ. See documentation for packages `spatstat` (Turner and Baddeley, 2005) and `sp` (Pebesma and Bivand, 2005) for ways to address this.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, pages 267–281, Budapest: Akademiai Kiado.

Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019). `inlabru`: an `R` package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6):760–766.

Baddeley, A., Chang, Y.-M., Song, Y., and Turner, R. (2013). Residual diagnostics for covariate effects in spatial point process models. *Journal of Computational and Graphical Statistics*, 22(4):886–905.

Baddeley, A., Rubak, E., Møller, J., et al. (2011). Score, pseudo-score and residual diagnostics for spatial point process models. *Statistical Science*, 26(4):613–646.

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns: (with discussion). *Australian & New Zealand Journal of Statistics*, 42(3):283–322.

Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):617–666.

Baddeley, A. J., Møller, J., and Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54(3):329–350.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society*, 70(4):825–848.

Berman, M. and Turner, T. R. (1992). Approximating point process likelihoods with GLIM. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):31–38.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98.

Bickel, P., Choi, D., Chang, X., Zhang, H., et al. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943.

Botella, C., Joly, A., Bonnet, P., Munoz, F., and Monestiez, P. (2021). Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. *Methods in Ecology and Evolution*, 12(5):933–945.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912.

Brix, A. and Diggle, P. J. (2001). Spatio-temporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63:823–841.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., and Bolker, B. M. (2017). `glmmTMB` balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., and Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 60(5):757–776.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical models in S*. CRC.

Chandler, M., See, L., Copas, K., Bonde, A. M., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., et al. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, 213:280–294.

Conn, P. B., Thorson, J. T., and Johnson, D. S. (2017). Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution*, 8(11):1535–1546.

Cooper, F., Pi, S.-Y., and Stancioff, P. N. (1986). Quantum dynamics in a time-dependent variational approximation. *Physical Review D*, 34(12):3831.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(1):209–226.

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

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: Volume II: General theory and structure*. Springer Science & Business Media.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Davis, P. J. and Rabinowitz, P. (2007). *Methods of numerical integration*. Courier Corporation.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563.

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, 23(12):1472–1484.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.

Elith, J. and 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.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368.

Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438.

Fithian, W. and Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, 7(4):1917.

Fletcher Jr, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., and Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, 100(6):e02710.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., and Sibert, J. (2012). AD model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2):233–249.

Funwi-Gabga, N. and Mateu, J. (2012). Understanding the nesting spatial behaviour of gorillas in the kagwene sanctuary, cameroon. *Stochastic Environmental Research and Risk Assessment*, 26(6):793–811.

Golding, N. (2019). `greta`: simple and scalable statistical modelling in `R`. *Journal of Open Source Software*, 4(40):1601.

Griewank, A. (1989). On automatic differentiation. *Mathematical Programming: Recent Developments and Applications*, 6(6):83–107.

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography*, 40(2):281–295.

Hager, T. and Benson, D. (2010). The eucalypts of the Greater Blue Mountains World Heritage Area: distribution, classification and habitats of the species of Eucalyptus, Angophora and Corymbia (family Myrtaceae) recorded in its eight conservation reserves. *Cunninghamia*, 11:425–444.

Hall, P., Ormerod, J. T., and Wand, M. P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica*, 21(1):369–389.

Hartig, F. (2017). Package 'dharma'.

Hastie, T. and Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography*, 36(8):864–867.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgaren, F., Nychka, D. A., Sun, F., and Zammit-Mangion, A. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425.

Hefley, T. J., Baasch, D. M., Tyre, A. J., and Blankenship, E. E. (2014). Correction of location errors for presence-only species distribution models. *Methods in Ecology and Evolution*, 5(3):207–214.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334.

Hogg, S., Wang, Y., and Stone, L. (2021). Effectiveness of joint species distribution models in the presence of imperfect detection. *Methods in Ecology and Evolution*.

Hui, F. K. C., You, C., Shang, H. L., and Müller, S. (2019). Semiparametric regression using variational approximations. *Journal of the American Statistical Association*, 114(528):1765–1777.

Illian, J. B., Sørbye, S. H., and Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics*, 6(4):1499–1530.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., and Stone, L. (2017). Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, 8(4):420–430.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). `TMB`: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 73(4):423–498.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233.

McCullagh, P. and Nelder, J. A. (2019). *Generalized Linear Models*. Routledge.

McInerny, G. J. and Purves, D. W. (2011). Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, 2(3):248–257.

Miller, D. A., Pacifici, K., Sanderlin, J. S., and Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10(1):22–37.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Myllymäki, M. and Mrkvička, T. (2019). `GET`: Global envelopes in `R`. *arXiv preprint arXiv:1911.06583*.

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):381–404.

Niku, J., Hui, F. K., Taskinen, S., and Warton, D. I. (2019). `gllvm`: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, 10(12):2173–2182.

Nychka, D., Furrer, R., Paige, J., and Sain, S. (2017). `fields`: Tools for spatial data. R package version 10.3.

Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4):315–331.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27.

Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792.

Ormerod, J. T. and Wand, M. (2010). Explaining variational approximations. *The American Statistician*, 64(2):140–153.

Ovaskainen, O. and Soininen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, 92(2):289–295.

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., and Collazo, J. A. (2017). Integrating multiple data sources in

species distribution modeling: a framework for data fusion. *Ecology*, 98(3):840–850.

Parmesan, C. (2006). Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology, Evolution, and Systematics*, 37:637–669.

Pearce, J. L. and Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, 43(3):405–412.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2):9–13.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259.

Phillips, S. J. and Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, 31(2):161–175.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., and McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, 5(5):397–406.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379.

Renner, I. W., Louvrier, J., and Gimenez, O. (2019). Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalized likelihood maximization. *Methods in Ecology and Evolution*, 10(12):2118–2128.

Renner, I. W. and Warton, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, 69(1):274–281.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society*, 71(2):319–392.

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., and Strokorb, K. (2015). Analysis, simulation and prediction of multivariate random fields with package `RandomFields`. *Journal of Statistical Software*, 63(8):1–25.

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656.

Shirota, S. and Gelfand, A. E. (2017a). Approximate Bayesian computation and model assessment for repulsive spatial point processes. *Journal of Computational and Graphical Statistics*, 26(3):646–657.

Shirota, S. and Gelfand, A. E. (2017b). Space and circular time log Gaussian Cox processes with application to crime event data. *The Annals of Applied Statistics*, 11(2):481–503.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):749–760.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., and O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, 43(10):1413–1422.

Simpson, D., Illian, J., Lindgren, F., Sørbye, S. H., and Rue, H. (2016). Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1):49–70.

Stoklosa, J., Daly, C., Foster, S. D., Ashcroft, M. B., and Warton, D. I. (2015). A climate of uncertainty: accounting for error in climate variables for species distribution models. *Methods in Ecology and Evolution*, 6(4):412–423.

Taylor, B., Davies, T., Rowlingson, B., and Diggle, P. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in `R`. *Journal of Statistical Software*, 63:1–48.

Taylor, B. M., Davies, T. M., Rowlingson, B. S., and Diggle, P. J. (2013). `lgcp`: An `R` package for inference with spatial and spatio-temporal log-Gaussian Cox processes. *Journal of Statistical Software*, 52(4):1–40.

Taylor, B. M. and Diggle, P. J. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, 84(10):2266–2284.

Thorson, J. T. (2019). Guidance for decisions using the vector autoregressive spatio-temporal (vast) package in stock, ecosystem, habitat and climate assessments. *Fisheries Research*, 210:143–161.

Thorson, J. T., Ianelli, J. N., Larsen, E. A., Ries, L., Scheuerell, M. D., Szuwalski, C., and Zipkin, E. F. (2016). Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, 25(9):1144–1158.

Thorson, J. T., Shelton, A. O., Ward, E. J., and Skaug, H. J. (2015). Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for west coast groundfishes. *ICES Journal of Marine Science*, 72(5):1297–1310.

Tobler, M. W., Kéry, M., Hui, F. K., Guillera-Arroita, G., Knaus, P., and Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, 100(8):e02754.

Turner, R. and Baddeley, A. (2005). `SPATSTAT`: an `R` package for analyzing spatial point patterns. *Journal of Statistical Software*, 12.

Tzeng, S. and Huang, H.-C. (2018). Resolution adaptive fixed rank kriging. *Technometrics*, 60(2):198–208.

Waagepetersen, R. (2004). Convergence of posteriors for discretized log Gaussian Cox processes. *Statistics & Probability Letters*, 66(3):229–235.

Wang, Y. and Blei, D. M. (2019). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Warton, D. I., Renner, I. W., and Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS one*, 8(11):e79168.

Warton, D. I. and Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, 4(3):1383–1402.

Watson, J., Joy, R., Tollit, D., Thornton, S. J., and Auger-Méthé, M. (2019). Estimating animal utilization distributions from multiple data types: a joint spatio-temporal point process framework. *arXiv preprint arXiv:1911.00151*.

Wickham, H. (2019). *Advanced r*. chapman and hall/CRC.

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., and McCarthy, M. A. (2019). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, 10(2):198–211.

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., and McCarthy, M. A. (2021). Defining and evaluating predictions of joint species distribution models. *Methods in Ecology and Evolution*, 12(3):394–404.

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J.-A., Guisan, A., et al. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews*, 88(1):15–30.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, 80(4):791–795.

Zammit-Mangion, A. and Cressie, N. (2017). FRK: An R package for spatial and spatio-temporal prediction with large datasets. *arXiv preprint arXiv:1705.08105*.

Zammit-Mangion, A., Dewar, M., Kadirkamanathan, V., and Sanguinetti, G. (2012). Point process modelling of the Afghan war diary. *Proceedings of the National Academy of Sciences*, 109(31):12414–12419.