Accounting for Preferential Sampling in the Statistical Analysis of Spatio-temporal Data

by

Joe Watson

MMath, University of Bath, 2016

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Statistics)

The University of British Columbia
(Vancouver)

December 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Accounting for Preferential Sampling in the Statistical Analysis of Spatio-temporal Data

submitted by **Joe Watson** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Statistics**.

Examining Committee:

Jim Zidek, Professor Emeritus, Statistics, UBC
Supervisor

Marie Auger-Méthé, Assistant Professor, Statistics, UBC
Co Supervisor

Lang Wu, Professor, Statistics, UBC
Supervisory Committee Member

Matias Salibian-Barrera, Professor, Statistics, UBC
University Examiner

Rebecca Tyson, Associate Professor, Mathematics, UBC Okanagan
University Examiner

Christopher K. Wikle, Curators Distinguished Professor, Statistics, University of Missouri
External Examiner

Additional Supervisory Committee Members:

Nancy Heckman, Professor, Statistics, UBC
Supervisory Committee Member

Abstract

Spatio-temporal statistical methods are widely used to model natural phenomena across both space and time. Example phenomena include the concentrations of airborne pollutants and the distributions of endangered species. A spatio-temporal process is said to have been preferentially sampled when the locations and/or times chosen to observe it depend stochastically on the values of the process at the chosen locations and/or times. When standard statistical methodologies are used, predictions of a preferentially sampled spatio-temporal process into unsampled regions and times may be severely biased. Preferential sampling within spatio-temporal data may be the rule rather than the exception in practice.

The work demonstrated in this dissertation addresses the issue of preferential sampling. We develop the first general framework for modelling preferential sampling in spatio-temporal data and apply it to historical UK black smoke measurements. We demonstrate that existing estimates of population-level black smoke exposures may be highly inaccurate due to preferential sampling. By leveraging the information contained in the chosen sampling locations, we can adjust estimates of black smoke exposure to the presence of preferential sampling. Next, we develop a fast, intuitive, powerful, and general test for preferential sampling. A user-friendly R-package we wrote performs the test. We demonstrate its utility in both a thorough simulation study and by successfully replicating previously-published results on preferential sampling. Finally, we adapt our ideas on preferential sampling to the setting of spatio-temporal point patterns. By considering the observed point pattern as a spatio-temporal thinned, marked log-Gaussian

Cox process, we show that preferential sampling can be directly accounted for within the model. Under certain assumptions, the true distribution of locations can then be attained. Using these ideas, we develop a framework for combining multiple data sources to estimate the spatio-temporal distribution of an animal. We then apply our framework to estimate effort-corrected space-use of an endangered ecotype of killer whales.

Ultimately, we hope that investigations into preferential sampling will become an essential component within spatio-temporal analyses, akin to model diagnostics. The methods developed in this dissertation are widely applicable, allowing researchers to routinely perform such investigations.

Lay Summary

Space-time statistical methods are widely used to model natural phenomena. Examples include the concentrations of airborne pollutants and the distributions of endangered species. Statistical models use data to describe these phenomena, and results from these models frequently inform governmental policy across a range of issues on matters of public health and the environment. However, it is common for the objectives underlying the data collection protocols to be related to the measured phenomenon. For example, governments preferentially situate air quality monitors near major sources of airborne pollutants. Ignoring these objectives when performing a statistical analysis can severely bias our understanding of a phenomenon. This dissertation develops tools for testing for the presence of preferential sampling and for subsequently adjusting conclusions to account for its existence. We demonstrate the utility of the methods in real-world case studies: predicting Great British air pollution concentrations, predicting lead concentrations across Galicia, and mapping killer whales.

Preface

This thesis was completed under the joint supervision of Prof. Jim Zidek and Prof. Marie Auger-Méthé. A version of Chapter 3 has been published [Watson J, Zidek JV, Shaddick G. A general theory for preferential sampling in environmental networks. The Annals of Applied Statistics. 2019;13(4):2662-700.] and versions of Chapters 4 and 5 have been submitted for peer review [Watson J. A fast Monte Carlo test for preferential sampling.] and [Watson J, Joy R, Tollit D, Thornton SJ, Auger-Méthé M. Estimating animal utilization distributions from multiple data types: a joint spatio-temporal point process framework.] respectively. The ideas for Chapter 3 have been jointly developed by Joe Watson, Prof. Jim Zidek and Prof. Gavin Shaddick with the majority of computational work and manuscript writing conducted by Joe Watson as lead author. Work seen in Chapter 4 was solely developed by Joe Watson. The ideas for Chapter 5 have been jointly developed by Joe Watson, Prof. Marie Auger-Méthé and coauthors with the majority of computational work and manuscript writing conducted by Joe Watson as lead author.

Table of Contents

List of Tables

List of Figures

Glossary

BCCSN BC Cetacean Sightings Network

BS Black Smoke

DAG Directed Acyclic Graphic

DFO Department of Fisheries and Oceans Canada

DIC Deviance Information Criterion

EPA United States Environmental Protection Agency

EU European Union

FLOPS Floating-Point Operarions

GB Great Britain

GLM Generalized Linear Model

GMRF Gaussian Markov Random Field

IID Independent and Identically Distributed

INLA Integrated Nested Laplace Approximation

IPP Inhomogeneous Poisson Process

LASSO Least Absolute Shrinkage and Selection Operator

LGCP Log-Gaussian Cox Process

MC Monte Carlo

MCAR Missing Completely At Random

MCMC Markov-Chain Monte Carlo

MSPE Mean Squared Prediction Error

NN Nearest Neighbour

NOAA National Oceanic and Atmospheric Administration

PDF Probability Density Function

PS Preferential Sampling or Preferentially Sampled depending on the context of the sentence

Q-Q Quantile-Quantile

SDM Species Distribution Model

SMRU Sea Mammal Research Unit

SPDE Stochastic Partial Differential Equation

SRKW Southern Resident Killer Whale

STGLMM Spatio-Temporal Generalized Linear Mixed-effects Model

UBC University of British Columbia

UD Utilization Distribution

UK United Kingdom

Acknowledgments

I am incredibly fortunate to have been immersed in the most welcoming, supportive, and diverse environment throughout my PhD journey. The open-door atmosphere enjoyed at UBC's Statistics Department is something that should be truly celebrated. Whilst there are numerous students, staff and faculty of the department that have helped to shape my PhD in some way, I would like to particularly thank a few individuals.

First and foremost I would like to thank my supervisor Dr. Jim Zidek and my cosupervisor Dr. Marie Auger-Méthé for their continuous support, inspiration and dedication throughout. I could not have asked for better supervisors! To Jim, I will greatly miss our weekly chats and your personal anecdotes (many of which involve the statistical 'Gods'). I thank you for inviting me to present at numerous conferences and workshops in my first year. It is thanks to you that I was able to begin developing my research network and to overcome a major fear of mine, public speaking, early on in my PhD. I hope to one day have a wealth of knowledge as great as yours. To Marie, your dedication and commitment to your students is truly special. I still don't know where you find your energy! It is thanks to you that I discovered a passion for statistical ecology, which I hope to pursue in the future. Your consistent faith in me as a researcher gave me the confidence I needed to pursue academia further.

During the final stages of my PhD journey, I was fortunate enough to have the two brilliant professors Dr. Nancy E. Heckman and Dr. Lang Wu provide feedback on my thesis. I thank you both for agreeing to spend your time and energy on my thesis! I truly appreciate the unique insights that

you both offer. To Lang, I would also like to thank you for introducing me to the literature on the use of joint models to account for response-biased sampling within biostatistical longitudinal studies. The majority of the work presented in this thesis was inspired by this literature. I would also like to express my deepest gratitude to all the members of my examining committee.

I would never have made it to UBC in the first place without the constant love and support of my parents, Lynn and Ian. I will be forever grateful to them for always listening when I need help, and, for providing advice that puts even the best agony aunt to shame! Finally, I thank Sammy Marsh for making my time in Vancouver so special.

In addition to the general acknowledgements above, I would also like to express my gratitude for the help I received with Chapters 3 and 5 that are largely based on papers. For the work in Chapter 3, I would like to thank the co-authors of the paper Jim Zidek and Gavin Shaddick. I would also like to thank the Associate Editor and the anonymous referees for their insightful comments. Their constructive feedback greatly helped to improve the focus of the paper in which the Chapter is based. For the work in Chapter 5, I would like to thank the co-authors of the paper Marie Auger-Méthé, Ruth Joy, Dominic Tollit and Sheila Thornton. I would like to thank the DFO, BCCSN, The Whale Museum and NOAA for access to sightings databases. I thank Jason Wood (SMRU), Jennifer Olson (The Whale Museum) and Taylor Shedd (Soundwatch) for their detailed insight into the operations of the whale-watch industry. Additionally, I would like to thank Eagle Wing Whale & Wildlife Tours for their substantial help with understanding the whale watching operations out of Victoria. This message of thanks extends to various other whale-watch companies who also provided assistance.

Dedication

To Sammy Marsh for helping
me see colour in the world on
the greyest of days.

Chapter 1

Introduction

"You can't fix by analysis what you bungled by design."
— Light, Singer and Willett (1990), page v

"*How were the data collected?*" "*Who collected the data?*" "*Why were these sample units chosen?*" "*Do the sample units reflect the population as a whole?*" These are fundamental questions that are asked by statisticians when they first encounter a set of claims and conclusions drawn from a new study. Analysts of the study are expected to justify how and why their sample units were chosen from the population. The reason for this scrutiny is simple and summarized well in the quote above. Inconsistencies between statistical inference and the truth are commonplace when a disconnect exists between the sample units and the population they were drawn from [Heckman, 1990]. Although methods exist for adjusting for poor design [Hernan and Robins, 2020], post hoc analysis can rarely fix these inconsistencies. It is the disconnect between sample units and the population that continues to draw skepticism on conclusions drawn from observational studies [Ebrahim and Smith, 2008].

Yet one domain of statistics has largely avoided such scrutiny despite its great importance and that is the domain of spatio-temporal modelling. Since Cressie's classic text [Cressie, 1993] laid its foundation, interest has expanded tremendously and numerous books have been written (e.g. Banerjee et al. [2015] and Cressie and Wikle [2011]). But when presented with spatio-

temporal data to model, few questions are typically asked by statisticians about how and why the sampled locations and times were chosen. Worse still, it is commonplace for these locations and times to be chosen to meet objectives related to the process being measured [Schumacher and Zidek, 1993]. For example, observations may be taken at locations and times where the process is expected to be 'highest' [EPA, 2005, Loperfido and Guttorp, 2008]. Intuitively, this can induce a dependence between the locations and times and the sampling process being measured. Surprisingly little is known about the possible consequences that a biased sampling design can have on the validity of statistical inference in spatio-temporal settings. Even fewer methods are available for adjusting inference to biases in the sampling design.

In spatio-temporal applications, a collection of (possibly noisy) observations of a spatio-temporal process are collected. We say that the data were preferentially sampled (PS), or identically that preferential sampling (PS) occurred, if a stochastic dependence exists between the choice of sampling locations and/or times with the underlying spatio-temporal process being measured [Diggle et al., 2010]. Thus PS is a special case of response-biased sampling. Interest in PS was sparked in 2010 by the landmark paper by Diggle et al. [2010]. The authors demonstrated that the consequences of PS on spatial inference can be severe, with both spatial prediction and parameter estimation affected.

Since that landmark paper, PS has been identified as a major concern across multiple fields including environmental statistics, ecology and econometrics [Fithian et al., 2015, Gelfand and Shirota, 2019, Paci et al., 2020, Pennino et al., 2019, Shaddick et al., 2016, Zidek et al., 2014]. Yet, thus far, PS has only been considered in the spatial-only setting, with no method yet proposed for modelling PS in the spatio-temporal setting. Furthermore, no general methodology exists for testing for PS. These methodologies are sorely needed. Most processes of interest are dynamic in time and recent computational, methodological, and software advances have taken place which have enabled spatio-temporal analyses to be performed with relative ease [Bakka et al., 2018]. Thus, the frequency of spatio-temporal analyses is expected to

increase into the foreseeable future.

In the remainder of this dissertation, we aim to: demonstrate that PS can be highly problematic for spatio-temporal analyses; present new methods for testing for PS; and present new methods for adjusting statistical inference to its presence. Chapter 2 provides a review of statistical methods for modelling spatio-temporal data and provides a clear definition of PS. Spatio-temporal data can be classified into one of three types: discrete-space spatio-temporal data, continuous-space spatio-temporal data, and spatio-temporal point-pattern data. Preferential sampling can affect the statistical analyses of all three types of data and different tools are required for modelling PS in each type. We briefly present toy examples of data generating mechanisms that give rise to preferential sampling in all three settings. For each setting, we then highlight the negative impact that PS can have on statistical inference when it is ignored. Finally, we introduce a popular class of hierarchical models, called spatio-temporal generalized mixed-effects models (STGLMMs). We show that these can provide a highly flexible framework for modelling preferentially-sampled spatio-temporal data of all three types. Then, we introduce a computational approach, called the integrated nested Laplace approximation (INLA), that can fit STGLMMs quickly and efficiently. Chapter 2 helps to provide context for the later Chapters.

Chapter 3 outlines the first general framework for modelling PS data in both discrete-space and continuous-space spatio-temporal settings. We demonstrate its utility by modelling historical black smoke levels in the United Kingdom. We show that the processes that determined where to place pollution monitoring sites may have had a severe impact on historical estimates of black smoke levels, including population-average exposure levels.

Chapter 4 outlines the first general test, insofar as we are aware, for preferential sampling in spatio-temporal data. We demonstrate that the test attains high power across numerous response types (continuous, count, etc.,), in both the spatial and spatio-temporal settings, for both continuous-space and discrete-space spatio-temporal data, even in settings with small sample sizes. Both the results presented in Chapter 3 and the results presented

in Diggle et al. [2010] on Galician lead concentration levels are replicated. Chapter 5 considers the effect of preferential sampling in point-pattern data. In particular, we consider the ecological application of estimating the space use of a highly mobile species, and develop a general modelling framework. Then, we apply it to a case study on killer whales (*Orcinus orca*). Finally, we provide a discussion of all the work and address the possible avenues for future research in Chapter 6.

Chapter 2

Background on Spatio-temporal Statistics

"Everything is related to everything else, but near things are more related than distant things."
— Tobler (1970)

Spatio-temporal statistics is the study of how to model and perform statistical inference on data collected in space and time. As Tobler's First Law of Geography states above, measurements taken close together in space and time are often more likely to be similar than those taken far apart. For conclusions to be valid, a statistical analysis must consider the additional autocorrelations that may exist between measurements taken across space and time. Methods for accounting for autocorrelations present within data collected across space and time are a primary focus of spatio-temporal statistics.

The statistical analysis of spatio-temporal data differs fundamentally from 'classical' statistics in many aspects. Firstly, in the spatial-only setting, it is typical for only a single replication of the process to have been observed [Cressie, 1993]. This 'sample size of one' scenario then requires that additional assumptions be placed on the underlying process to provide a foundation for statistical inference. Concepts such as stationarity

help to create 'pseudo-replicates' of the process which enable various of its characteristics to be statistically identified. Secondly, unlike many domains of 'classical' statistics, it is rare for observed spatio-temporal data to be assumed to be an independent and identically distributed (IID) sample. Instead, it is commonly believed that additional spatial, temporal, and spatio-temporal correlations exist due to a multitude of latent processes [Le and Zidek, 2006]. These processes are believed to drive the response and make observations taken close together in space and/or time appear more similar than those that are highly separated in space and/or time. Adjusting inference to account for the effects of these residual correlations requires careful treatment by the analyst.

In this dissertation, we consider descriptive models. Mechanistic models require a deeper understanding of the subject-specific matter so that a system of equations can be developed for capturing the dynamics of the spatio-temporal phenomenon through time [Wikle, 2015]. On the other hand, descriptive models attempt only to describe the first two moments of the spatio-temporal phenomenon with mean and covariance functions. By taking a covariance-based approach, descriptive models have the benefit of being highly general. However, descriptive models risk painting an oversimplified picture of complicated phenomena and incorporating subject knowledge into descriptive models can prove challenging [Wikle, 2015]. Despite these limitations, we only consider descriptive models hereafter.

The literature on descriptive models for spatio-temporal data is vast [Blangiardo and Cameletti, 2015, Cressie and Wikle, 2011, Le and Zidek, 2006]. Multiple methods including kernel density smoothers [Hallin et al., 2004], splines [Wood et al., 2017], and random fields [van Lieshout, 2019] have all been proposed to capture additional spatio-temporal correlations unaccounted for in standard analyses. Throughout this dissertation, we consider only the use of Gaussian processes, also known as Gaussian random fields in the context of spatio-temporal applications. Gaussian processes have the advantages of being extremely flexible and highly data-driven, without the need for additional tuning or bandwidth parameters that often require ad-hoc methods for their estimation. Furthermore, they are

extremely useful for uncertainty quantification, in that they naturally provide a full distribution at prediction locations instead of merely providing a point estimate.

Throughout this section we adapt the definitions found in van Lieshout [2019]. We denote a Gaussian process at a point in space $\mathbf{s} \in \Omega$ and time $t \in \mathcal{T}$ as $Z(\mathbf{s}, t)$. Formally, a Gaussian process on the dense index set $\mathcal{I} = (\Omega \times \mathcal{T}) \subset \mathbb{R}^2 \times \mathbb{R}$ is defined as follows:

Definition 2.1 *The family $(Z(\boldsymbol{s}, t))_{(\boldsymbol{s},t)\in\mathcal{I}}$ indexed on $\mathcal{I}$ is a Gaussian process (or Gaussian random field) if for any finite set of indices (i.e. points in space and time) $(\boldsymbol{s}_1, t_1), ..., (\boldsymbol{s}_n, t_n)$, the random vector $(Z(\boldsymbol{s}_1, t_1), ..., Z(\boldsymbol{s}_n, t_n))^T$ is multivariate Gaussian distributed.*

The above definition can be scaled up to higher dimensions, or to the spatial-only setting with $\mathcal{T}$ a singleton. The multivariate Gaussian distribution is fully determined by its mean and covariance matrix. This property carries over to the Gaussian process, with only the mean function and covariance function required to fully determine the Gaussian process.

Definition 2.2 *The mean function at any point in space and time $(\boldsymbol{s}, t)$ is denoted $\mu(\boldsymbol{s}, t)$ and is defined as:*

$$\mu(\boldsymbol{s}, t) = E(Z(\boldsymbol{s}, t)).$$

The covariance function at any two points in space and time $(\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j)$ is denoted $\Sigma(\boldsymbol{s}_i, \boldsymbol{s}_j; t_i, t_j)$ and is defined as:

$$\Sigma(\boldsymbol{s}_i, \boldsymbol{s}_j; t_i, t_j) = Cov(Z(\boldsymbol{s}_i, t_i), Z(\boldsymbol{s}_j, t_j)).$$

The only restriction required for $\Sigma(\cdot, \cdot; \cdot, \cdot)$ to be a valid covariance function of a Gaussian process is that it be positive-definite [van Lieshout, 2019]. This implies that for all collections of locations and times, the resulting covariance matrix must be a positive semi-definite matrix. A valid covariance function that is commonly used in spatial analyses is the Matérn covariance.

The Matérn covariance depends on three parameters, and different combinations of these can result in a Gaussian process that can exhibit a wide range of properties. For example, the orders of mean-square differentiability, and the minimum distance at which two points become 'approximately independent' can be readily defined by the parameters [Diggle et al., 2007]. For a given set of parameters, the value of the function evaluated at two points depends only on their interpoint distance. The Matérn covariance function between two points separated by a distance d units, with non-negative parameter values σ, ν, and ρ is defined:

$$C(d;\sigma,\nu,\rho) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{\rho}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{d}{\rho}\right),$$

where $\Gamma(\cdot)$ is the gamma function and K_ν is the modified Bessel function of the second kind.

For a given set of n values $\mathbf{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$ observed at locations and times $\mathcal{I} = (\mathbf{s}_1, t_1), ..., (\mathbf{s}_n, t_n)$, the probability density function (pdf) of a multivariate Gaussian distribution with given mean and covariance functions μ and Σ can be specified. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the n-vector of mean values and the $n \times n$ covariance matrix appropriately evaluated at $\mathcal{I}$. The pdf is:

$$\pi(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}-\boldsymbol{\mu})\right).$$

Here, we have used π to denote a probability density, T as a superscript to denote a matrix or vector transpose, and the vertical bar to be read as 'conditional upon'. The multivariate Gaussian distribution above has many nice properties. For example, if two variables y_i, y_j are uncorrelated (i.e. $\boldsymbol{\Sigma}_{i,j} = 0$), then they are statistically independent. A dual property can be deduced from the precision matrix $\boldsymbol{\Sigma}^{-1}$, defined as the inverse of the covariance matrix. The precision matrix is often given the notation Q. If $Q_{i,j}$, the $(i,j)^{\text{th}}$ entry of the precision matrix Q, is identically zero then the variables y_i, y_j are conditionally independent, given the values of all the

other variables $\mathbf{y}_{-(i,j)} = \mathbf{y} \setminus (y_i, y_j)$ [Rue and Held, 2005]. Thus $Q_{i,j} = 0$ implies the following Markovian property:

$$\begin{aligned}
\pi(y_i|y_j, \mathbf{y}_{-(i,j)}) &= \frac{\pi(y_i, y_j|\mathbf{y}_{-(i,j)})}{\pi(y_j|\mathbf{y}_{-(i,j)})} && \text{(Bayes' rule)} \\
&= \frac{\pi(y_i|\mathbf{y}_{-(i,j)})\pi(y_j|\mathbf{y}_{-(i,j)})}{\pi(y_j|\mathbf{y}_{-(i,j)})} && \text{(by conditional independence)} \\
&= \pi(y_i|\mathbf{y}_{-(i,j)})
\end{aligned}$$

This latter property has been exploited for model specification and for making advances in computational efficiency [Blangiardo and Cameletti, 2015, Lindgren et al., 2011b]. The conditional independence allows analysts to specify covariance matrices that are guaranteed to be positive definite, whilst being interpretable and justifiable on logical grounds [Besag, 1974b]. Whilst the true latent process being modeled is unlikely to satisfy this conditional independence property, the property provides a useful mathematical structure for approximating the latent dependency structures.

For example, a set of 'neighbours' $N(i)$ may be defined for each location $\mathbf{s}_i$ and time t_i, such that the variable y_i can be assumed to be conditionally independent of all the other variables $y_j : j \notin N(i)$, given its neighbors $y_k : k \in N(i)$. This can equivalently be specified as $Q_{i,j} = 0 : \text{iff } j \notin N(i)$. This is called the local Markov property. Popular criteria used for defining whether or not two locations/regions are neighbours include: whether or not they share a common contiguous border and whether or not their inter-point distance is less than some threshold.

Classes of flexible sparse precision matrices have been developed for modelling general spatio-temporal phenomena. When a sparse precision matrix is used, computational tricks involving sparse matrix algebra can then be applied for implementing a Gaussian process in a fast and memory-saving manner [Rue and Held, 2005]. The main gains in computational speed arise from the computation of the sparse matrix determinants and inverses. In fact, the computational order of evaluating the above pdf for a spatio-temporal

Gaussian process defined with a sparse precision matrix, instead of with a dense covariance matrix, can be reduced from $\mathcal{O}(n^3)$ floating-point operations (flops) to $\mathcal{O}(n^{3/2})$ flops [Lindgren et al., 2011b] for processes defined in continuous-space, or to $\mathcal{O}(n)$ flops [Katzfuss et al., 2020] for processes defined in discrete-space.

Gaussian processes have proven to be an effective tool for modelling residual spatio-temporal correlations in general spatio-temporal applications [Diggle et al., 2007]. As we will discuss later, the methods from spatio-temporal statistics largely fall into three categories, depending on the nature of the data and the inferential goals [van Lieshout, 2019]. The first category of methods are referred to as discrete-space spatio-temporal methods [Blangiardo and Cameletti, 2015]. These are concerned with the modelling of data collected at discrete 'sites'. These 'sites' are typically well-defined areal units such as pixels or census districts. Thus, Ω is considered a finite set for these methods. Observations are typically aggregated measures or counts, with the goal of inference being noise-reduction or the 'smoothing' of the process [van Lieshout, 2019].

The second set of methods is known as geostatistical methods and their spatio-temporal extensions, sometimes referred to as continuous-space spatio-temporal methods [Diggle et al., 2007]. These are concerned with the modelling of point-referenced data, collected in continuous-space. Thus, Ω is considered dense for these methods. Spatio-temporal interpolation is often a primary objective of these methods.

The final class of methods, referred to as spatio-temporal point process methods are concerned with the modelling of a spatio-temporal point-pattern, or point-pattern for short. The point-patterns we consider for modeling are sets of geographical points and times deemed to be random realizations from some process of interest [Baddeley et al., 2015, Illian et al., 2008]. The points may refer to events of interest (e.g. diseases, earthquake epicentres, etc.,), with the goal being the explanation of the driving forces behind the events and the identification of regional 'hotspots'. This is the only class of methods out of the three to consider the locations as random. Marked point processes consider the setting where the point-pattern, and a

set of observations taken at each point, are both considered random variables [Schlather et al., 2004].

Whilst the earlier definition of a Gaussian process was given for a dense index set, the index set, $\mathcal{I}$, may also be a finite set. Once again, the main requirement is that the covariance matrix must be strictly positive definite. Numerous classes of valid Gaussian processes have been developed that can be defined on discrete sets. An especially popular subclass are Gaussian Markov random fields (GMRFs). The addition of the 'Markov' name arises from the Markovian conditional independence property exhibited by precision matrices. As with their continuous counterparts, GMRFs are completely specified by their mean vector and covariance matrix. Popular examples include conditional autoregressive and simultaneous autoregressive models [Besag and Kooperberg, 1995, van Lieshout, 2019].

For the remainder of this Chapter, we present how PS can arise in all three types of spatio-temporal data, and how it can impact statistical inference. To achieve this, we present a toy example of PS for each type of data. In each example, we first generate two similar-sized datasets: one from a PS data generating mechanism, and one from a data generating mechanism free from PS. We then fit 'standard' models to both datasets and investigate the conclusions drawn from each model. This helps us to isolate the impacts that PS has on the statistical inference. The 'standard' statistical models used to perform the statistical inference are all examples of a popular class of models referred to as spatio-temporal generalized linear mixed-effects models (STGLMMs). We close the Chapter by rigorously defining this class of models, before providing details on a computational tool called INLA that is commonly employed to fit STGLMMs. This class of models and computational tools is used throughout the later Chapters. Note that throughout this dissertation, we consider spatial statistics, the study of how to model and perform statistical inference on data collected across space, to be a subfield of spatio-temporal statistics.

2.1 Preferential sampling in discrete-space spatio-temporal data

For general discrete-space spatio-temporal data, we assume we have a finite index set $\mathcal{I}$ of size M, which defines the population of discrete areal units $A_i \subset \Omega : i \in \{1, ..., M\}$. We assume a finite set of N times $t_j \in \mathcal{T} : j \in \{1, ..., N\}$, defining the temporal domain. These may define points in time, or time intervals, in which case $t_j \subset \mathcal{T}$. We assume that observations are taken at a subset of the areal and temporal units. We collect these observations into a response vector, denoted $\mathbf{y}$. Throughout the dissertation, we vectorize all spatio-temporal terms for convenience, and avoid matric distributions. We denote $\mathbf{Y}$ as the random variable associated with the response. We will often refer to the areal units as sites.

By Tobler's First Law of Geography, introduced earlier, we assume that the observations $\mathbf{y}$, taken close together in space and/or time, are more likely to be similar. Throughout this dissertation, we use a Gaussian Process for capturing these additional spatio-temporal correlations. In discrete space, we use a GMRF and denote it $\mathbf{Z}$. We assume $\mathbf{Z}$ is valid, and has been specified with $MN \times MN$ covariance matrix $\mathbf{\Sigma} = Q^{-1}$ and a constant mean MN-vector with entries β_0. In general, covariates collected at spatial site and time, denoted $\mathbf{x}_{i,j}$, may also be included in a model for $\mathbf{Y}$.

We now give a toy example to illustrate both the appearance and the consequences of PS in the discrete-space setting. We repeat this for the other two data types later. For simplicity, we consider the spatial-only setting throughout all examples.

A toy example

For the toy example in this section, we assume that no covariates are present and we assume that the response vector $\mathbf{Y}$ is a set of noise-free observations of the Gaussian process $\mathbf{Z}$. We assume that values of the response y_i are observed at a subset of the M sites. Let R_i denote the indicator variable for site $A_i \in \mathcal{I}$, with $R_i = 1$ indicating that we observe the process at site i. We take a Bayesian approach to statistical inference throughout the dissertation.

With this in mind, and with the above assumptions, the model we use for generating the data in the toy example is:

$$\begin{aligned}
(Y_i|R_i = 1) &= Z_i \qquad\qquad i \in \{1, ..., M\} \\
[Z_1, ..., Z_M]^T &= \mathbf{Z} \backsim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \\
R_i &\backsim Bernoulli(p_i) \\
\Theta = (\beta_0, \boldsymbol{\theta}, p_i) &\backsim \text{Priors.}
\end{aligned}$$

The Bernoulli process defining the site-selection indicators is rarely stated within a statistical analysis of spatio-temporal data. This suggests most statistical analyses are conditioned on the observed locations or sites chosen for sampling as being fixed. This is equivalent to the assumption that the probability function p is independent of $\mathbf{Z}$. Explicitly defining the Bernoulli sampling process helps to highlight the assumptions being made on the missingness mechanism of the data. For example, the assumption of a constant probability ($p_i \equiv p$), implies the sites are assumed by the analyst to be 'missing completely at random' (MCAR) [Wu, 2009]. Under this MCAR assumption, the Bernoulli process may be ignored when modelling $\mathbf{y}$ without biasing the statistical inference.

In this dissertation, we investigate the impact that different functional forms of p_i can have on the statistical inference of $\mathbf{Y}$. By our earlier definition, when p_i is related to $\mathbf{Z}$ (e.g. through a logit link), then a stochastic dependence exists between p_i and $\mathbf{Z}$ and thus the discrete-space spatio-temporal data have been preferentially sampled.

We define 1078 discrete spatial units and generate a single realisation of the GMRF model above to form the complete data $\mathbf{Z}$. Next, we generate the two datasets in the spatial-only setting by changing only the specification of p_i. The first sampling scheme sees the value of Z_i appear linearly in the logit of p_i with a positive parameter. This preferentially selects for observation the sites with the highest values of Z_i, and preferentially censors the spatial units with the lowest values of Z_i. Plots of the true field and the 184

preferentially-selected areal units are shown in panels A and B of Figure 2.1 respectively. Data are recorded more often in areal units situated in the South-Western corner of the region. These are precisely the areal units with the highest values of the process. The converse is seen in the South-Eastern and South-Central regions.

The second sampling scheme sees $p_i \equiv p$. Under this scheme, sites are sampled completely at random. This provides an example of a MCAR site-selection process, free of preferential sampling. The units are all equally-likely of being selected and hence the region is expected to be consistently represented across space. Under this sampling scheme, we should be able to condition on the selected sites as fixed without biasing inference. For brevity, we omit a plot of these chosen areal units. We fix the number of selected areal units to be 184 to remove any effects of sample size from the conclusions. In both cases, we fit the correctly specified model for $\mathbf{Y}$, albeit with the selection process determining R_i ignored. As mentioned earlier, the selection process is typically ignored in practice. This allows us to highlight the impacts that PS can have on the statistical inference of discrete spatial data. These concepts naturally transfer into the spatio-temporal setting.

The posterior mean values of the field are then estimated across all 1078 areal units. These are shown in Panels C and D of Figure 2.1 for the MCAR and PS datasets respectively. Clearly the posterior estimates of the field from the PS data consistently overestimate the true values throughout the data-sparse regions. Conversely, those from the MCAR data appear to reflect the true values much better. This disparity can best be summarized by looking at the posterior distributions of the intercept β_0 from the two models. The posterior mean and 95% credible intervals from the models fit to the PS and the MCAR data are respectively -0.378 (-0.449, -0.307) and -0.795 (-0.852, -0.738). For comparison, the values for the model fit to the complete data are -0.8 (-0.802, -0.798). The posterior distribution of the intercept is positively-biased for the PS data. This bias is a defining characteristic of PS in discrete-space spatial data, and the bias carries over to the spatio-temporal setting. In Chapters 3, we develop a general framework for adjusting a statistical analysis of spatio-temporal data to PS. In

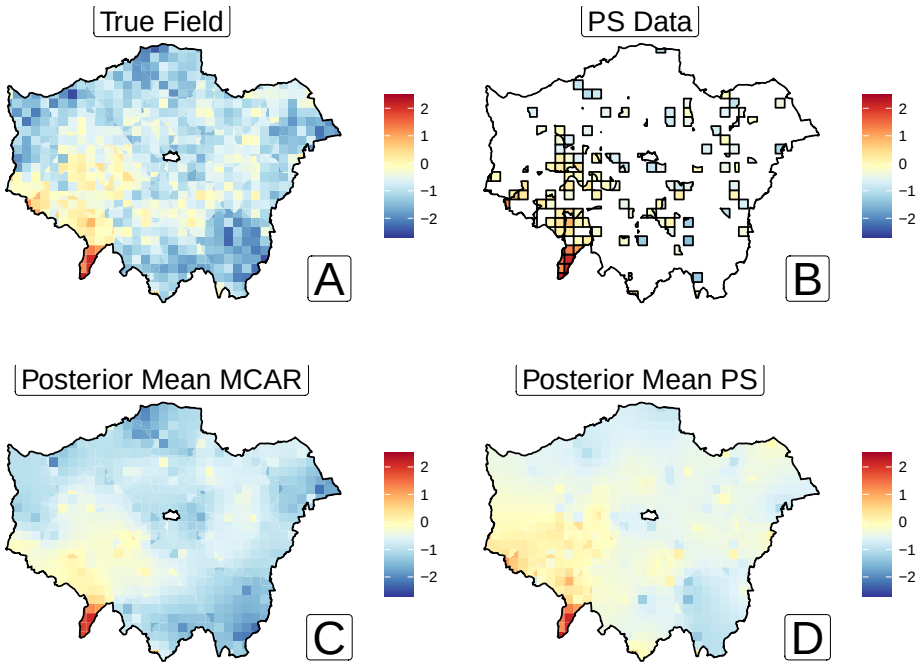

Figure 2.1: Plots of the toy simulation study. Panel A shows the complete dataset simulated from a Gaussian Markov random field. Panel B shows the realisation of the preferentially-sampled Bernoulli sampling process. Panel C shows the posterior means of the field from the model fit to the data sampled from the MCAR process. Panel D shows the posterior means of the field from the model fit to the preferentially-sampled data. Observe how poorly D characterizes the field in A compared with C

Chapter 4, we develop a test for PS in spatio-temporal data. Both apply to the discrete-space setting.

2.2 Preferential sampling in continuous spatio-temporal data

For general continuous-space spatio-temporal data, the index set $\mathcal{I}$ is a dense subset $\Omega \times \mathcal{T}$ of $\mathbb{R}^2 \times \mathbb{R}$. Data are collected at a finite set of spatio-temporal

locations, denoted $(\mathbf{S}, \mathbf{T}) = (\mathbf{s}_1, t_1), ..., (\mathbf{s}_n, t_n)$ and vectorized into a response vector $\mathbf{y}$. Once again, we will include a Gaussian process $Z(\mathbf{s}, t)$ within the model used to describe $\mathbf{Y}$. Within the STGLMM framework introduced later, the random variables of the response at each space-time coordinate $(\mathbf{s}, t)$, denoted $Y(\mathbf{s}, t)$, are then assumed to be independently distributed, given $Z(\mathbf{s}, t)$. The Gaussian process $Z(\mathbf{s}, t)$ is well-defined at all locations and times within $\Omega \times \mathcal{T}$ with mean function $\mathbf{0}$ and covariance function $\Sigma(\cdot, \cdot; \cdot, \cdot)$. In general, covariates $\mathbf{x}(\mathbf{s}, t)$ may be available at each of the finite set of locations. Often, the goal of spatio-temporal analyses of continuous-space data is spatio-temporal interpolation or prediction.

As before, we now present a toy example. Again, we consider the spatial-only setting and ignore covariates.

A toy example

We assume that a finite set of locations, denoted $\mathbf{S}$, are generated from a point process. Details of point processes are found in the next section. At each of these locations $\mathbf{s}_i \in \mathbf{S}$, a noise-free observation of $\mathbf{Z}(\mathbf{s}_i)$ is made and stored in the vector $\mathbf{y}$. Under the above assumptions, the model we use for generating the data in the toy example is:

$$\begin{aligned}
(Y(\mathbf{s}) | \mathbf{s} \in \mathbf{S}) &= Z(\mathbf{s}) \\
\mathbf{Z}(\mathbf{S}) &\backsim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \\
\mathbf{S} &\backsim \text{Point process}(\lambda(\mathbf{s})) \\
\Theta = (\beta_0, \boldsymbol{\theta}, \lambda) &\backsim \text{Priors.}
\end{aligned}$$

As before, analysts rarely state their assumptions about the sampling process that generated the set of locations at which observations of the process were made. Instead, analysts typically condition on the locations $\mathbf{S}$ as fixed when performing a statistical analysis of a continuous-space spatial dataset [Diggle et al., 2010]. In this dissertation, we argue that the assumed sampling processes should be explicitly defined. Clearly stating our beliefs

on how the locations were sampled helps to highlight the possible consequences that conditioning on the locations as fixed, and hence ignoring the sampling process, can have on the statistical inference of $\mathbf{Y}$. For example, if we know *a-priori* that all regions were sampled with equal intensity and hence that $\lambda(\mathbf{s}) \equiv \lambda \; \forall \mathbf{s} \in \Omega$, then the sampling process may be ignored within a statistical analysis of $\mathbf{Y}$ without biasing the inference. However, when $\lambda(\mathbf{s}) \propto \mathbf{Z}$, the data are said to have been preferentially sampled and ignoring the sampling process can bias the statistical inference about $\mathbf{Y}$ [Gelfand et al., 2012].

We simulate a Gaussian process with a spatial Matérn covariance function on the unit square with parameters chosen to ensure the spatial range and marginal standard deviation are equal to 0.7 and 1 respectively. The realisation is shown in panel A of Figure 2.2. The values of the field appear larger in the West and smaller in the East. Next, we define the two sampling schemes that generate the locations at which $Z(\mathbf{s})$ is observed without error.

In the first scheme, sampling is preferential. The sampling locations are a realisation of an inhomogeneous Poisson point process. The value of the field at location $\mathbf{s}$, $Z(\mathbf{s})$, is included linearly within the log intensity, $\log(\lambda(\mathbf{s}))$, of the point process with a positive coefficient. Under this sampling scheme, a high density of sampling locations is expected in regions where $Z(\mathbf{s})$ is high (e.g. the West). Conversely, in regions where $Z(\mathbf{s})$ is small, the density of sampling locations is expected to be low (e.g. the East). A plot of the 24 chosen locations is shown in Panel C of Figure 2.2. Indeed, the expected behaviour is seen.

The second sampling scheme can be thought of as the continuous analogue of MCAR. The locations at which to sample the field are a realisation of a homogeneous Poisson point process. Thus, $\lambda(\mathbf{s}) \equiv \lambda$. Under this scheme, the density of sampling locations is expected to be constant throughout the unit square. Furthermore, the average intensity of this sampling scheme is set equal to that of the previous sampling scheme to encourage a similar sample size to be realized. Indeed, 27 locations are chosen to observe the process under the second sampling scheme and are shown in the bottom right plot of Figure 2.2. No obvious spatial trend in the density of the points

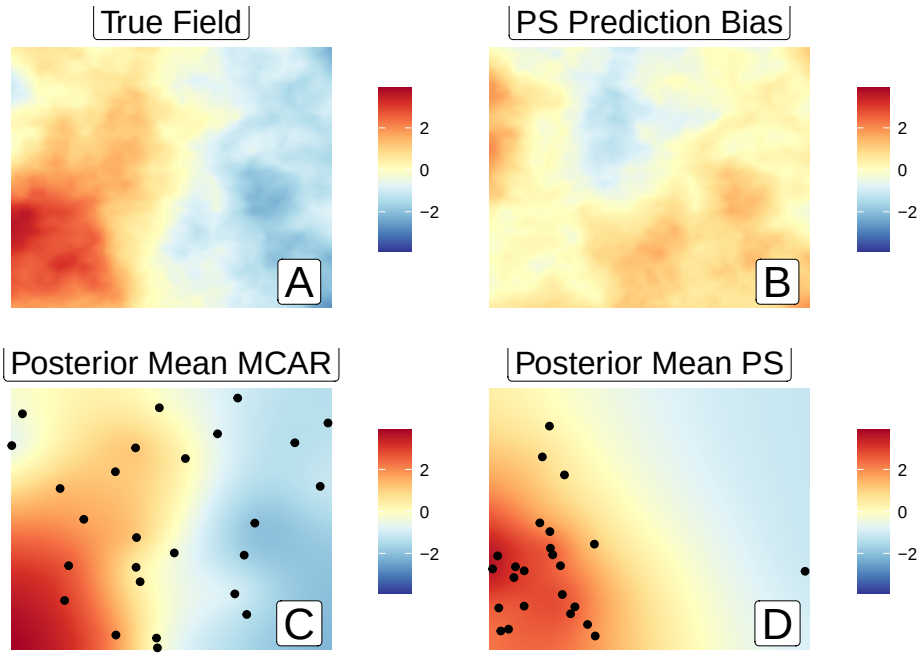

Figure 2.2: A series of plots showing the toy simulation study. Panel A shows the complete data that were simulated from a Gaussian random field. Panel B shows the bias of the posterior mean from the model fit to the preferentially-sampled data. Panel C shows the posterior means of the field from the model fit to the MCAR data. The locations at which samples of the field were taken are shown in black. Panel D shows the posterior means of the field from the model fit to the preferentially-sampled data. The locations at which samples of the field were taken are shown in black.

can be seen.

Next, the correctly-specified model for $\mathbf{Y}$ is fit to both datasets, albeit with the sampling process ignored by conditioning on the locations as fixed to reflect a typical geostatistical analysis. For both models, the posterior mean values of the field are then estimated across a grid of pixels placed across the unit square. The posterior means are shown in colour behind the points in panels C and D of Figure 2.2. It can immediately be seen that

the posterior mean values from the model fit to the preferentially-sampled dataset do not capture the lower tail or the contrast of the field.

To better highlight the impacts that preferential sampling has on inference, we compute the prediction bias from the model fit to the preferentially-sampled data. This is shown in panel B of Figure 2.2. It can be seen that the model overestimates the true value of the field in almost the entirety of the unit square. Once again, the abilities of the models to capture the global average value of the field can be assessed through summary statistics. The posterior average prediction bias, averaged across the pixels, is 0.01 and 0.20 for the models fit to the MCAR and preferentially-sampled datasets respectively. This prediction bias is the norm when standard statistical models are fit to preferentially sampled data [Diggle et al., 2010]. This includes spatio-temporal data. This demonstrates that we risk biasing our inference if we ignore the sampling processes that generated the locations chosen to observe the process.

In Chapters 3, we develop a general framework for adjusting a statistical analysis of spatio-temporal data to preferential sampling. In Chapter 4, we develop a test for preferential sampling in spatio-temporal data. Both apply to the continuous-space setting.

2.3 Preferential sampling in spatio-temporal point-pattern data

Background on log-Gaussian Cox processes

Point processes are used for modelling point-pattern data. Whereas in the previous subsection a point process was used to characterize the points at which the realization of $\mathbf{Z}$ was observed, in this section the realization is the points themselves. To guarantee that a valid stochastic process is defined for characterising the points, a set of underlying assumptions is required. The first definition defines collections of point configurations that can be more easily modelled. In particular, the definition restrict us to consider only configurations that have countably infinite numbers of points.

Definition 2.3 *The family $N^{lf}(\mathbb{R}^2 \times \mathbb{R})$ of locally finite point configurations in $\mathbb{R}^2 \times \mathbb{R}$ consists of all subsets $(\boldsymbol{S}, \boldsymbol{T}) \subset \mathbb{R}^2 \times \mathbb{R}$ such that for every bounded Borel set $A \subset \mathbb{R}^2 \times \mathbb{R}$, finitely many points are contained within the intersection $(\boldsymbol{S}, \boldsymbol{T}) \bigcap A$.*

The next definition defines a point process on the family of locally finite point configurations.

Definition 2.4 *A point process $(\boldsymbol{S}, \boldsymbol{T})$ on $\mathbb{R}^2 \times \mathbb{R}$ with realisations in $N^{lf}(\mathbb{R}^2 \times \mathbb{R})$ is a random, locally finite configuration of points such that for all bounded Borel sets $A \subset \mathbb{R}^2 \times \mathbb{R}$, the number of points of $(\boldsymbol{S}, \boldsymbol{T})$ that fall within A is a finite random variable that we denote $N_{S,T}(A)$.*

Whilst many classes of point process models exist, we focus our attention on the flexible class of inhomogeneous spatio-temporal Poisson point processes (IPPs). IPPs can be made even more flexible through the addition of Gaussian processes, and can then be computed by approximating their likelihood with models falling within the STGLMMs framework introduced later [Simpson et al., 2016]. Note that we remove the subscript $\mathbf{T}$ from the random variable $N_{S,T}(\cdot)$ for readability.

Definition 2.5 *A spatio-temporal point process $(\boldsymbol{S}, \boldsymbol{T})$ on $\mathbb{R}^2 \times \mathbb{R}$ is an inhomogeneous spatio-temporal Poisson point process with intensity $\lambda(\boldsymbol{s}, t) : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^+$ if:*

- *For every bounded Borel set $A \subset \mathbb{R}^2 \times \mathbb{R}$, $N_S(A)$ is Poisson distributed with mean $\Lambda(A) = \int_A \lambda(\tilde{\boldsymbol{s}}, \tilde{t}) d\tilde{\boldsymbol{s}} d\tilde{t}$,*
- *For any k disjoint bounded Borel sets $A_1, ..., A_k, k \in \mathbb{N}$, the random variables $N_S(A_1), ..., N_S(A_k)$ are independent,*
- *$\lambda(\cdot, \cdot)$ is integrable on bounded sets.*

The IPP is a very flexible model. The intensity $\lambda(\cdot, \cdot)$ may be allowed to vary across space and time according to a set of linear and nonlinear

covariates, including splines. The likelihood of a point-pattern $\mathbf{Y} = (\mathbf{S}, \mathbf{T})$ observed within bounded spatial and temporal domains Ω and $\mathcal{T}$ is:

$$\pi(\mathbf{Y}|\lambda(\cdot,\cdot)) = \exp\left\{|\Omega||\mathcal{T}| - \int_{\Omega}\int_{\mathcal{T}} \lambda(\mathbf{s},t)\mathrm{d}t\mathrm{d}\mathbf{s}\right\} \prod_{(\mathbf{s}_i,t_i)\in\mathbf{Y}} \lambda(\mathbf{s}_i,t_i), \qquad (2.1)$$

with $|\Omega|$ and $|\mathcal{T}|$ denoting the area and length of the study region and temporal domain respectively.

However, the stringent mean-variance relationship assumed on the counts $N_S(A)$ is undesirable in many settings. Overdispersion, which in this setting occurs when the variance of the counts exceeds the mean, frequently occurs in practice. Furthermore, additional spatio-temporal clusters may form due to the presence of unmeasured processes that drive the intensity. To alleviate these concerns, the class of Cox spatio-temporal point processes extends the Poisson process by allowing $\lambda(\cdot,\cdot)$ to be a realisation of a random field.

If we assume the random field is a realisation of a log-Gaussian process, then we have the class of log-Gaussian Cox spatio-temporal point processes (LGCP herafter). LGCPs can be specified more simply in hierarchical form:

Definition 2.6 *$(\boldsymbol{S}, \boldsymbol{T})$ on $\mathbb{R}^2 \times \mathbb{R}$ is an LGCP with intensity $\lambda(\boldsymbol{s},t) : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^+$, Gaussian process $\boldsymbol{Z}$, and covariates $\boldsymbol{x}(\boldsymbol{s},t) \in \mathbb{R}^p$, if conditioned upon $Z(\boldsymbol{s},t)$:*

- *$log(\lambda(\boldsymbol{s},t)) = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}(\boldsymbol{s},t) + Z(\boldsymbol{s},t)$,*
- *$(\boldsymbol{S}, \boldsymbol{T})$ is an inhomogeneous spatio-temporal Poisson point process with intensity $\lambda(\cdot,\cdot)$.*

LGCPs are an especially popular class of point process. The hierarchical definition given above makes them relatively easy to fit compared with other competing models [Simpson et al., 2016]. Furthermore, the flexibility of Gaussian processes carries over, allowing for point processes with highly flexible properties to be specified [Yuan et al., 2017]. Marginally, the LGCP can have very general mean-variance relationships due to the various covariance structures that can be specified on the Gaussian process. Furthermore,

stochastic dependence can exist between the counts in disjoint regions [Baddeley et al., 2015]. This property can account for the additional clustering that is often present in real point-pattern data.

One of the most useful properties of the IPP and the LGCP within ecological [Fithian et al., 2015] and epidemiological [Johnson et al., 2019] applications is its ability to serve as an integrative modelling framework for jointly modelling datasets that were collected following different protocols. In addition to modelling the direct observations of points in space and/or time, the LGCP allows for the simultaneous modelling of aggregated counts and/or binary indicator variables of the occurrences of points within well-defined spatial units and time intervals [Hefley and Hooten, 2016]. In fact, the conditional likelihoods of these different observation processes can all be derived from the fact that the number of points within any bounded Borel set A, conditioned on knowing the Gaussian process $\mathbf{Z}$, is Poisson distributed (Definitions 2.5 and 2.6). In particular, Definition 2.5 implies that $\forall A = (\mathbf{S}, \mathbf{T}) \subset \Omega \times \mathcal{T}$:

$$(N_S(A)|\mathbf{Z}) \backsim \text{Poisson}(\Lambda(A)) \tag{2.2}$$

$$\mathbb{P}(N_S(A) > 0|\mathbf{Z}) = 1 - \mathbb{P}(N_S(A) = 0|\mathbf{Z}) = 1 - \exp(\Lambda(A)). \tag{2.3}$$

Equation 2.2 shows that aggregates of point process data into counts within well-defined spatial and temporal units may be data modelled as Poisson counts, conditional upon $\mathbf{Z}$. If, instead of counts, the point process data are instead summarized by binary indicator variables of occurrence (i.e. $\mathbb{I}(N_S(A) > 0)$), then equation 2.3 shows that the occurrence events may be modelled as a collection of Bernoulli random variables with probability of success equal to $1 - \exp(\Lambda(A))$, conditional upon $\mathbf{Z}$.

In either case, conditional upon knowing $\mathbf{Z}$, the likelihoods of data collected from different observation processes may be combined. We will see that many of these aggregated data types will naturally take the form of spatio-temporal generalized linear mixed-effects models (STGLMMs). To compute the likelihood of the LGCP, additional work is required. The inte-

gral seen in (2.1) is intractable, and therefore computational approximations are required to compute the likelihood [Simpson et al., 2016]. Approximations are often derived by taking aggregates of the point process over very small regions. Poisson regression and logistic regression methods based on (2.2) and (2.3) are two such examples [Baddeley et al., 2015]. These computational approximations to LGCPs fall within the STGLMMs framework and thus the computational method (INLA) that we introduce in section 2.4 can be used for fitting LGCPs.

Preferential sampling in spatio-temporal point-pattern data

Preferential sampling within spatio-temporal point-pattern data can best be summarized by a two-step process. In the first step, a spatio-temporal point-pattern $(\mathbf{S}, \mathbf{T})$ is generated. In the second step, a process acts upon $(\mathbf{S}, \mathbf{T})$, causing each space-time point $(\mathbf{s}_i, t_i) \in (\mathbf{S}, \mathbf{T})$ to be either retained or discarded. The process in the second stage is known as a 'thinning process' and it determines which points of $(\mathbf{S}, \mathbf{T})$ are kept and which are discarded prior to analysis.

Only the retained points, denoted $(\mathbf{S}_0, \mathbf{T}_0) \subset (\mathbf{S}, \mathbf{T})$ are made available to the analyst. Preferential sampling then occurs when the probability that a point at location $(\mathbf{s}, t) \in \Omega \times \mathcal{T}$ is retained, denoted $p(\mathbf{s}, t)$, depends upon the underlying intensity $\lambda(\mathbf{s}, t)$ at that location. When this occurs, the thinning process is stochastically dependent on the underlying intensity $\lambda(\mathbf{s}, t)$. As before, the statistical inference of preferentially-sampled point-pattern data may be biased. When the thinning process discards the points uniformly throughout $\Omega \times \mathcal{T}$, then the intensity $\lambda(\mathbf{s})$ may be recovered from the partially-observed data without adjustment, albeit with a negatively-biased intercept within the log-linear model for $\lambda(\mathbf{s}, t)$. The magnitude of the bias depends on the proportion of points that are discarded.

In practice, the thinning process may be deterministic and retain points with certainty within certain subregions of Ω, and time intervals of $\mathcal{T}$ denoted Ω_0 and $\mathcal{T}_0$ respectively. When this occurs, the analysis must be suitably adjusted to match the partial observations [Chakraborty et al., 2011]

Again, if Ω_0 and $\mathcal{T}_0$ are chosen where $\lambda(\mathbf{s}, t)$ is high (or low), then preferential sampling occurs. Alternatively, the thinning process may be stochastic, with $p(\mathbf{s}, t)$ driven by factors that affect the visibility or detectability of the points [Fithian et al., 2015]. For example, weather conditions and/or terrain ruggedness may need to be considered in cases where the points represent the presence of animals at points in space and time.

In either case, thinned point processes allow for the modelling of partially-observed point processes and the intensity of a thinned point process can be derived [Hefley and Hooten, 2016]. The intensity of the observed point-pattern, $\lambda_{obs}(\mathbf{s}, t)$, is the product of the true data-generating intensity, $\lambda(\mathbf{s}, t)$, and $p(\mathbf{s}, t)$, reflecting the imperfect detectability through space and time. Consequently, if $p(\mathbf{s}, t)$ can be estimated well, then the bias due to imperfect detection may be partially controlled for in an analysis.

Once again we simulate a toy example in the spatial-only setting without covariates to highlight the impacts that ignoring the sampling process can have on the statistical inference of spatial point-patterns.

A toy example

We sample a complete point-pattern from an LGCP. To do this, we simulate a Gaussian process with spatial Matérn covariance on the unit square with spatial range equal to 0.7 and standard deviation equal to 1. Next, we generate a realisation of a Poisson point process in the unit square with intensity equal to the exponential of the random field, scaled by 500. The realized field is shown in the top left plot of Figure 2.3, with the realisation of the 867 points shown in black. These form the complete set of points. There is a clear increase in the density of the points within the central-Western region, precisely where the field is largest.

Next, we assume that the points are imperfectly detected according to two thinning processes. In the first case, the point-pattern is preferentially sampled according to a detection probability surface $p(\mathbf{s}) = C\lambda(\mathbf{s})$, assumed to be unknown to the analyst. Each of the points from the complete pattern are either retained or discarded following independent Bernoulli trials

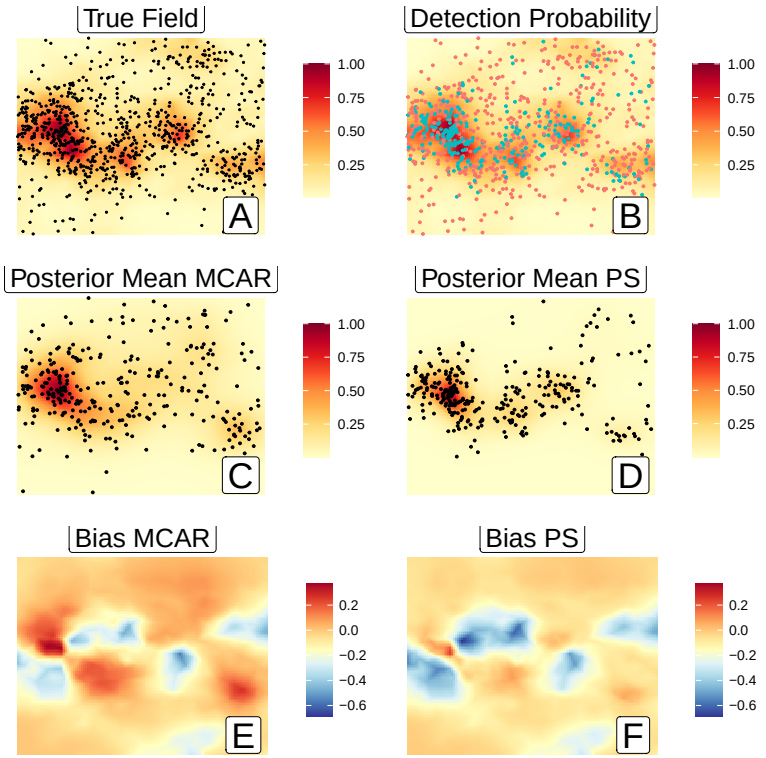

Figure 2.3: Plots showing the toy simulation study. Panel A shows the complete point-pattern data that was simulated from a log-Gaussian Cox process. The normalized intensity surface that generated the points is shown behind the points as a raster in colour. Panel B shows the detection probability surface used to preferentially thin the point-pattern. The complete point-pattern is included. The retained and discarded points from a single realisation of a Bernoulli process with probabilities given by the detection probability surface are shown in blue and red respectively. Panel C shows the normalized posterior means of the intensity from the log-Gaussian Cox process model fit to the MCAR data which are shown in black. Panel D shows the normalized posterior means of the intensity from the log-Gaussian Cox process model fit to the preferentially-sampled dataset which are shown in black. Panels E and F show the prediction biases of the normalized intensities from the models fit to the MCAR and preferentially-sampled datasets respectively. The model fit to the PS data underestimates the field almost entirely throughout Ω.

with success probabilities following $p(\mathbf{s})$. The constant C is chosen to ensure $p(\mathbf{s}) \in [0,1]$. Panel B of Figure 2.3 shows $p(\mathbf{s})$, with the retained and discarded points plotted in blue and red respectively in the figure. It is immediately apparent that the points are only retained in regions where $\lambda(\mathbf{s})$ is high; a telltale behaviour of preferential sampling.

The second sampling scheme is representative of a MCAR scheme in the context of point-patterns. Each point is equally likely to be retained (i.e. $p(\mathbf{s}) = p$). For comparative purposes, we fix the total number of points at 303 to match the number of points sampled under the previous sampling scheme. Under this sampling scheme, each region is equally under-represented. Thus, the spatial density of the retained point-pattern, $(\mathbf{S}_0, \mathbf{T}_0)$, throughout Ω is expected to closely match that of the complete pattern, with the majority of global features retained. In fact, the intensity of $(\mathbf{S}_0, \mathbf{T}_0)$ is identically $p\lambda(\mathbf{s})$. We use this fact for model comparison purposes next.

Log-Gaussian Cox processes are fit to both point-patterns, with $p(\mathbf{s})$ assumed constant, and hence the sampling processes ignored. To compare the results from the two sampling schemes, we compute the posterior mean intensities (i.e. the fields) from the models and then normalize to the unit interval. Panel A of Figure 2.3 shows the true normalized intensity. Panels C and D of Figure 2.3 show the normalized posterior mean intensity from the model fit to the preferentially-sampled data and the MCAR datasets respectively. The retained points are plotted in black. It can be seen that the model fit to the preferentially-sampled data dramatically underestimates the field throughout Ω. In general, apart from the major hotspots, the global characteristics of the field are missed. Conversely, this is not seen from the model fit to the MCAR data.

Panels E and F in Figure 2.3 present the prediction bias of the normalized field from the two models. The consistent underestimation of the field around the central band of Ω from the PS model can be seen. In summary, the preferential detection probability surface leads to a model that badly characterizes the field. The average prediction bias throughout Ω is -0.12 and -0.05 for the PS and MCAR models respectively, demonstrating once again that large biases in the predictions of the field are a direct conse-

quences of PS. This is a clear demonstration that the sampling scheme must be considered within a point process analysis.

In practice, it is is typically assumed that either the points were perfectly detected, the points were retained with constant $p(\mathbf{s}, t) \equiv p$, or that a set of covariates are available that can fully explain the $p(\mathbf{s}, t)$. In Chapter 5 we relax these assumptions and also allow for the underlying process $Z(\mathbf{s}, t)$ to drive the sampling intensity. Thus we allow for preferential sampling to be accounted for directly within a modelling framework.

2.4 Spatio-temporal generalized linear mixed-effects models

In the previous three sections, we saw examples of PS across all three types of spatio-temporal data. All of these models can be considered within the class of spatio-temporal generalized linear mixed models (STGLMMs), defined now. We focus on STGLMMs as they provide a very flexible class of models that can suitably model a large range of phenomena [Gómez-Rubio, 2020]. STGLMMs are relatively easy and fast to implement using the approximate INLA approach discussed later in this Section. Furthermore, the models used to describe preferential sampling in the three data types: the Bernoulli process, the point process, and the thinning process can all be modelled within the STGLMMs framework [Gómez-Rubio, 2020]. Thus, we can develop joint models for PS that fall within the STGLMMs framework. We demonstrate this later in this Section.

The models seen in Sections 2.1 - 2.3 all ignored covariates. Covariates are frequently available in practice and can be easily included within the STGLMM framework. Furthermore, the models in Sections 2.1 and 2.2 assumed that the Gaussian process $\mathbf{Z}$ was observed without noise. In practice, most data $\mathbf{Y}$ will have large amounts of measurement error and noise and will not be reasonably modelled as Gaussian. It may also be impossible to transform the responses to approximate Gaussianity without any additional skewness, heavy tails, or heteroscedasticity remaining. A prime example of such data was seen with the aggregated point process data in Section 2.3

where the responses were binary and count data. A vast library of statistical distributions exist that can account for the quirks and individualism that real data possess [Krishnamoorthy, 2016]. The STGLMM framework provides a convenient hierarchical modelling framework that can place suitable distributions on the responses $\mathbf{Y}$.

SGLMMs, the spatial-only equivalent to STGLMMs, were popularized by the papers of Besag et al. [1991] and Diggle et al. [1998] in the discrete-space and continuous-space settings respectively. The basic idea of SGLMMs is to construct a generalized linear model of the response variable, conditional upon a series of spatially correlated random effects. Linear combinations of these random effects alongside covariates are then included within the linear predictor of the generalized linear model (GLM) describing the transformed expectation of the response. The GLM set-up assumes that the responses are conditionally independent of each other, given the random effects. STGLMMs then extend the SGLMM framework by allowing for spatio-temporally correlated random effects. The general set-up of STGLMMs is:

$$
\begin{aligned}
Y_{i,j} &\backsim f_Y(g(\mu_{i,j}), \boldsymbol{\theta}_Y), \quad f_Y \backsim \text{density} \\
g(\mu_{i,j}) = \eta_{i,j} &= \mathbf{x}_{i,j}^T \boldsymbol{\gamma} + \sum_{k=1}^{q} u_{i,j,k} \beta_k(\mathbf{s}_i, t_j) \\
\beta_k(\mathbf{s}_i, t_j) &\backsim N(\mathbf{0}, \Sigma_k(\boldsymbol{\theta_k})) \quad k \in \{1, .., q\} \\
\Theta = (\boldsymbol{\theta_Y}, \boldsymbol{\gamma}, \boldsymbol{\theta_1}, ..., \boldsymbol{\theta_q}) &\backsim \text{Priors} \\
\mathbf{x}_{i,j} \in R^p, \quad \mathbf{u}_{i,j} &\in R^q.
\end{aligned}
$$

The above framework is evaluated on finite spatial and temporal index sets $i \in \{1, ..., M\}$ and $j \in \{1, ..., N\}$. This will either correspond to the population of discrete spatial units, the set of all observation and prediction locations being considered for a continuous-space spatio-temporal analysis, or the set of all spatial and temporal units used to approximate the computation of the IPP's likelihood (2.1). The function g is known as the link

function and maps the support of the expectation of the response onto the real line. The linear predictor η contains a linear combination of covariates $\mathbf{x}$ and latent effects $\beta_k(\cdot,\cdot)$, possibly scaled by covariates $\mathbf{u}$. These latent effects are zero-mean Gaussian distributed, possibly with white noise, spatial, temporal and/or spatio-temporal covariance structures $\Sigma_k(\theta_k)$ assumed on them. We change the notation from $\mathbf{Z}$ to β to highlight the fact that Gaussian processes can also be used for capturing spatio-temporal changes in the effects of covariates. This is done in Chapter 3.

For this dissertation, we work in the Bayesian setting. We choose this approach for two reasons. First, in spatio-temporal settings, it is often the case that numerous combinations of hyperparameters θ_k that define the latent effects $\beta_k(\cdot,\cdot)$ within an STGLMM may explain a given dataset 'well'. Put differently, in typical settings, the hyperparameters may only be weakly identified [Zhang, 2004]. The Bayesian approach offers a natural method for averaging over this (potentially large) set of competing models. By placing additional prior distributions on all the hyperparameters, Bayesian methods naturally perform a weighted-average across all the possible predictive distributions defined by the set of possible values that the hyperparameters can take. The weights of each model are proportional to the posterior distribution of the hyperparameters in light of the data and the assumed priors [Robert, 2007]. Conversely, competing methods such as empirical Bayes or type-II maximum likelihood-based approaches condition on the estimated hyperparameters as known when forming model-based predictions of both η and μ [Le and Zidek, 2006]. Thus, they may miss (potentially large) uncertainties in the hyperparameter estimates.

Secondly, in data-sparse settings, there may also be an additional advantage offered by Bayesian methods. When additional knowledge is known a-priori about the true levels of complexity present in the spatio-temporal phenomenon being studied, then hyperparameters may naturally be constrained to take 'sensible' values through the use of penalized complexity prior distributions [Fuglstad et al., 2018, Simpson et al., 2017]. These can help to reduce the risk of over-fitting that can plague flexible modelling frameworks. Whilst penalty functions such as the LASSO penalty [Tibshi-

rani, 1996], may be added to frequentist analyses to reduce the risk of overfitting, the Bayesian approach offers a more flexibile and intuitive framework for incorporating prior knowledge to achieve this goal.

Given the hierarchical specification, Markov chain Monte Carlo methods (MCMC) provide a convenient and popular method for fitting STGLMMs [Gelfand et al., 2010]. Unfortunately, MCMC methods can be computationally-demanding and time-consuming without the development of bespoke efficient MCMC implementations. Thus we do not consider these approaches and instead look at the recently-developed Integrated Nested Laplace Approximation (INLA) approach [Rue et al., 2009] for model-fitting in Section 2.4.1. We discuss in depth the benefits and limitations of using the INLA approach to fit STGLMMs in Section 2.4.2.

2.4.1 Incorporating preferential sampling within STGLMMs

Another major advantage of the STGLMMs framework for modelling a response vector $\mathbf{Y}$ is that it may also be used for modelling the preferential sampling process. Thus, a joint modelling framework may be developed for accounting for preferential sampling in all spatio-temporal data settings.

We consider the discrete-space and continuous-space settings separately from the point-pattern setting for reasons that will be made clear. For the discrete-space and continuous-space settings, let the STGLMM for $\mathbf{Y}$ be defined as above. Next, define the $R_{i,j}$ as the random variables used for defining the PS. In the discrete-space setting, the $R_{i,j}$ will be indicator variables describing the site-selection process that determines which of the areal units contain data. In the continuous-space setting, multiple choices of $R_{i,j}$ are available for approximating the point process (LGCP) that determines the selection of locations and times $(\mathbf{S}, \mathbf{T})$ chosen to observe the spatio-temporal phenomenon. Two approximations are commonly used. The first method sets $R_{i,j}$ as binary indicator events and then approximates the LGCP with a logistic regression STGLMM [Warton et al., 2010]. The second approach sets $R_{i,j}$ as counts and then approximates thr LGCP with a Poisson STGLMM [Baddeley et al., 2015]. We use the former in Chapter 3.

Now, we define the model for PS as:

$$
\begin{aligned}
R_{i,j} &\backsim f_R(h(m_{i,j}), \boldsymbol{\theta}_R), \;\; f_R \backsim \text{density} \\
h(m_{i,j}) = \zeta_{i,j} &= \mathbf{x}_{i,j}^T \boldsymbol{\alpha}_1 + \mathbf{w}_{i,j}^T \boldsymbol{\alpha}_2 + \sum_{k=1}^{q} v_{i,j,k} \beta_k(\mathbf{s}_i, t_j) \\
\Theta_R &= (\boldsymbol{\theta_R}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \backsim \text{Priors} \\
\mathbf{w}_{i,j} &\in R^r, \;\; \mathbf{v}_{i,j} \in R^q.
\end{aligned}
$$

We jointly fit the above model with the earlier STGLMM for $\mathbf{Y}$. If we assume the joint model truly generates the data, then preferential sampling can arise in two distinct ways. The necessary approach required for adjusting for preferential sampling differs across the two. The first way that preferential sampling can occur is when one or more latent effects $\beta_k(\mathbf{s}, t)$ are shared between $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$. This occurs when both $v_{i,j,k}$ and $u_{i,j,k}$ are nonzero for some $k \in \{1, ..., q\}$. This is the form of preferential sampling seen in all three toy examples. Accounting for PS under this assumed data generating mechanism requires the joint model to be fit.

The second way that preferential sampling can occur is when the PS can be explained by the available covariates $\mathbf{x}$. This occurs when one or more covariate within $\mathbf{x}$ is shared between the $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$. Formally, this happens when both $\alpha_{1,j}$ and γ_j are nonzero for the true data generating mechanism, for some $j \in \{1, ..., p\}$. When this occurs, but when none of the latent effects are shared in $\boldsymbol{\zeta}$, and hence when $v_{i,j,k} = 0 \;\; \forall (i, j, k)$, then controlling for PS is greatly simplified. Here, unlike with the previous form of preferential sampling, a joint model is not required for controlling for PS. Instead, the offending covariate/s within $\mathbf{x}$ need only be included in their correct functional forms within the model for $\mathbf{Y}$ to control for PS.

As noted earlier, the above statements cover only the discrete-space and continuous-space settings. In point-pattern data, we only have a single source of information associated with each data point, the space-time location $(\mathbf{s}_i, t_i)$. Contrast this with the previous discrete-space and continuous-

space settings. Here, two sources of information are available at each sampled location $(\mathbf{s}_i, t_i)$. First, are the values of the response $y(\mathbf{s}_i, t_i)$. Second, are the locations and times $(\mathbf{s}_i, t_i)$ themselves. Thus, two separate models: one for describing $\mathbf{Y}$, and the other for describing the characteristics of the sampling process through $R_{i,j}$, could be constructed.

Given the availability of only a single source of information, we are unable to construct additional variables $R_{i,j}$. Instead, we are left with needing to construct a suitable model for the thinned point-pattern $\mathbf{Y} = (\mathbf{S}_0, \mathbf{T}_0)$ that can incorporate PS. To achieve this, we can take a model based approach. We may construct a thinning process, denoted $p(\mathbf{s}, t)$, that acts multiplicatively on the intensity of the points. Under this assumed thinned point process model, the marginal intensity of the observed point-pattern takes the form $\lambda_{obs}(\mathbf{s}, t) = \lambda(\mathbf{s}, t)p(\mathbf{s}, t)$ [Fithian et al., 2015]. Thus, instead of building a joint model with two separate likelihoods, we instead fit a joint model with two processes contained within a single LGCP likelihood.

In particular, the model becomes:

$$
\begin{aligned}
Y_{i,j} &\backsim f_Y\left(g(\mu_{i,j})h(m_{i,j}), \boldsymbol{\theta}_Y\right), \;\; f_Y \backsim \text{density} \\
g(\mu_{i,j}) = \eta_{i,j} &= \mathbf{x}_{i,j}^T\boldsymbol{\gamma} + \sum_{k=1}^{q} u_{i,j,k}\beta_k(\mathbf{s}_i, t_j) \\
h(m_{i,j}) = \zeta_{i,j} &= \mathbf{w}_{i,j}^T\boldsymbol{\alpha}_2 \\
\beta_k(\mathbf{s}_i, t_j) &\backsim N(\mathbf{0}, \Sigma_k(\boldsymbol{\theta_k})) \quad k \in \{1, .., q\} \\
\Theta &= (\boldsymbol{\theta_Y}, \boldsymbol{\gamma}, \boldsymbol{\theta_1}, ..., \boldsymbol{\theta_q}, \boldsymbol{\alpha}_2) \backsim \text{Priors} \\
\mathbf{x}_{i,j} &\in R^p, \;\; \mathbf{u}_{i,j} \in R^q, \;\; \mathbf{w}_{i,j} \in R^r.
\end{aligned}
$$

In the above model, the LGCP likelihood is approximated using the f_Y density. The true intensity $\lambda(\mathbf{s}, t)$ is again modelled through the $g(\mu_{i,j})$ terms, with the thinning process $p(\mathbf{s}, t)$ modelled with the $h(m_{i,j})$ terms. Note how neither the latent effects $\beta_k(\mathbf{s}, t)$, nor the covariates $\mathbf{x}$, are present in the thinning process. This is because the above model is plagued by identifiability issues. In fact, without additional prior knowledge available

on the thinning function, neither the covariates $\mathbf{x}$, nor the latent effects $\beta_k(\mathbf{s},t)$ present in $\boldsymbol{\eta}$ may be shared with the thinning process (see Chapter 5). Note that the above model remains identifiable when there is correlation between $\mathbf{w}$ and $\mathbf{x}$ [Warton et al., 2013]. In Chapter 5, we consider the scenario where additional strong prior knowledge is available.

When the above model truly describes the PS, then the PS may be controlled for by including the covariates $\mathbf{w}$ in their correct functional forms, without any additional knowledge required of the thinning function. Computationally, the simplest choice is to build a log-linear model for both the intensity and the thinning process. One example is to use the Poisson approximation to the LGCP likelihood and to set $h \equiv g \equiv \log$. Under this choice, f_Y is a Poisson density function and the model remains within the STGLMM class. In fact, this leads to a single linear predictor, with the $\mathbf{w}_{i,j}^T \boldsymbol{\alpha}_2$ terms added to each $\eta_{i,j}$.

However, this does not lead to a strictly valid model. The log link function for h can lead to thinning probabilities $p(\mathbf{s},t)$ being predicted outside of the interval $[0,1]$. Regardless, this choice is often made [Fithian et al., 2015]. Alternatively, a more suitable link function (e.g. logit, probit, etc.,) may be specified for h and hence for $p(\mathbf{s},t)$. However, this leads to $\mathbf{w}$ being included nonlinearly within the linear predictor for $\boldsymbol{\eta}$ and thus takes the model outside of the STGLMMs framework. This leads to additional computational challenges, with one proposed solution to linearise the model via Taylor series expansions [Bachl et al., 2019].

Large computational advantages can be attained if we can remain within the class of STGLMMs. One such computational approach that can fit STGLMMs efficiently, called INLA, is summarized next.

2.4.2 INLA

The multivariate Gaussian prior distribution assumed on both the latent effects $\beta_k(\cdot,\cdot)$ and the parameter vectors $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}_1$, and $\boldsymbol{\alpha}_2$ allows the above STGLMM framework to be implemented quickly and accurately using the approximate INLA approach [Rue et al., 2009]. INLA is a novel technique

for approximating posterior distributions. It is based on the Laplace approximation method and is applicable to a wide class of models called latent Gaussian models, of which STGLMMs defined above is a subclass.

Laplace approximation

The basic idea of the Laplace approximation in statistics is to approximate an arbitrary distribution with a multivariate Gaussian distribution, centred at the mode. The mean and covariance matrix of the Gaussian distribution is derived from the 2-term Taylor series expansion of the original distribution about its posterior mode. The covariance is set equal to the inverse of the matrix of second derivatives.

In particular, for arbitrary multivariate density $\pi(\mathbf{x})$ with mode $\tilde{\mathbf{x}}$ and with ∇ denoting the gradient operator and $H(\mathbf{x})$ denoting the Hessian matrix of second derivatives:

$$\begin{aligned}\log(\pi(\mathbf{x})) &\approx \log(\pi(\tilde{\mathbf{x}})) + (\mathbf{x}-\tilde{\mathbf{x}})^T \nabla \log(\pi(\tilde{\mathbf{x}})) + \frac{1}{2}(\mathbf{x}-\tilde{\mathbf{x}})^T \nabla^2 \log(\pi(\tilde{\mathbf{x}}))(\mathbf{x}-\tilde{\mathbf{x}}) \\ &= \log(\pi(\tilde{\mathbf{x}})) + \frac{1}{2}(\mathbf{x}-\tilde{\mathbf{x}})^T H(\tilde{\mathbf{x}})(\mathbf{x}-\tilde{\mathbf{x}}) \qquad \text{gradient 0 at mode.}\end{aligned}$$

Taking the exponential of both sides of the equation:

$$\begin{aligned}\pi(\mathbf{x}) &\approx \pi(\tilde{\mathbf{x}}) \exp\left(\frac{1}{2}(\mathbf{x}-\tilde{\mathbf{x}})^T H(\tilde{\mathbf{x}})(\mathbf{x}-\tilde{\mathbf{x}})\right) \\ &= \tilde{\pi}(\mathbf{x}),\end{aligned} \tag{2.4}$$

we obtain the multivariate normal approximation to $\pi(\mathbf{x})$ with mean $\pi(\tilde{\mathbf{x}})$ and covariance matrix $-H(\tilde{\mathbf{x}})^{-1}$. The accuracy of the approximation depends strongly on the distribution being approximated. Furthermore, the accuracy of the approximation deteriorates as the distance of $\mathbf{x}$ from the mode increases. In the STGLMM setting, the target distribution that we will commonly wish to approximate will be the posterior distribution of the latent effects. When the values of the hyperparameters are known, the

Laplace-approximation method provides a good approximation [Rue et al., 2009].

INLA

In practice, the hyperparameters are rarely known. In general, the above Laplace approximation applied to an arbitrary posterior distribution will have a mode and Hessian matrix that will depend upon the hyperparameters. Furthermore, to compute the posterior marginal distribution of the latent effects in an STGLMM, we will need to integrate over the posterior distribution of the hyperparameters. The INLA approach offers a computationally fast approach for doing this. Let $\mathbf{y}, \boldsymbol{\beta}$, and $\boldsymbol{\theta}$ denote the response, latent effects and hyperparameters respectively. For notational simplicity, we use $\pi(\cdot)$ throughout to denote probability density.

For any given set of hyperparameters $\boldsymbol{\theta}$, the posterior distribution of the latent effects given the response and hyperparameters can be written as follows:

$$\pi\left(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\theta}\right) = \frac{\pi\left(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}\right)}{\pi\left(\boldsymbol{\theta}|\mathbf{y}\right)}.$$

Next, the above equation can be rearranged to provide an expression for the posterior distribution of the hyperparameters, given the response:

$$\pi\left(\boldsymbol{\theta}|\mathbf{y}\right) = \frac{\pi\left(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}\right)}{\pi\left(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\theta}\right)}. \tag{2.5}$$

Finally, the numerator may be rearranged as follows:

$$\begin{aligned}\pi\left(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}\right) &= \frac{\pi\left(\beta,\boldsymbol{\theta},\mathbf{y}\right)}{\pi(\mathbf{y})} \\ &\propto \pi\left(\beta,\boldsymbol{\theta},\mathbf{y}\right) = \pi\left(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\beta}|\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right).\end{aligned}$$

Given that the denominator $\pi(\mathbf{y})$ does not depend on either the unknown

latent effects or hyperparameters, we see that the joint posterior distribution is proportional to the product of three probability density (or probability mass) functions that can each be evaluated exactly. Thus, an approximation to equation 2.5 can be formulated as follows:

$$\pi\left(\boldsymbol{\theta}|\mathbf{y}\right) \approx \frac{\pi\left(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}\right)}{\tilde{\pi}\left(\tilde{\boldsymbol{\beta}}|\mathbf{y},\boldsymbol{\theta}\right)} = \tilde{\pi}\left(\boldsymbol{\theta}|\mathbf{y}\right). \tag{2.6}$$

Here, $\tilde{\pi}\left(\tilde{\boldsymbol{\beta}}|\mathbf{y},\boldsymbol{\theta}\right)$ denotes the Laplace approximation to the posterior of the latent effects, evaluated at the posterior mode $\tilde{\boldsymbol{\beta}}$ and conditional on $\boldsymbol{\theta}$. The posterior mode can be found using standard optimization techniques. In settings where the assumed distribution of the response is Gaussian, the Laplace approximation to the posterior of the latent effects is exact, due to the self-conjugacy property of the Gaussian. As an aside, the INLA approach can be iterated one step further to offer an improved approximation to the posterior of the latent effects. However, this comes at a high computational cost and the gains in accuracy are small [Krainski et al., 2018, Taylor and Diggle, 2014]. Thus, we do not consider this further.

Equation 2.5 may now be used to approximate the posterior marginal distributions of the latent effects, with the uncertainties from the hyperparameters integrated out. The approximate posterior distribution $\tilde{\pi}\left(\boldsymbol{\theta}|\mathbf{y}\right)$ is searched across its support to find the regions of highest probability density. Next, a grid of G integration points $\boldsymbol{\theta}_g : g \in \{1,...,G\}$ is chosen, with each integration point having corresponding volume Δ_g. Then, the following approximation to the integral can be formed:

$$\begin{aligned}
\pi\left(\boldsymbol{\beta}|\mathbf{y}\right) &= \int_{\text{supp}(\boldsymbol{\theta})} \pi\left(\tilde{\boldsymbol{\beta}}|\mathbf{y},\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}|\mathbf{y}\right)\mathrm{d}\boldsymbol{\theta} \\
&\approx \int_{\text{supp}(\boldsymbol{\theta})} \tilde{\pi}\left(\tilde{\boldsymbol{\beta}}|\mathbf{y},\boldsymbol{\theta}\right)\tilde{\pi}\left(\boldsymbol{\theta}|\mathbf{y}\right)\mathrm{d}\boldsymbol{\theta} \\
&\approx \sum_{g=1}^{G} \tilde{\pi}\left(\tilde{\boldsymbol{\beta}}|\mathbf{y},\boldsymbol{\theta}_g\right)\tilde{\pi}\left(\boldsymbol{\theta}_g|\mathbf{y}\right)\Delta_g.
\end{aligned} \tag{2.7}$$

Discussion of Laplace approximations and INLA

In theory, if a set of regularity conditions holds on both the latent effects and the response distribution [Schervish, 2012], including the condition that the dimension of the latent effects is fixed, then the Laplace approximation can become very good in large sample settings due to the Bernstein-von Mises theorem. Simply put, the Bernstein-von Mises theorem says that under a set of regularity conditions, the posterior distribution of the latent effects will asymptotically converge to that of a multivariate normal distribution in distribution [Schervish, 2012]. Thus, the use of the Laplace approximation in conjunction with a sensible method of choosing the hyperparameters (e.g. maximum likelihood) is justified on asymptotic grounds in certain settings.

However, the above standard conditions that justify the use of the Laplace approximation based on asymptotic grounds do not hold in most spatio-temporal settings. Often it is the case that the dimension of the latent effects $\boldsymbol{\beta}$ grows with n, instead of being fixed. Thus, observations rarely accumulate around each location in the domain that index $\boldsymbol{\beta}$ and instead $\dim(\boldsymbol{\beta})/n \approx C$, with C constant. Rue et al. [2009] discusses this and concludes that the accuracy of the Laplace (and INLA) approximations to both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ "seems to be directly related to the 'actual' dimension" of $\boldsymbol{\beta}$ and suggest that the effective number of parameters be calculated to investigate the approximation accuracy.

Furthermore, the approximation accuracy of INLA also depends upon the size of G in (2.7), the method used to set-up the grid points, and whether or not the improved approximation to $\pi\left(\tilde{\boldsymbol{\beta}}|\mathbf{y}, \boldsymbol{\theta}_g\right)$ is used. In practice, the accuracy of the approximation is most sensitive to the choice of method used to set-up the grid points [Krainski et al., 2018]. The INLA approximation is the one we use throughout the remaining chapters of this dissertation, with the actual computational procedures both designed and implemented in the R package *R-INLA* [Lindgren et al., 2011b, Rue et al., 2009, 2017]. The feasibility of the INLA approach is limited to settings where the dimension of $\boldsymbol{\theta}$ is no more than about 15 [Rue et al., 2009]. In simulation studies, the INLA approach has been found to perform comparatively as well as MCMC

methods in LGCP analyses [Taylor and Diggle, 2014].

While INLA is a highly general method suited for fitting STGLMMs quickly, efficiently, and with relative ease, limitations do exist. First, INLA performs inference on the posterior marginal distributions of individual model parameters within (2.7), instead of on the joint posterior distribution of the model parameters [Gómez-Rubio, 2020]. Knowledge of the joint posterior distribution is required to perform inference on functions of multiple parameters (e.g. the pairwise differences between parameters). However, methods developed by Rue et al. [2009] allow for samples to be drawn from an approximate joint posterior distribution of the latent effects and hyperparameters, using a fitted INLA model. These samples then enable approximate joint inference to be performed.

Additionally, by default INLA is unable to fit: models with multiple levels of hierarchy, general mixture models, or models with missing covariates. Recently, a hybrid MCMC-INLA hybrid approach was proposed by Gómez-Rubio and Rue [2018] that allows for all of the above model features to be fit using INLA and can also be used to obtain the joint posterior distribution of a chosen set of parameters. Crucially, only standard MCMC algorithms such as Metropolis-Hastings are typically required, making the development of computer code relatively easy.

2.4.3 Applicability of the methods in practice

We have briefly introduced the field of spatio-temporal statistics and its three branches: discrete-space, continuous-space, and point-pattern data. Next, we presented examples of preferential sampling in all three data settings, and ultimately demonstrated that prediction bias is a common consequence of ignoring the sampling process. Then, we introduced STGLMMs, a flexible framework for modelling all three types of spatio-temporal data. We discussed how the sampling process can be modelled within the STGLMMs class with relative ease. Finally, we introduced the computational approach, INLA, that can fit STGLMMs both quickly and in a memory-efficient manner.

In the following three Chapters, we build upon the methods introduced in this Chapter. Throughout, we demonstrate that: preferential sampling is prevalent in real-world data by demonstrating its presence in three real-world datasets; the magnitude of prediction bias can be large and that preferential sampling should not be ignored; the methods developed in this dissertation can be applied to very large datasets quickly using standard software thanks to the INLA computational approach.

Thus, we hope to leave readers with a clear awareness of the potential deleterious effects that preferential sampling can have on their statistical inference of spatio-temporal data. We aim to convince readers that the problem of preferential sampling must not be ignored within any statistical analysis of spatio-temporal data. We provide readers with a suite of tools for both testing for the presence of preferential sampling in their data, and for subsequently adjusting their inference to its presence.

Chapter 3

A General Theory for Preferential Sampling in Environmental Networks

"close attention should be given to densely populated areas within the region, especially when they are in the vicinity of heavy pollution."
— The United States' EPA Monitoring Network Design QA Handbook Vol II Section 6.0, guidelines for selecting the number and locations of air pollution samplers

A preview

This Chapter presents a general framework for modeling the PS of locations and times chosen to monitor a spatio-temporal process. We focus on environmental spatio-temporal processes, although the framework can be generalized beyond this. The framework is applicable in both the discrete–space and continuous–space settings seen in the previous Chapter (Sections 2.1 and 2.2). As discussed in the previous Chapter (Section 2.4.1), we construct binary indicator variables $R_{i,j}$ for both the discrete–space and continuous–space settings. In the continuous–space setting, we choose to discretize the space into a large set of discrete spatial 'sites' and use the logistic regression

approximation to a LGCP [Warton et al., 2010].

Our framework considers the joint distribution of an environmental process with a site–selection process that considers where and when sites are placed to measure the process. By sharing random effects between the two processes, the joint model is able to establish whether or not site placement was stochastically dependent on the environmental process under study. Furthermore, if stochastic dependence is identified between the two processes, then inferences about the probability distribution of the spatio–temporal process will change, as will predictions made of the process across space and time. The embedding into a spatio–temporal framework also allows for the modelling of the dynamic site–selection process itself. Real world factors affecting both the size and location of the environmental monitoring network can be easily modelled and quantified.

We then apply this framework to a case study involving particulate air pollution over the UK where a major reduction in the size of a monitoring network through time occurred. It is demonstrated that a significant response–biased reduction in the air quality monitoring network occurred, namely the relocation of monitoring sites to locations with the highest pollution levels, and the routine removal of sites at locations with the lowest. We also show that the network was consistently unrepresentative of the levels of particulate matter seen across much of GB throughout the operating life of the network. Finally we show that this preferential sampling of monitoring sites may have led to a severe over–reporting of the population–average exposure levels experienced across GB. This discovery could have great impacts on estimates of the health effects of black smoke levels.

3.1 Introduction to Chapter 3

This Chapter concerns preferential sampling (PS), where the locations of sites selected to monitor a spatio–temporal environmental process $Z_{st},\ s \in \Omega,\ t \in \mathcal{T}$, depend stochastically on the process they are measuring. Thus PS is a special case of response–biased sampling. The space–time point is defined $(\mathbf{s}, t) \in \Omega \times \mathcal{T}$, with Ω denoting the spatial domain of interest and

$\mathcal{T}$ the temporal domain. Purely spatial processes (i.e. when $|\mathcal{T}| = 1$), and purely temporal processes (i.e. when Ω is ignored) are two special cases.

Spatial sampling network designers must specify a set of time points $T \subset \mathcal{T}$ at which to observe Z and at each time $t \in T$, a finite subset of sites $S_t \subset \Omega$ at which to do so. Generally the temporal domain $\mathcal{T}$ would be a finite set as for practical reasons Z must be a time–averaged quantity. The designer may select the network sites in a preferential way to meet specified objectives [Schumacher and Zidek, 1993], although attaining those objectives may present its own challenges [Chang et al., 2007]). Moreover, the suitability of the network for achieving its initial objectives may decline over time as in the case of the air quality monitoring network for Metro Vancouver [Ainslie et al., 2009]. In some cases, the objectives may not be well prescribed in which case evidence suggests that in these cases administrators may select monitoring sites preferentially [Shaddick and Zidek, 2014]. Finally, the data provided by networks for one purpose may be used for another purpose and this mismatch may cause problems. For example, urban air pollution monitoring sites provide the information needed to detect noncompliance with air quality standards [EPA, 2005, Loperfido and Guttorp, 2008]. However, these measured values of Z would tend to overestimate the overall levels of the air pollutant throughout Ω and thus render the data unsuitable for assessing the impacts of Z on human health and welfare. In such cases networks well designed for one purpose may be seen as preferentially sampled when the data they yield are used for another purpose.

A variety of approaches can be taken for modelling PS and mitigating its effects in a spatio–temporal process framework. The choice of framework depends on contexts and purposes. Subsection 3.2.1 reviews some of these approaches along with their associated references. Two different situations are encountered. In what might be called the retrospective approach, all the process data are available for use in assessing and mitigating the impact of PS at any given time $t \leq \max(T)$. Such impacts could, for example, distort estimates of model parameters, spatial predictions, temporal forecasts, trends, and risk assessments. A special case is where $|T| = 1$ and $Z_{sT}, s \in \Omega$ is a random spatial field. Since data are not collected over time, strong

assumptions must be made about the PS process that yields the network of sites. The data cannot be used to build an emulator of the actual selection process itself, since the requisite data are not yet available when the spatial sites are selected. But it might be assumed that the future latent data does reflect the past during the period under which the network was designed.

In the prospective case, the selection of network sites at time $t \in T$ may be based on process observations up to and including time $t-1$. In this case, the propensity to preferentially select sites at time t can be estimated without benefit of having the data for time t. The temporal model can then be sequentially updated at time $t+1$ and the process model could adapt quickly to abrupt changes rather than projecting long term trends.

We develop a general modelling framework for the retrospective case, that enables a researcher to determine if the locations of the monitoring sites that form an operational network have been selected preferentially through time (i.e. if response–biased selection occurred). Furthermore, unlike with the spatial–only data, our framework applied to spatio–temporal data allows for a site–selection process emulator to be developed. The population of all site locations considered for selection at any time $t \in T$ is defined as $\mathcal{P} \subset \Omega$. $\mathcal{P}$ must be specified a–priori, as the model framework does not consider locations outside of the fixed (pre–specified) population $\mathcal{P}$ in the site–selection process. But within that framework both static and mobile monitoring networks are admitted. Importantly, depending on the choice of population $\mathcal{P}$, different insights into the nature of PS can be explored.

Defining the population of sites considered for selection throughout ($\Omega \times \mathcal{T}$) has been an issue of fundamental importance for all previous work on PS. This is especially true for the model framework introduced in this Chapter. Depending on the choice of population, different insights into the nature of PS can be obtained and spatial predictions may change dramatically. We consider two populations in this Chapter, however, more can be thought of and implemented to suit the needs and knowledge of the researchers. In one of the cases considered in this Chapter, that population is considered to consist of all sites that have been deemed worthy of being monitored at some times $t \in T$. We refer to these as the observed sites. In another case,

pseudo–sites are also included uniformly throughout Ω. These pseudo–sites have never been monitored but are considered important for characterizing the field itself and for investigating the impacts of PS. The name pseudo–sites follows from presence-only applications in statistical ecology, where such sites are often referred to as pseudo–zeros [Fithian and Hastie, 2013, Warton et al., 2010]. We opt for the name pseudo–sites to distinguish these locations from the traditional 'data-locations' and 'prediction-locations' terminology used in classical geostatistics. This is because in many applications, not all prediction locations can also be pseudo site locations. For example there may be regions in $A \subset \Omega$ across which we wish to predict the field, yet know with certainty that a site could not have been considered for selection for reasons unrelated to the process being measured. Possible reasons include the presence of a physical barrier (e.g. a mountain range) or a political barrier (e.g. a militarized zone) that could make the placement of a monitoring site impossible. Note that in all cases our population of sites $\mathcal{P}$ is finite. This assumption of a finite population is in contrast to the spatial continuum assumed by point process models, although parallels between the methodologies exist and are discussed at length in this Chapter.

A Bayesian model is introduced for the joint distribution of the response vector (Y_{st}, R_{st}). R_{st} is a binary response for the site–selection process, which is 0 or 1 according to whether or not a monitoring site is absent or present at the space–time point $(\mathbf{s}, t) \in \mathcal{P} \times T$, with $\mathcal{P} \subset \Omega$ a fixed population of site locations under consideration. The resulting model when fitted, identifies the effects of PS if any, on inferences about the population mean of the process underlying Y. For brevity, we denote this population's mean by '$\mathcal{P}$–mean'. By sharing random effects across the two processes, the stochastic dependence (if any) between $Y_{\mathbf{s},t}$ and $R_{\mathbf{s},t}$ can be quantified, and subsequently the model can adjust the space–time predictions according to the nature of PS detected.

Moreover it yields an emulator of the dynamic preferential site–selection process as the operational monitoring network (denoted by S_t) evolves over time. The factors affecting the initial site placements can be allowed to differ from those affecting the retention of existing sites in the network. The

dynamic model allows for an assessment of the degree to which preferentiality is determined not just by stochastic processes underlying Y, but by other factors that might include for example the administrative processes involved in the establishment of a monitoring site. Two examples considered in this Chapter are political affinity for environmental monitoring and budgetary constraints in an attempt to emulate the site–selection process, although more can be hypothesized and included. A key result described in the Chapter is the ability to use the *R-INLA* software package with the SPDE approach Lindgren et al. [2011b], Rue et al. [2009, 2017] to fit the joint distributions proposed in our framework. This ensures inference remains feasible, even for space–time applications with many thousands of pseudo–site locations.

Finally, we fit our model framework to a real case study: a large scale air pollution monitoring network in the UK that monitored black smoke (BS hereafter) levels for more than fifty years. This case study provides an ideal data example for our model since the network underwent a constant, dramatic re–design through time and furthermore, the locations of the observed sites appear to largely under-represent rural regions of Great Britain (GB hereafter). We consider two populations $\mathcal{P}$ of sites. First, we consider $\mathcal{P}_1$ to be the locations at which a site was operational at some $t \in T$ (i.e. observed sites only). Here, we ultimately wish to see the effects of PS, if any, on estimates of the $\mathcal{P}_1$–mean, as well as investigate if the network evolved preferentially. Our second population $\mathcal{P}_2$ includes thousands of uniformly located ('pseudo') sites placed approximately 5km apart from each other throughout GB. Since we uniformly cover GB, from this population we are able to assess if the observed sites were preferentially placed within GB (i.e. Ω), and then preferentially retained in the network. We can then evaluate the effects of PS on the $\mathcal{P}_2$–mean (i.e. the average across GB). These two choices of population help to address two distinct questions.

3.2 Modelling frameworks

This section describes a very general framework in which PS can be explored depending on the purpose of that exploration. It begins in Subsection 3.2.1 with a review of some existing theory.

3.2.1 Review of related work

Most work on PS is set in the geostatistical framework where T consists of a single time point so for expository simplicity we temporarily drop the subscript t in this context. In geostatistics PS has a long history. For example Isaaks and Srivastava [1988] describe the deleterious impact to variogram estimates when " the data locations... are preferentially located in high– or low–valued areas", in particular because the "preferentially clustered data" can lead to a "destructuring" of the variogram. In fact this concern about clustered data goes back to Switzer [1977]. Olea [2007] reviews the history of PS, in particular with respect to the clustering due to it. However interest in this topic has spread to a variety of subject areas (see for example Michalcová et al. [2011], Zoltán et al. [2007]).

Interest in the statistical science community seems to have been sparked by the paper of Diggle et al. [2010] (hereafter DMS). DMS defines the PS of a space–time field succinctly as the property $[Z,S] \neq [Z][S]$. Here Z denotes the spatial field and S the locations. The square bracket notation can be read as the "probability distribution of". DMS notes that when sampling is non–preferential, S can be regarded as fixed; inferences about Z and its distribution can then be based on conditional distributions given S. The authors also note that non–PS differs from "uniform sampling" when for a given sample size, every possible realization of S is equally likely. DMS assumes that conditional on S and the Gaussian process $Z_s,\ s \in S$, the measured values of Z denoted by Y are mutually independent Gaussian random variables with mean $\mu + Z_s$. At the same time, conditional on Z, S is assumed to be an inhomogeneous Poisson point process (IPP) with intensity function $\lambda(s) = \exp\{\alpha + \beta Z_s\},\ s \in \Omega$. The parameter β represents the degree of PS – with $\beta > 0$, implying large values of Z_s are associated with an

increased chance of inclusion of a sample in a local neighbourhood around s in S. As noted by Professor Dawid in his discussion of DMS, this model cannot represent the real site selection process since the network designers would not know anything about Z until the sites had been established and their measured values were available. Thus this model cannot be viewed as a site–selection emulator since perfect knowledge surrounding Z prior to measurement cannot be assumed. Nevertheless in a post–hoc analysis of those data, the PS model can be fitted and so capture the impact of the real selection process on inferences made about Z and its probability distribution.

The IPP model was used subsequent to the publication of DMS by other investigators in a similar way but in a fully Bayesian model for inference. More specifically Gelfand et al. [2012] replaces $\alpha + \beta Z_s$ in DMS's intensity function by (in our notation) $\alpha + \alpha_1^T \mathbf{X}_s$ where $\mathbf{X}$ denotes a vector of observable covariates. This change makes the model more like a possible model for the real process. Note that without the inclusion of the process Z_s inside the linear predictor of the Poisson process model, they assume a missing–at–random missingness mechanism, with no further dependence existing between the site locations and the underlying process Z_s when conditioned on the included covariates $\mathbf{X}_s$. Thus this would no longer be considered PS by our earlier definitions. Pati et al. [2011] also includes the covariate vector and replaces $\alpha + \beta Z_s$ by $\alpha + \alpha_1^T \mathbf{X}_s + \beta \xi_s$ so that the effect of the observable covariates is incorporated in the PS model. The $\{\xi_s\}$ are referred to as a "residual process" and so unlike DMS, these authors are not making PS depend directly on the process Z. A second residual process η is added to the measurement model so conditional on ξ, η, X and S the $\{Y_s\}$ are assumed to be independently distributed with mean $\mu + \alpha_1^T \mathbf{X}_s + \beta \xi_s + \beta_1 \eta_s$. Thus it would seem that in effect that the process model is being represented by $Z_s = \alpha_1^T \mathbf{X}_s + \beta \xi_s + \beta_1 \eta_s$ while the potential PS derives from only a subcomponent of that process.

The need to include covariates (predictors) is well recognized in DMS and its ensuing discussions, so Gelfand et al. [2012] and Pati et al. [2011] are welcome additions to the geostatistical literature on PS. But none of

these models include as we do in this Chapter, residual terms that represent the ill–defined administrative and other processes involved in actual site selection. These terms are not subcomponents of Z and yet the case study presented in this Chapter suggests that these residuals play a significant role in PS. Furthermore the point process model on which the above models are based will not be suitable in all applications such as that in Conn et al. [2017] about mapping species abundance in ecology. That paper presents a general theory for PS where Ω consists of a finite set of points and the response distributions are non–Gaussian to include such things as count data.

3.2.2 A general retrospective modelling framework

In this section we introduce the general model framework and its purpose, before implementing it on a real case study in Section 3.4. First, we carefully define the population of locations $\mathbf{s} \in \mathcal{P} \subset \Omega$ to consider for selection at some or all $t \in T$. The size and placement of this population may substantially affect the resulting inference. In many cases, either the precise locations of all sites under consideration at each $t \in T$ will be known, or there will be a clearly defined population of locations at which interest lies in estimating the space–time field and/or its corresponding population summary statistics. This case is Population 1 ($\mathcal{P}_1$) considered in our later application. For the second population ($\mathcal{P}_2$) used in our later analysis, we consider all possible points $\mathbf{s} \in \Omega$ to be the population.

Computational considerations lead us, for Population 2, to approximate this by the placement of pseudo–sites in a high density regular grid, thus placing a pseudo site approximately every 5km in Ω. This is similar in flavour to the discretized computational lattice used in the log Gaussian Cox Process (LGCP hereafter) approach by DMS [Diggle et al., 2010]. In fact, as the density of pseudo–sites under consideration in Ω increases, the resulting logistic regression likelihood converges towards a (scaled) Poisson point process likelihood. Parameter estimates and their standard errors converge to those from the Poisson point process too. However, the accuracy

of this approximation depends on the density and placement of the pseudo–sites [Fithian and Hastie, 2013, Warton et al., 2010]. We discuss this in depth later. The LGCP idea has also been considered further, but the need to explicitly add a third likelihood to the joint model to capture the retention process in spatio–temporal applications may make this approach less desirable in some scenarios.

Note that the space–time field represented as $Z_{i,t}$ in previous work, is represented in our model framework as a sum of latent random effects. The benefit of this is that each of the components making up the space-time field may be allowed to have a unique influence on the site–selection process. An example of where this could be beneficial is the case where site–selection is driven only by the subset of the latent random effects that act on a particular spatial scale. We let $\mathcal{P}$ denote the set of site locations in the population and define M to be the number of sites (i.e. $M = |\mathcal{P}|$). Note the interpretation of the $\mathcal{P}$–mean differs substantially across these populations. The $\mathcal{P}_1$–mean can be interpreted as the network average, whilst the $\mathcal{P}_2$–mean can be interpreted as the GB–average (the mean of the space–time field across GB).

We let $Y_i(t)$ denote a spatio–temporal observation process (continuous, count, etc.) at site i, that is at location $\mathbf{s}_i \in \mathcal{P} \subset \Omega$, at time $t \in T$. We let $R_i(t)$ denote the random selection indicator for site $\mathbf{s}_i \in P$ at time t, with 1 meaning the site was operational at this time. We let $t_1, ..., t_N$ denote the (finite) N observation times, and let $r_{i,j} \in \{0, 1\}$ denote the realisation of $R_i(t_j)$ for site $\mathbf{s}_i \in P$ at time t_j, $i \in \{1, ..., M\}, j \in \{1, ..., N\}$. The subscript j will act as a pointer to the desired time. Then our general model framework can be written as follows:

$$
\begin{aligned}
(Y_{i,j}|R_{i,j}=1) &\backsim f_Y(g(\mu_{i,j}),\boldsymbol{\theta}_Y), \;\; f_Y \backsim \text{density} \\
g(\mu_{i,j}) = \eta_{i,j} &= \mathbf{x}_{i,j}^T\boldsymbol{\gamma} + \sum_{k=1}^{q_1} u_{i,j,k}\beta_k(\mathbf{s}_i,t_j) \\
R_{i,j} &\backsim \text{Bernoulli}\,(p_{i,j}) \\
h(p_{i,j}) = \nu_{i,j} &= \mathbf{v}_{i,j}^T\boldsymbol{\alpha} + \sum_{l=1}^{q_2} d_l \sum_{k=1}^{q_1} w_{i,j,l,k}\beta_k(\mathbf{s}_i,\phi_{i,l,k}\,(t_j)) + \sum_{m=1}^{q_3} w^{\star}_{i,j,m}\beta^{\star}_m(\mathbf{s}_i,t_j) \\
\beta_k(\mathbf{s}_i,t_j) &\backsim \text{(possibly shared) latent effect with parameters } \boldsymbol{\theta_k} \quad k\in\{1,..,q_1\} \\
\beta^{\star}_m(\mathbf{s}_i,t_j) &\backsim \text{site–selection only latent effect with parameters } \boldsymbol{\theta}^{\star}_m \quad m\in\{1,..,q_3\} \\
\Theta &= \left(\boldsymbol{\theta_Y},\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\theta_1},...,\boldsymbol{\theta_{q_1}},d_1,...,d_{q_2},\boldsymbol{\theta}^{\star}_1,...,\boldsymbol{\theta}^{\star}_{q_3}\right) \backsim \text{Priors}
\end{aligned}
$$

$$
\mathbf{x}_{i,j}\in R^{p_1},\;\; \mathbf{u}_{i,j}\in R^{q_1},\;\; \mathbf{v}_{i,j}\in R^{p_2},\;\; \mathbf{W}_{i,j}\in R^{q_2\times q_1},\;\; \mathbf{w}^{\star T}_{i,j}\in R^{q_3}
$$

The above framework is set up to allow for a large degree of modelling flexibility for spatial, temporal and spatio–temporal applications. Note that the two functions g and h are known as link functions. These relate the expected value of the response to the linear predictor. Popular choices of h for the Bernoulli likelihood are the logit, complementary log-log and probit functions. In our later analysis, we will generate our zeros (or pseudo–sites) with an approximately constant intensity across Ω. Thus in our case the logit link is the suitable choice for link function since it exploits a natural connection between the conditional logistic regression and the loglinear Poisson point process model we are approximating when we condition on the total count [Baddeley et al., 2015].

We now dissect the model term–by–term. Firstly, consider the observation process Y. We allow for any distribution to be chosen as the likelihood for the observation process. This allows a range of different data types (e.g. continuous, count, etc.,) to be modelled, including those that exhibit a range of features such as skewness, heavy tails and/or over-dispersion. In the linear predictor $\eta_{i,j}$, we may include a linear combination of fixed covariates

$\mathbf{x}_{i,j}$ with a linear combination of q_1 latent effects $\beta_k(\mathbf{s}_i, t_j)$. These q_1 random effects can include any combination of spatially–correlated processes (such as Gaussian [Markov] random fields), temporally correlated processes (such as autoregressive terms), spatio–temporal processes and IID random effects. Note that we include the additional fixed covariates $\mathbf{u}_{i,j}$ to allow for spatially–varying coefficient models, as well as both random slopes and/or scaled random effects to be included. The flexibility here allows for areal data to be modelled too, simply by changing the definition of $\mathbf{s}_i$ from being a point to representing a well-defined area.

Next, we consider the site–selection process $R_{i,j}$. As before, in the linear predictor $\nu_{i,j}$, we may include a linear combination of fixed covariates $\mathbf{v}_{i,j}$ with a linear combination of latent effects. This time however, the latent effects appearing in the observation process $Y_{i,j}$ are allowed to exist in the linear predictor of the selection process $R_{i,t}$. This sharing of the latent effects across the two processes allows for stochastic dependence to exist between the two processes and hence enables us to investigate whether we have a missing–not–at–random mechanism. Note that the matrix $\mathbf{W}_{i,j}$ is fixed beforehand, and allows for q_2 linear combinations (possibly scaled by covariates) of the latent effects from the $Y_{i,j}$ process to be copied across. The parameter vector $\mathbf{d}$ determines the degree to which each shared latent effect (or combination of) affects the R process and therefore measures the magnitude and direction of stochastic dependence between the two models term–by–term. We denote this term by $\mathbf{d}$ in recognition of the landmark paper by Diggle et al. [2010]. Finally, as seen in Pati et al. [2011], we allow q_3 latent effects, independent from the $Y_{i,j}$ process to exist in the linear predictor. This allows us to extract as many sources of variation from the site–selection process as possible, reducing the risk of over–estimating the magnitude of the d_l terms, and thus the stochastic dependence between the two processes.

For added flexibility we allow temporal lags in the stochastic dependence. This allows the site–selection process to depend upon the realized values of the latent effects at any arbitrary time in the past, present or future. Thus this framework allows for both proactive and reactive site–selection to oc-

cur. For example, if for a pollution monitoring network, site–selection were desired near immediate sources of pollution (say for exceedance detection), then we may view as reasonable, a model that allows for a dependence between the latent field at the previous time step as a site–selection emulator. In this case, we would select as the temporal lag function, $\phi_{i,l,k}(t_j) = t_{j-1}$. We define this to be reactive selection, where placement depends only on past realisations of the space–time field. Say instead, site placements were desired near areas forecast to increase in industrialisation (and hence pollution emission). Then a model allowing for dependence with future values of the latent process may be suitable. To achieve this we would select $\phi_{i,l,k}(t_j) > t_j$. We define this to be proactive site selection. Models with mixtures of reactive and proactive site selection could also be admitted and fit under this framework since a unique temporal lag function $\phi_{i,l,k}(t)$ is allowed for each latent effect shared between the linear predictors.

Also of interest is the possibility of setting $w_{i,j,l,m} = 0$ for some values of the subscripts to allow for the directions of preferentiality to change through time. For example, the initial placement of the sites might be made in a positively (or negatively) preferential manner but over time the network might be redesigned so that sites were later placed to reduce the bias. To capture this, it would make sense to have a separate PS parameter d estimated for time $t = 1$ and for times $t > 1$ to capture the changing directions of preferentiality through time. This can easily be implemented. Furthermore, we may wish to set $w_{i,j,l,m} = 0$ for certain values of the subscripts to see if the effects of covariates and/or the effects of PS differs between the initial site placement process and the site retention process.

Clearly the above modelling framework has potential for over–fitting and model non–identifiability among others things. Thus careful choice of prior distributions, linear constraints on the latent effects (e.g. sum–to–zero constraints) and exploratory analysis is vital to fully utilize this model framework.

3.3 Case study: the data

Annual concentrations of BS were obtained from the UK National Air Quality Information Archive (www.airquality.co.uk). Set up in 1961, this archive was the world's first coordinated archive of national air pollution monitoring networks. While it was being established, the network increased in size and the initial growth was quite rapid; from 800 sites in 1962, 1159 sites in 1966, to 1275 sites in 1971 (see Fig 3.1). After this initial period the overall size of the network declined due to rationalization and in response to changing levels of air pollution; in 1976 there were 1235 operational sites, 563 in 1986, 225 in 1996, and 65 in 2006.

Site locations (at a 10 m resolution) and annual average concentrations of BS (μgm^{-3}) were obtained from monitoring sites. For the reasons given by Shaddick and Zidek [2014], we restrict ourselves to only the sites operating between April 1966 and March 1996 and with data capture of at least 75%, equivalent to 273 days a year (as stated in the EC directive 80/779/EEC Colls [2002]). The locations of all these sites (i.e the population $\mathcal{P}_1$ considered in this Chapter) can be seen in Fig 3.3. It can be seen immediately that a high density of sites are located near many major industrial cities such as London and the Midlands, with almost no sites located in the relatively sparsely populated north of Scotland.

The decline in concentrations during this time period was most dramatic. Annual recorded network means fell from 80 μgm^{-3} in 1966 to 31 in 1976, 19 in 1986, 9 in 1996 and 5 μgm^{-3} in 2006. Fig 3.2 shows a random sample of site–specific log–transformed annual BS levels. Concentrations of BS were typically highest in areas where the use of coal for domestic heating was relatively widespread, such as in parts of Yorkshire and within large cities.

Along with these large changes in concentrations, the dramatic changes in the size of the network can be seen in Fig 3.1 which shows the number of operational sites with at least 75% (annual) data capture vs. year within the chosen study period. The initial increase in the size of the network can clearly be seen followed by the long–term reduction in the number of sites over time. Also evident is the marked reduction of the network in the early

1980s when there was an almost 50% reduction in the number of sites as the network was reorganized owing to falling urban concentrations. With such a dramatic drop in the size of the network, one must ask how the network reduction was chosen. Fig 3.2 shows a plot of a random sample of 30 sites' (log-transformed) black smoke trajectories. From this plot there appears to be evidence that the sites that remained in the network until the end were those providing the highest measurements. Thus, we can see clear evidence for a response–biased network reduction process (i.e PS).

Thus, we have a dataset that exhibits three interesting features:

1. A high density of monitoring sites near major industrious regions, and hence near potential sources of BS. Conversely, an under-representation of the rural areas of Northern Scotland, Wales and Cornwall (Fig 3.3), and hence areas with low expected BS.

2. A large change in concentrations of BS throughout the period of study, resulting in a rapidly evolving latent spatio–temporal process (Fig 3.2).

3. A network whose size dramatically changes through time (Fig 3.1).

4. A network that underwent a biased redesign through time (Fig 3.2), with the sites providing the smallest BS readings being dropped from the network.

These four features provide the perfect opportunity for the model framework to both detect and attempt to correct for the effects of PS made within the network. In particular, depending on our choice of $\mathcal{P}$, we are investigating whether or not informative dropout/inclusion occurred in the operational network S_t through time, and/or whether the network of observed sites is representative of Great Britain (GB) as a whole.

Note that the same exploratory analysis was conducted as in Shaddick and Zidek [2014], and a quadratic temporal effect was found suitable to both fit the data and also provide a non–complex relationship to explain the observed decline in (log transformed) concentrations over time. Variograms were constructed for each year separately and for the average over all years,

both on the original data and on the residuals from the temporal model; a spatial model from the Matérn class seemed an appropriate choice.

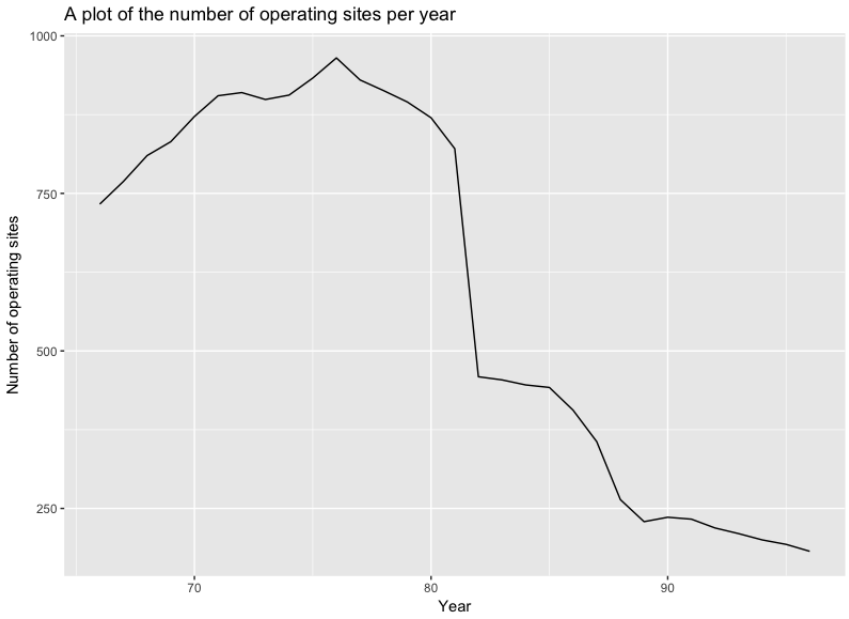

Figure 3.1: A plot showing the number of the monitoring sites that are operational at each year and have data capture of at least 75%. Notice the sharp drop in the size of the network in 1982. Note that a total of 1466 sites were operational at some point in time.

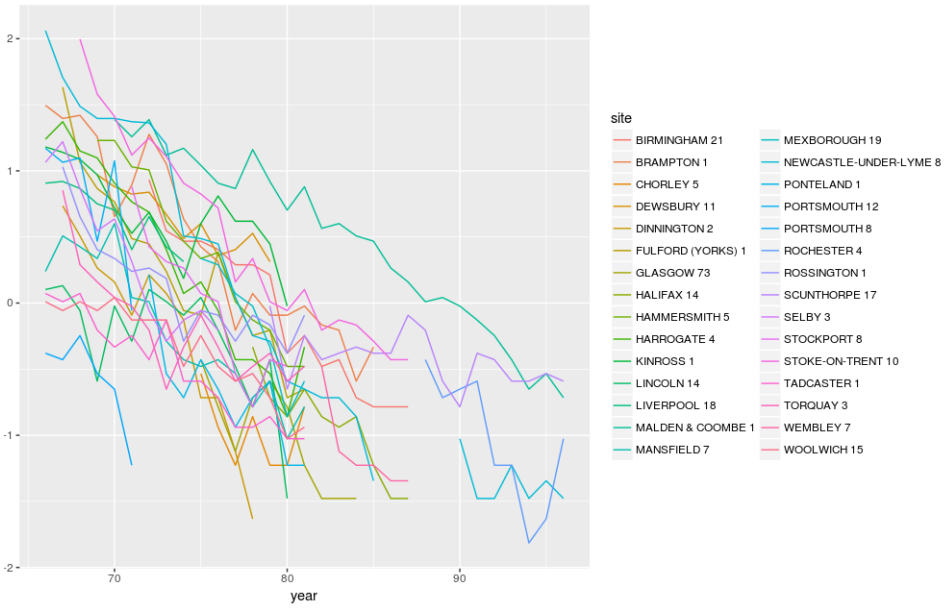

Figure 3.2: A plot showing the mean black smoke level on a log transformed scale for 30 randomly chosen sites. Missing line segments indicate the site was offline that year. Notice that the sites reporting the lowest values tend to be removed from the network earliest.

3.4 Modelling

We build one model from the general framework introduced in Section 3.2. We fit and present the results from three implementations of this model to display the features of the modelling framework. The three implementations are developed through a combination of imposing strict constraints on the PS parameters (i.e. by imposing point mass priors on the $\mathbf{d}$ parameter vector), and changing the population under consideration. These three implementations clearly demonstrate the ability of the model framework to both detect, and adjust for, PS. Furthermore, they highlight the components of the model involved with the PS detection and correction. This helps to demystify the method and to avoid it being seen as a black–box approach.

The joint model developed incorporates the effects of selection by sharing the random effects present in the observation process with the site–selection process. In particular, the selection process is allowed to use information from both spatially varying Gaussian processes and spatially–uncorrelated site–specific effects, to determine the site selection probabilities each year. If PS is detected, then this model should help to de–bias predictions of the $\mathcal{P}_1$ and $\mathcal{P}_2$–means relative to those reported from the raw data, by moving their point predictions against the direction of preferentiality. The magnitude of this movement is dependent upon: the flexibility of the model, the magnitude of the estimated PS parameters d_β, d_b, and the choice of $\mathcal{P}$. This fact is clearly demonstrated by the results from the three implementations.

The same joint model, and computational mesh is used across all three implementations. The differences seen in the results come only from the different assumptions placed upon the site–selection processes and populations. In the first implementation, the site–selection process is forced to be independent from the pollution process through the point mass prior at 0 imposed on d_β, d_b. In other words we constrain the PS parameters to be zero. Consequently the subsequent inference from this model will ultimately be equivalent to the inference from a model without any site–selection process component. In the second and third implementations, we remove this constraint, and two different choices of $\mathcal{P}$ are made to address two alternative scenarios.

All modelling is performed in R-INLA with the SPDE approach [Lindgren et al., 2011a, Rue et al., 2009, 2017]. This enables the rapid computation of approximate Bayesian posterior distributions for both the model variables and latent effect predictions. It does this by approximating the spatio–temporal processes with a Gaussian Markov random field (GMRF) representation by solving an SPDE on a triangulation grid. Details can be found in Lindgren et al. [2011a]. Due to the large size of the dataset and the desired spatial prediction, MCMC approaches without sophisticated approximations or efficient implementation could become infeasible. This is due to the computationally expensive operation of inverting large, dense spatial covariance matrices being required at each MCMC iteration to evaluate the

likelihood. The SPDE approach, by developing a GMRF representation to the spatial fields, only requires the computationally cheaper operations of computing the inverse and the determinants of sparse precision matrices – a task that is made possible with numerical sparse matrix libraries.

3.4.1 Data cleaning

A few data cleaning steps were carried out before fitting the modelling. Due to the right skewness of the black smoke observation distribution, we applied the natural logarithmic transformation to the values to make the observation distribution more Gaussian in shape. Since the natural logarithm is a non–transcendental function, meaning in particular that its series representation contains an infinite series of powers of its argument, we first divided each value by the mean of all the recorded black smoke levels to make the response dimensionless. This ensures not only that the inference remains valid, but also readily interpretable as they are in effect compared to a natural origin [Shaddick and Zidek, 2014]. Next, we scaled the Eastings and Northings coordinates by the standard deviation of the Eastings, and re–scaled the years to lie in the interval [0,1].

3.4.2 Observation process

The following model for the observation process is used for all three implementations seen shortly. The specification follows from Shaddick and Zidek [2014] and is formulated as follows. Let $Y_{i,j}$ denote the observed log black smoke ratio at site i, situated at $\mathbf{s}_i$, at time t_j $i \in \{1, ..., M\}, j \in \{1, ..., N\}$. Let $t_j^\star$ denote the j^{th} time–scaled observations that lie in the interval $[0, 1]$. Let $R_{i,j}$ denote the random selection indicator for site i at time t_j. Let $R_{i,j} = 1$ or 0 depending on whether or not the site was operational in that year and provided the minimum number of readings outlined earlier. Note that there are 1466 sites that record at least one annual reading, and $N = 31$.

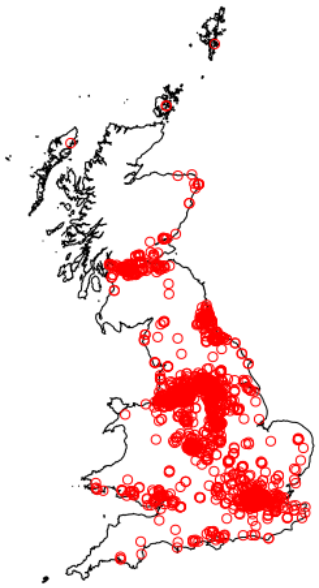

Figure 3.3: A plot of Great Britain, with the locations of the observed sites, and hence $\mathcal{P}_1$ shown. Notice the high density of sites placed in the most population-dense regions.

$$
\begin{aligned}
(Y_{i,j}|R_{i,j}=1) &\backsim \mathrm{N}\left(\mu_{i,j},\sigma_\epsilon^2\right) \\
\mu_{i,j} &= (\gamma_0+b_{0,i}+\beta_0(\mathbf{s}_i))+(\gamma_1+b_{1,i}+\beta_1(\mathbf{s}_i))t_j^\star \\
&\quad +(\gamma_2+\beta_2(\mathbf{s}_i))(t_j^\star)^2 \\
[\beta_k(\mathbf{s}_1),\beta_k(\mathbf{s}_2),...,\beta_k(\mathbf{s}_m)]^T &\backsim^{IID} \mathrm{N}\left(\mathbf{0},\Sigma(\zeta_k)\right) \quad \text{for} \quad k\in\{0,1,2\} \\
[b_{0,i},b_{1,i}] &\backsim^{IID} \mathrm{N}\left(\mathbf{0},\Sigma_b\right) \quad \Sigma_b=\begin{bmatrix}\sigma_{b,1}^2 & \rho_b \\ \rho_b & \sigma_{b,2}^2\end{bmatrix} \\
\Sigma(\zeta_k) &= \text{Matérn}\left(\zeta_k\right) \\
\theta=\left(\sigma_\epsilon^2,\gamma,\zeta_k,\sigma_{b,1}^2,\rho_b\right) &\backsim \text{Priors.}
\end{aligned}
$$

The choice of the observation process model is explained as follows. The

sources of variation can be broken up into three components: global variation, independent site–specific variation and smooth spatially correlated variation. To ensure model identifiability, we enforced sum–to–zero constraints on all random effects (β and b), and furthermore we did not estimate spatially–uncorrelated random effects b at locations with no observations. Note that in the notation of Section 3.2, the b and $\beta_k(\mathbf{s}_i)$ terms are examples of the $\beta(\mathbf{s}, t)$ latent effects and thus $q_1 = 5$. For readability we choose to separate the notation for these effects. Note that, whilst the b terms are assumed independent between sites, the terms $b_{0,i}, b_{1,i}$ are assumed a-priori to be a realisation from a (possibly-correlated) multivariate Gaussian distribution with covariance matrix Σ_b.

The global temporal trend is captured by the γ_k terms since these parameters remain constant across the sites. As in Shaddick and Zidek [2014], when comparing various models for the first (non-joint) implementation, more complex temporal relationships (such as splines) were not favoured by multiple model selection criteria including DIC. Secondly, the independent site–specific variations are captured by the IID random intercepts and random slopes ($b_{0,i}, b_{1,i}$). In geostatistical terms, the b terms act as nugget effects for their corresponding $\beta_k(\mathbf{s})$ terms. The (nugget–free) $\beta_k(\mathbf{s})$ terms then capture the smooth spatially–correlated variation. Models without the b terms showed large residual site–specific errors. Thus it appears that small–scale factors may be a large source of variability in the measured black smoke trajectories, independent from the regional location alone. Note that separate spatially–correlated Gaussian fields for each year were tested (i.e using a separate $\beta_{0,j}(\mathbf{s})$ field for each year), but did not improve the model fit.

The intuition behind the short scale b terms in the model is as follows. An observation tower close to a large source of black smoke (e.g. a road, a polluting factory, or a power station) would likely yield a much higher annual reading than placing it say half a kilometer away from such a source. Since this spatial scale is much smaller than that captured by the $\beta_k(\mathbf{s})$ processes, these differences will not be accounted for without either including covariates that capture the causes of these effects (e.g. distance from the

nearest pollutant source), or by allowing each site to have it's own deviation from the smoothly predicted field via either a fixed or random, site–specific effect. Note that spatially–uncorrelated random quadratic slopes $b_{2,i}$ were not found to improve the model fit with respect to DIC under the first implementation and actually led to a large instability in the predictions of sites that took fewer measurements. It appears that the inclusion of these terms led to some over–fitting.

The choice of priors for the hyperparameters θ were made to make them as weakly informative as possible and hence to reduce their effects upon the posterior results, but also to bound their values inside sensible limits. Despite the fact that previous analyses have been made on this dataset, we only use vague information from these results when constructing the priors. We discuss the details of the chosen priors in the supporting material found in Appendix A.1.

3.4.3 Site–selection process

The following model for the site–selection process is used for all three implementations with the aim of emulating the complex decision–making processes that occurred when setting up the monitoring network. Let: $R_{i,j}$ denote the random selection indicator for site i at time t_j; Let $R_{i,j} = 1$ or 0 depending on whether or not the site was operational in that year and provided that the minimum number of readings outlined earlier is attained. Let $r_{i,j} \in \{0,1\}$ denote the realisation of $R_{i,j}$ for site i at time $t_j^\star$, $i \in \{1,...,M\}, j \in \{1,...,N\}$. Finally, $\mathbf{s}_i$ denotes the location (the scaled Eastings and Northings coordinates) of site i. The model is then:

$$
\begin{aligned}
R_{i,j} &\backsim \text{Bernoulli}\left(p_{i,j}\right) \\
\text{logit}p_{i,1} &= \alpha_{0,0} + \alpha_1 t_1^\star + \alpha_2 (t_1^\star)^2 + \beta_1^\star(t_1) \\
&\quad + \alpha_{rep} I_{i,2} + \beta_0^\star(\mathbf{s}_i) \\
&\quad + d_b \left[b_{0,i} + b_{1,i}(t_1^\star)\right] \\
&\quad + d_\beta \left[\beta_0(\mathbf{s}_i) + \beta_1(\mathbf{s}_i)(t_1^\star) + \beta_2(\mathbf{s}_i)(t_1^\star)^2\right] \\
\text{for } \; j \neq 1 \quad \text{logit}p_{i,j} &= \alpha_{0,1} + \alpha_1 t_j^\star + \alpha_2 (t_j^\star)^2 + \beta_1^\star(t_j) \\
&\quad + \alpha_{ret} r_{i,(j-1)} + \alpha_{rep} I_{i,j} + \beta_0^\star(\mathbf{s}_i) \\
&\quad + d_b \left[b_{0,i} + b_{1,i}(t_{j-1}^\star)\right] \\
&\quad + d_\beta \left[\beta_0(\mathbf{s}_i) + \beta_1(\mathbf{s}_i)(t_{j-1}^\star) + \beta_2(\mathbf{s}_i)(t_{j-1}^\star)^2\right] \\
I_{i,j} &= I\left[\left(\sum_{l \neq i} r_{l,j-1} I\left(||s_i - s_l|| < c\right)\right) > 0\right] \\
\left[\beta_0^\star(\mathbf{s}_1), ..., \beta_0^\star(\mathbf{s}_m)\right]^T &\backsim \text{N}\left(\mathbf{0}, \Sigma(\zeta_R)\right] \\
\Sigma(\zeta_R) &= \text{Matérn}\left(\zeta_R\right) \\
\left[\beta_1^\star(t_1), ..., \beta_1^\star(t_T)\right]^T &\backsim \text{AR1}\left(\rho_a, \sigma_a^2\right) \\
\theta_R = \left[\alpha, d_b, d_\beta, \rho_a, \sigma_a^2, \zeta_R\right] &\backsim \text{Priors.}
\end{aligned}
$$

The first rows of the linear predictors comprise the global effects of time on the log odds (and thus eventually the probability) of selection. We allow for a quadratically changing global log odds of selection with time, and allow for a global first–order autoregressive deviation from this quadratic change (denoted by $\beta_1^\star(t_j)$). This term represents the change in time of both the political and public moods regarding the need for maintaining the overall network size. New governments may well prioritize public spending on the environment in different ways and furthermore, the public's approval of environmental spending likely changes in light of new knowledge. Additionally, large changes in the size of the public monitoring network can be

seen around 1982 (see Fig 3.1). Here a sharp decrease in the size of the network occurred, reducing the number of sites by almost half. The smooth quadratic effect of time clearly would not suffice to capture this short term trend and thus a random effect seems compelling, especially one that is able to adequately capture this short term change (i.e. overdispersion), such as the autoregressive term we used.

The second rows of the linear predictors represent the site–specific factors influencing the log odds ratio in favour of a site's inclusion in the network S_j at time t_j. Firstly, α_{ret} represents what we call the "retention effect". This term reflects how the probability a site is selected in a given year, changes conditional upon its inclusion in the network in the previous year. Since large costs can be incurred in setting up monitoring sites at new locations, it is plausible that network designers would favour the maintenance of existing sites over their replacement at new site locations, even if the conditions at other sites (represented by the other terms in the linear predictor) are more favourable. In fact, it is this indicator variable that determines whether or not the linear predictor corresponds to the site-placement process or the site-retention process. If we wanted to investigate the possibility that the effects of PS or covariates were different between the two processes, then we could include additional product terms between the various effects and $r_{i,j-1}$ to capture this change. Here, we share all parameters across the two processes and allow only a unique intercept to exist between the processes. This is discussed in depth later.

In contrast, α_{rep} captures the repulsion effect. $I_{i,j}$ denotes an indicator variable that determines whether or not another site in the network placed within a distance c from site i was operational at the previous time t_{j-1}. Plausibly network designers would not want to place sites close to an existing site. Conversely, there may be unmeasured regional confounders affecting the localized site–selection probabilities (e.g. population density) that may lead to additional clustering that cannot be explained by the model without the inclusion of the confounder. This parameter should help to capture any additional clustering that may be present. We choose the hyperparameter c to be 10km.

Finally, there may be a larger motivation to place more/fewer sites in certain areas of the UK throughout T, that cannot be explained by the other terms in the model. This may be due to population density or due to increased/decreased political incentives in this area. We attempt to capture such spatially–varying area effects in the $\beta_0^\star(\mathbf{s})$ field. This can be viewed as a spatially–correlated correction field similar to that used by Pati et al. [2011]. Note that this is fixed in time with the aim of avoiding identifiability issues.

Whilst it may appear that we have included a lot of effects in the site–selection process, it is of paramount importance to adequately capture and remove as many sources of variability from the site–selection process as possible. The preferentiality parameters should therefore only act upon the residual signal, after such effects have been removed. Since we are dealing with a large quantity of spatio-temporal data, we are able to learn the temporal features affecting site–selection and thus we can attempt to emulate the true process itself. This is in stark contrast with the spatial setting. By removing large sources of variability from the site–selection process first, we reduce the risk of over–estimating the stochastic dependence between the selection and observation processes and hence reduce the risk of over-adjusting our parameter estimates and predictions.

The third and final rows of the linear predictor represent the preferentiality parameters of the selection process, following the work of Diggle et al. [2010]. We decide to separate the preferentiality into two sources: small–scale deviations from the localized average black smoke levels, and the medium–scale regional deviations from the UK–wide annual black smoke levels. In recognition of the landmark paper by Diggle, we denote the two parameters by d_b, d_β respectively. Since we have constrained both the $[b_{0,i}, b_{1,i}]$ terms and the $\beta_k(\mathbf{s})$ processes to sum to zero, the terms being multiplied by d_b, d_β represent deviations from the $\mathcal{P}$–mean. Both of these effects are allowed to affect site selection independently. The interpretation of these PS parameters depends largely upon the choice of the population $\mathcal{P}$. All PS effects detected are after controlling for the other site–selection effects.

In consideration of the discussions following Diggle et al. [2010], for $j > 1$

site selections made at time t_j involve estimated black smoke levels based on observations made at the previous time t_{j-1}. Thus in our model we do not assume the network designers formulate site selection decisions based on black smoke forecasts into the future or for the current unobserved year, but on predicted quantities at the previous time step. Therefore in our framework, we model the site–selection as being reactive for times $t_j : j > 1$. Using the notation from 2.2, $\phi_{i,l,k}(t_j) = t_{j-1} \forall i, l, k$ and $t_j > 1$. If the true selection mechanism is believed to be different, then the change of paradigm is trivial. For computational savings, we base the site selection at time 1 to be based on the estimated field at time 1 (i.e. $\phi_{i,l,k}(t_1) = t_1$). Our choice of priors are discussed in depth in the supporting material found in Appendix A.1.

3.4.4 Three implementations

For Implementation 1 we constrain the PS parameters d_b, d_β to equal 0. Thus Implementation 1 incorporates the prior assumption that no stochastic dependence between the site–selection process and the observation process was present and thus that no PS occurred. A direct result of this independence assumption is that the posterior distribution of the observation process Y is the same, regardless of the specification of either the site–selection model terms, or the choice of the population of sites $\mathcal{P}$ to consider for selection. Thus, the results from Implementation 1 will match those of a typical spatio–temporal analysis that ignores site–selection. This will be used as our baseline for comparison.

For Implementation 2, we remove the zero constraints on the PS parameters, imposing instead weakly informative Gaussian priors with mean 0 and variance 10. For Implementation 2, we consider only the 1466 observed site-locations for selection at each time $t \in T$. We define this as Population 1, $\mathcal{P}_1$ and thus $M = |\mathcal{P}_1| = 1466$. Population 1 is shown as the red circles in Fig 3.3.

For Implementation 3, we replace the zero constraints with the same Gaussian priors, but consider a different population of sites for selection

at each year, $\mathcal{P}_2$. For $\mathcal{P}_2$ thousands of pseudo–sites are also considered for selection at each time step along with the observed sites from $\mathcal{P}_1$. We ensure the locations of the pseudo–sites are uniformly distributed throughout Great Britain (GB) and placed with high density. It has been shown that estimates and corresponding standard errors of all (non-intercept) parameters converge toward those of the equivalent IPP as the number of pseudo–sites tends towards infinity, so long as the density of the points is uniform (in probability) [Fithian and Hastie, 2013, Warton et al., 2010]. Thus there is some duality with the approach of DMS [Diggle et al., 2010] and our Implementation 3. The locations of $\mathcal{P}_2$ are shown in Fig 3.4.

For Implementation 2 we aim to see if the network evolved preferentially. That is, out of the observed sites, were sites added and dropped from the network in a manner that was dependent upon the value of the latent black smoke process and hence missing not at random (MNAR). Under Population 1, since we do not consider locations within the unsampled regions for selection, no additional information is being added to the unsampled regions. Hence we do not expect the estimates of BS to change much at these locations unless estimates of the site–trajectories and hence the $\mathcal{P}_1$–mean change. Furthermore, we are unsure if the joint model will substantially adjust estimates of the $\mathcal{P}_1$–mean, even if PS is detected. This is since results from a small simulation study we conducted suggest that if we have a case where we fit an inflexible temporal model to a dataset whose sites have a long average consecutive lifetime, estimates will remain largely the same due to the over–determined nature of the problem. In fact, the sites in the dataset provide an average of 12 consecutive years of readings, with the minimum consecutive lifetime of a site being 6 years. Additionally the deviation from the quadratic trend is typically small (Fig 3.2). Thus we may expect only a small change to the results seen from Implementation 1.

For Implementation 3 we investigate if the network of operational sites at each time $S_t : t \in T$ is being located throughout GB (Ω) in a preferential manner. Thus the interpretation of preferential (i.e. response–biased) network evolution is lost under this choice of population. Instead, these PS parameters d_β, d_b now measure the degree to which the operational network

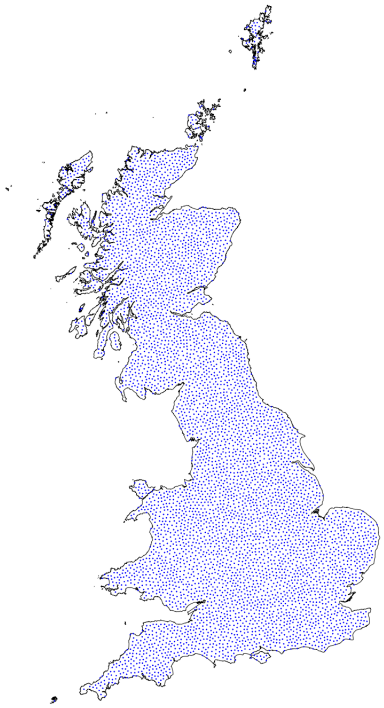

Figure 3.4: A plot of the locations of all sites considered for selection in Population 2. The locations are shown as blue dots, many of which are in regions of low human population density.

(S_t) is preferentially located in Ω through time $\mathcal{T}$. This is due to our second population $\mathcal{P}_2$ covering Ω uniformly and hence considering each point $\mathbf{s} \in \Omega$ as being equally likely to be sampled a-priori. This is unlike Population 1, which did not include large areas of unsampled Scotland, Wales and Cornwall for selection at each time $t \in T$. Thus Population 2, by adding additional information to the unsampled regions via the site–selection process, should inform the joint model about the appropriate adjustment of BS estimates in the unsampled regions according to the nature of PS detected.

Put differently, the joint model will extrapolate any associations detected between the site–selection process and the underlying latent effects into the unsampled regions.

In fact, hidden away in the details of Implementation 3 is the fact that the Bernoulli random variable models two processes simultaneously. Implementation 3 can be considered as being a joint model with three processes: an observation process, an initial site–placement process and a site–retention process. The latter two are fit using only one Bernoulli likelihood. The initial site–placement process is fit using a conditional logistic regression approximation to a log-Gaussian Cox process, and is similar to that seen in Diggle et al. [2010]. The site–retention process is modeled as a Bernoulli random variable. Inside the linear predictor of the Bernoulli likelihood, the indicator variable $r_{i,(j-1)}$ points the linear predictor towards the site–placement process when it is equal to 0 or towards the site–retention process when it is equal to 1. In our example, we only allow for a unique intercept to exist across the two processes, sharing the remaining parameters. Thus we assume that the effects of all the covariates and the effects of PS are constant across the two processes. This assumption can be relaxed by including interaction effects between $r_{i,(j-1)}$ and the other parameters, including the PS parameters.

Note that care is required to ensure that only the pseudo–sites contribute a zero to the Bernoulli likelihood for the site–placement process across all years. Furthermore, for our application, we must ensure that only the sites that have been removed from the network in year j contribute a zero to the Bernoulli likelihood for the site–retention process at year j. This ensures that no site in the network was ever re-installed after its removal, a fact seen in our data. Clearly then, the choice of zeros here is application–dependent. Additional details are given in the supporting material found in Appendix A.1.

The ability of our joint model to adjust estimates of the pollution process at a point $\mathbf{s}$ depends upon the distance of the point from the nearest monitoring site in the network. For pseudo–sites further from an observed site than the effective range of the spatially varying β processes, essentially all

the degrees–of–freedom of the spatially–varying quadratic terms $\beta_k(\mathbf{s})$ are available for use in fitting the site–selection process to make the posterior probability of repeated non-selections (i.e. the $r_{i,j} = 0$'s) of the pseudo–site high. Since we have no black smoke observations here, the fitting of the quadratic slopes to these pseudo–sites is therefore an under–determined problem. Thus we would expect the estimates of black smoke here to be different. For pseudo–sites very close to an observed site (i.e. well within the effective range), we would expect the estimates at the pseudo–site locations to remain largely unchanged, since the problem remains over–determined. For pseudo–sites within the effective range of, but not immediately next to an observed site, we expect estimates to change moderately since the problem is weakly-determined.

3.4.5 Model identifiability issues

When fitting a model this large, issues around model identifiability commonly arise, namely the possibility of the data providing information about the model parameter values through the likelihood. We assessed these issues with two approaches. First, we enforced sum–to–zero constraints on all the random effects to ensure they are simply localized deviations about a global trend. As discussed in the supporting material found in Appendix A.1, we placed penalized complexity priors [Fuglstad et al., 2019, Simpson et al., 2017] on the Matérn parameters of the Gaussian processes to provide some prior information on the range and scale, while reducing the possibility of overfitting the data.

To confirm that we had fully resolved the model identifiability issues, we then conducted a small simulation study. We sampled the data from various models similar in form to the joint model introduced in sections 3.4.2 and 4.3 to see if the posterior estimates of both the parameters and the space–time field covered the true values. Interestingly, for a much smaller dataset, we found no identifiability issues except for the range parameter on the $\beta_0^\star(\mathbf{s})$ process. Here the mean squared error of the point estimates of this parameter were very high relative to the other parameters, although

the nominal coverage levels and bias remained good. This could be a sign of identifiability issues surrounding this effect, or perhaps could be due to the difficulty with estimating a Matérn field using only small amounts of binary point data. All other parameter estimates in the simulation studies, as well as posterior predictions were good. Of most interest was the model's capability to detect the preferentiality parameters d_β, d_b with high precision, negligible bias and with posterior credibile intervals attaining nominal coverage levels.

Interestingly, we experience the same difficulties with identifying the $\beta_0^\star(\mathbf{s})$ process in our case study. Our estimated marginal distribution for the range parameter of the $\beta_0^\star(\mathbf{s})$ process in the UK black smoke case study was found to have a 95% posterior credible interval of (0.03, 1.18). Given that we scaled the coordinates, these distances imply that the model encountered difficulties with estimating this parameter. Importantly, the posterior means of the standard deviation of this effect were around 0.03 with 95% credible intervals lying in the region of between 0.00 and 0.08. Thus, ultimately this effect has minimal impact upon the model fit.

We also assessed the ability of the joint model framework, under simulated PS settings, to de-bias estimates of site–specific trajectories and network averages (equivalent to the $\mathcal{P}_1$–mean). Two such simulation studies considered distinct temporal trends. The first fixed the temporal component to be rigid, the second allowed for a flexible nonlinear trend. In particular, we witnessed that under a rigid (spatially-varying) linear slopes model, when the average of the consecutive lifetimes of the sites is high, the bias induced in the site–specific estimates and the $\mathcal{P}_1$–mean that occurs from ignoring the site–selection process is almost zero. This is due to the problem being over-determined – only a few observations of the process at each site are required for the model to accurately forecast/backcast estimates throughout T. This is similar to what is seen in the case of the UK black smoke dataset. Conversely, when the temporal trend is highly nonlinear and the average consecutive lifetimes of the sites are short, the biases in parameter estimates, site–specific predictions and estimates of the $\mathcal{P}_1$–mean through time can all be high if we ignore the site–selection process. To provide

a 'highly nonlinear' trend, we opted to use an independent realisation of a Matérn field for each of the 30 simulated 'years'. The insights from these two scenarios help explain the results seen shortly in Implementation 2. They also hint that changes to inference under Implementation 2 $\mathcal{P}_1$ would be highest for applications with mobile monitoring sites.

3.5 Results

We focus our attention upon the following issues and objectives:

1. Do implementations 2 and/or 3 detect that, within the network of observed sites (i.e. Population 1), the sites have been preferentially added and removed even after controlling for the various covariates included in the site–selection process? If so, has this been done based upon short–range, site–specific deviations from the regional mean black smoke, and/or medium-range regional deviations from the annual $\mathcal{P}$– mean?

2. When considering Implementation 3, does the model detect that the network of operational sites S_t have been preferentially located within GB (Ω) through time, even after controlling for the various covariates included in the site–selection process?

3. Do estimates of the black smoke annual means in GB (i.e. the $\mathcal{P}_2$–mean) change significantly when we consider the stochastic dependence between the placement of the sites and the black smoke field?

4. If we backcast and/or forecast the predictions at all observed site locations (i.e. $\mathbf{s} \in \mathcal{P}_1$) at all times, how do the estimated black smoke levels differ between the operational (S_t) and offline sites (S_t^C)? Do these differences change in time, and if so, does the apparent priority of site placement change through time?

5. Given the original purpose of the air quality network for monitoring the progress achieved by the Clean Air Act in reducing the population exposure levels to both black smoke and sulphur dioxide [McMillan and Murphy, 2017], if we average the estimated black smoke field across

Great Britain's population, do the estimated population-average exposure levels change between the implementations?

6. Considering the 1980 EU black smoke guide value of 34 μgm^{-3}, how does the estimated proportion of GB exceeding this value change through time? What are the differences across the three implementations? Furthermore, how do estimates of the proportion of the population exposed to BS levels above this value change under the three implementations?

Parameter	Implementation 1	Implementation 2	Implementation 3
d_γ	0 (0)	0.62 (0.17)	2.77 (0.01)
d_b	0 (0)	0.06 (0.04)	0.12 (0.01)
β_0	96.50	94.94	21.87
(trans scale)	1.15 (0.02)	1.13 (0.01)	-0.34 (0.09)
ρ_b	-0.77 (0.02)	-0.76 (0.02)	-0.78 (0.00)
α_{ret}	-	6.18 (0.06)	6.47 (0.06)
α_{rep}	-	0.08 (0.11)	0.82 (0.10)

Table 3.1: A table showing the posterior mean and standard deviations for parameter estimates for the three implementations. Note that the top row estimates of β_0 have been transformed back onto the original data scale.

3.5.1 Implementation 1 – assuming independence between Y and R

If we assume independence between Y and R, the posterior results about the observation process Y from Implementation 1 are identical to those that would have been discovered from fitting only the observation process (i.e. fitting only the Y model). As expected, especially high values of black smoke are predicted to exist around the North West and Yorkshire areas of England in 1966. This area covers the major cities of Liverpool, Manchester, Leeds and Sheffield, all industry–heavy cities at the time under study. By 1996 the relative levels of black smoke in these areas are far reduced and

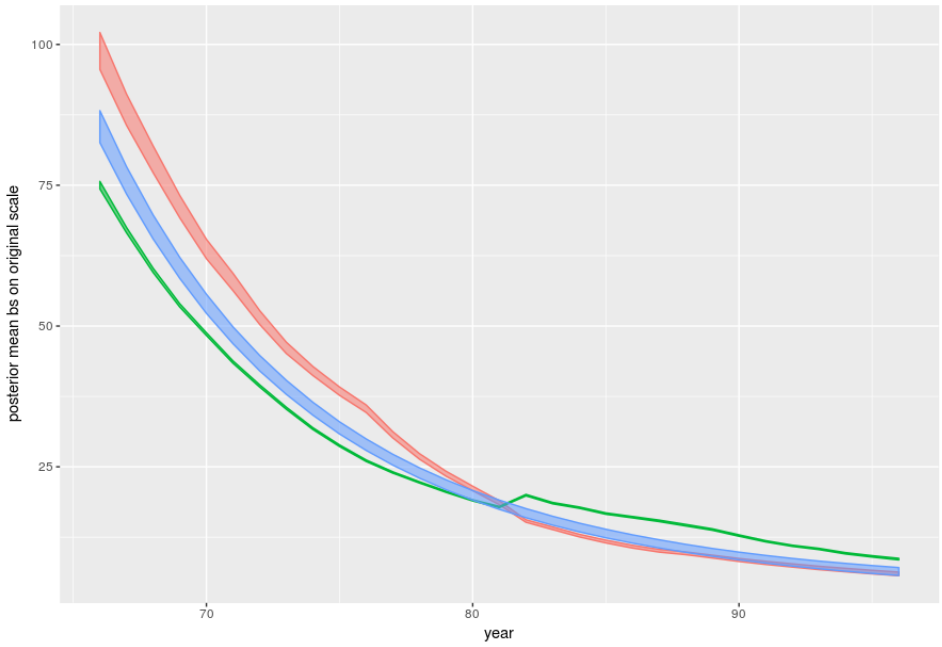

Figure 3.5: Implementation 1. In green are the model–estimated BS levels averaged over sites that were selected in $\mathcal{P}_1$ (i.e. operational) at time t. In contrast, those in red are the model–estimated BS levels averaged over sites that were not selected in $\mathcal{P}_1$ (i.e. offline) at time t. Finally, in blue are the model–estimated BS levels averaged across Great Britain. Also included with the posterior mean values are their 95% posterior credible intervals. If printed in black-and-white, the green band is initially the lower line, the red band is the upper line and the blue band is initially the middle line. Notice the change in the ordering of the values that occurs in 1982.

exceeded by the Greater London area. Counter–intuitively however, the estimated black smoke levels in the Scottish Highlands, an area with almost no manufacturing or industry are predicted to be relatively high (see Fig A.1 in the Appendix) across all time periods. This result is a direct consequence of the absence of monitoring sites in this area (see Fig 3.3), along with a lack of informative covariates included in the observation process Y for this region.

A typical location in the unsampled regions of the Scottish Highlands, Cornwall or The Borders (the Northernmost and South-Westernmost regions of Fig 3.3) sees their distance to the nearest site in $\mathcal{P}_1$ typically exceeding the estimated spatial ranges of the random fields. Consequently, model–estimates in such areas essentially equal the average of the observed pollution levels (i.e. the $\mathcal{P}_1$–mean). This feature can immediately be seen to be problematic since it is likely that the true black smoke levels will be below the $\mathcal{P}_1$–mean in these regions. As well, large standard errors (i.e. posterior pointwise standard deviations) for the predicted black smoke levels are found in these regions due to their lack of monitoring sites (see Fig A.1).

Next, we consider the model–estimated black smoke levels for all the observed site locations (i.e. Population 1) in Fig 3.5 at every time point. To investigate Objective 4, for each $t \in T$ we split the observed sites into the operational sites S_t and offline sites S_t^C. The set of operational sites S_t are defined to be the sites in Population 1 that recorded the minimum number of observations that year. The set of offline sites S_t^C are defined to be the sites in Population 1 that failed to record this minimum number of observations that year. Note that $S_t \bigcup S_t^C = \mathcal{P}_1$ and $S_t \bigcap S_t^C = \emptyset$.

Here we can see that from Implementation 1, that it appears the sites were initially placed in regions with below–average black smoke levels between 1966 – 1980 (see Fig 3.5). This is inferred from the posterior mean black smoke levels – they are significantly lower for the operational sites compared with the estimated GB–average. The lack of additional information for the unsampled regions of GB makes the estimates in these areas equal to the $\mathcal{P}_1$–mean and thus the GB–average is nearly identical to the $\mathcal{P}_1$–mean. Over time, the posterior means for the black smoke levels at the

operational and offline sites converge, before the direction of preferentiality changes in 1982. The latter was the year a major network redesign was initiated, removing almost half of the operational sites (see Fig 3.1). Here we see strong evidence the sites that remained in the network after this redesign were in locations with black smoke levels above the $\mathcal{P}_1$–mean. This is due to the posterior mean black smoke levels being significantly higher for the operational sites compared with the offline sites.

Thus from looking at the results from Implementation 1 alone, we gain some insight about Issues 1 and 4. It appears that the sites were preferentially sampled in almost all time periods. Initially the operational sites appear to have been placed in regions with black smoke levels below the $\mathcal{P}_1$–mean, before being placed in regions with levels above the $\mathcal{P}_1$–mean after the major network redesign in 1982. These results are significant with respect to 95% credible intervals. However, we still have doubts about the predicted black smoke levels in regions of GB known to have little industry or population density – two major sources of black smoke. Since these regions cover large percentages of the surface area of GB, the effect of over–estimating the predictions in these areas would be a marked increase in the estimated GB–average black smoke level. Implementation 3 attempts to rectify this problem by extending the definition of $\mathcal{P}$ into these regions.

3.5.2 Implementation 2 – $\mathcal{P}_1$

Firstly, we consider the posterior parameter estimates for the two sources of preferentiality (see Table 1). These are denoted by d_β, d_b, the medium–range and short–range preferentialities respectively. Only the former effect d_β was detected to be significantly nonzero with a posterior estimated value of 0.66 and a 95% posterior credible interval of (0.34, 0.99). The posterior estimate of the short–range preferentiality was 0.06 with a 95% posterior credible interval of (-0.01, 0.15). Thus in both cases the direction of preferentiality was positive, suggesting that year–by–year, the site placements are positively associated with the relative levels of black smoke at the site location, especially with the regional–average level.

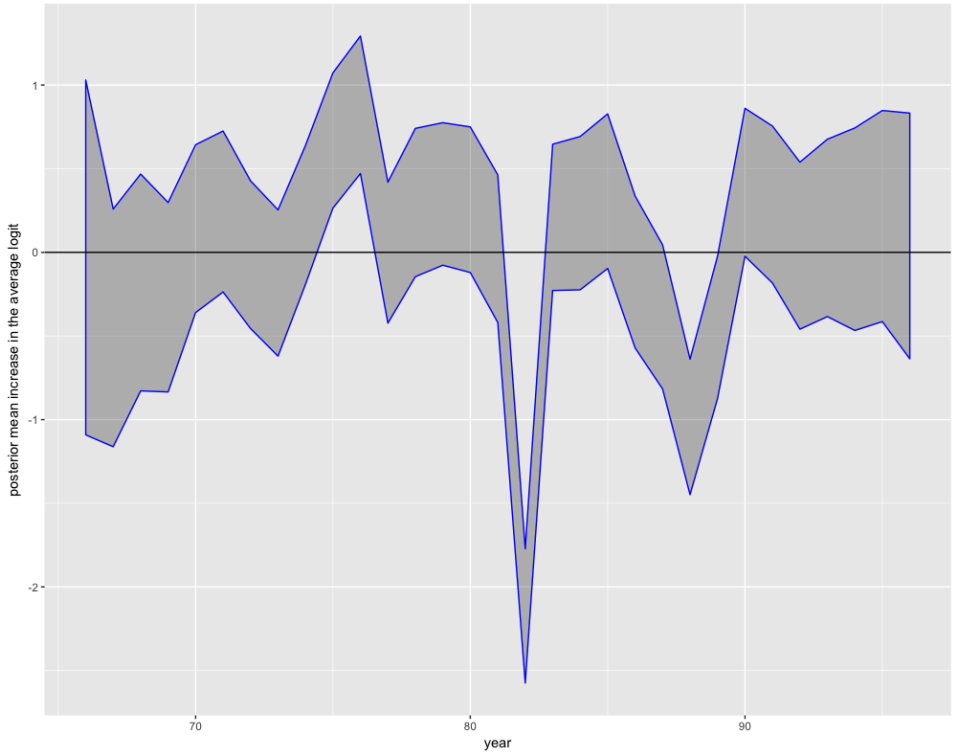

Figure 3.6: A plot of the year–by–year change in the logit of selection captured by the autoregressive $\beta_1^\star(t)$ process in the R process in Implementation 2. Notice the sharp negative signal seen in 1982. Note that the plot for Implementation 3 is almost identical.

Interestingly however, despite this reasonably strong evidence of PS, the posterior predictions of black smoke levels are almost identical to those from Implementation 1. Fig A.5 and Fig A.2 both appear strikingly similar to those from Implementation 1 (Fig 3.5 and Fig A.5). In particular, no obvious changes in the estimated BS levels are seen across the unsampled regions of the Scottish Highlands or the foot of Cornwall. Furthermore, the posterior mean black smoke level averaged across GB remains largely the same throughout time relative to the predictions from Implementation 1.

Thus it appears that despite the joint model detecting PS under Population 1, little–to–no change in the posterior estimates is seen in either the

GB–average levels or the individual site–specific BS trajectories. This is in stark contrast with the observed de–biasing of the regional mean witnessed shortly under Implementation 3. The explanation for these two results may be best explained in terms of the two different populations $\mathcal{P}_1, \mathcal{P}_2$ of sites under consideration for selection.

For $\mathcal{P}_1$, since the sites considered for selection at each time t are only the locations in which an operational site is placed at any time $t \in T$, no information about the selection of sites has been added to the never–sampled regions in Ω. Consequently, when estimating the levels of black smoke via the estimation of the latent Gaussian fields in these regions, we have no additional information about the possible values they could take. Thus, model–based estimates in these unsampled regions will tend towards the predicted global mean levels, which in this case is precisely the $\mathcal{P}_1$–mean (the average taken across the network of observed locations). Furthermore, given the high average lifetime of the monitoring sites, estimates of the site–specific trajectories and hence the $\mathcal{P}_1$–mean barely change under the joint model due to the over–determined nature of the estimation. This is in stark contrast with $\mathcal{P}_2$ or when a point process approach is taken. These place zero counts throughout the domain Ω and hence add additional information into the never–sampled and hence under–determined regions. The lack of change in estimates of the $\mathcal{P}_1$–mean is not a problem with the model. The quadratic model showed a good model fit and we therefore see the inability of the model to change the longitudinal trajectories at the observed site locations for this dataset as proof of the model's robustness – we would almost certainly be concerned if the estimates changed dramatically at the site locations.

If instead, when forming our predictions of black smoke at these never–sampled locations, the model had the additional information that no site was selected here at this time (i.e. $R_{i,j} = 0$ at site $\mathbf{s}_i \in \Omega \setminus \mathcal{P}_1$), then this would provide the model with additional information about the likely values of black smoke at this location. For example, if PS were detected by the model, such that locations in regions with above average black smoke were estimated to have a site with higher probability (i.e if $d_\beta > 0$), then knowl-

edge that a site was not placed at a given location would provide (albeit only slight) evidence for the model that the black smoke level here is below the operational network average. Suppose instead that we have a whole region such as the Highlands, with no monitoring sites present at any time. Estimates of black smoke across this region could then be considerably below the average of the predicted levels at the observed site locations throughout time, depending upon the magnitude of PS detected. This idea of filling the region with zeros to indicate non–selection is the basis of the paper of Diggle et al. [2010] and the approach taken in Implementation 3.

For datasets where the average lifetimes of the monitoring sites are shorter, the measurement error is higher, and/or the functional form of the temporal trend is of higher order, then this joint model framework would have a greater capacity to change estimates of site–specific trajectories, the $\mathcal{P}_1$–mean and hence predictions throughout Ω. This was seen in our simulation study. However, for many applications involving data collected from static monitors, little will change in inferences under a joint model with population $\mathcal{P}_1$. An example of where large differences may be witnessed is for data collected over time from mobile monitors whose locations change at each time step. In this setting we would have a very sparse data setup, with only a single observation of the process' trajectory obtained at each location. The large under–determined missing–data problem here would present the perfect opportunity to assess the ability of the joint model framework to adjust the inference.

After the extensive network redesign in 1982, the autoregressive $\beta_1^\star(t)$ process captured a sharp decline in the average logit for site selection in 1982 (see Fig 3.6). This process may be reflecting, among other things, the year–by–year changes in public and political moods towards pollution monitoring. The 95% posterior credible intervals do not cover 0 and thus the drop of over half of the network in 1982 appears to be a significant event in the lifetime of the network.

Turning our attention now to the estimated parameters of the site selection process $R_{i,j}$, no clear repulsion effect α_{rep} was detected ($\alpha_{rep} = 0.08$ 95% CI (-0.14, 0.31)). This implies that any clustering or repulsion effects

witnessed in the data with respect to $\mathcal{P}_1$ can be attributed to the levels of black smoke alone. On the contrary, the retention effect was found to be very large 6.18 (95% CI (6.07, 6.29)), in agreement with common sense. This finding indicates that there is a clear incentive (possibly financial) for site–selectors to maintain sites in their current locations instead of relocating them each year.

In summary, for this dataset Implementation 2 does not lead to changes in site–specific trajectories, nor does it lead to changes in estimated BS levels in unsampled regions of GB. However, we do still gain some useful insights. We find that the site–selection was in fact preferentially made (i.e. response–biased), and that the extent of this PS could not be attributed to chance alone. Furthermore, we were able to investigate the impact of other factors, such as retention effects and changing political affinities for the network expansion on the evolving operational network S_t. We have presented future applications where the results from implementations 1 and 2 may not agree so closely.

3.5.3 Implementation 3 – $\mathcal{P}_2$

Firstly, we consider the posterior parameter estimates for the two sources of PS (see Table 1). These are denoted by d_β, d_b, the medium-range and short-range preferentiabilities respectively. The posterior estimated value of d_β was 2.77 with a 95% posterior credible interval (2.76, 2.79). The posterior estimate of the short-range preferentiality was 0.12 with a 95% posterior credible interval (0.11, 0.13). Thus in both cases the direction of preferentiality was significantly positive, suggesting that year-by-year, the site placements were positively associated with the relative levels of black smoke at the site location, both locally and regionally.

Fig A.3 shows a striking difference in the appearance of the estimated black smoke field through time. A direct consequence of the strong PS detected is the dramatic drop in the posterior predictions of black smoke levels in undersampled regions of GB relative to Implementation 1. Fig A.3 shows a huge drop in estimated levels in the unsampled regions of Northern

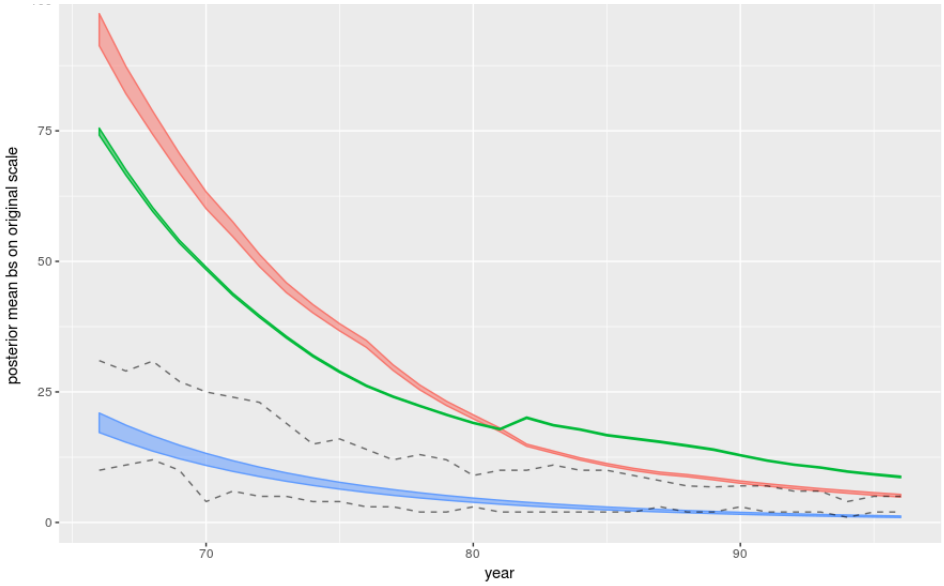

Figure 3.7: Implementation 3. In green are the model–estimated BS levels averaged over sites that were selected in $\mathcal{P}_1$ (i.e. operational) at time t. In contrast, those in red are the model–estimated BS levels averaged over sites that were not selected in $\mathcal{P}_1$ (i.e. offline) at time t. Finally, in blue are the model–estimated BS levels averaged across Great Britain. Also included with the posterior mean values are their 95% posterior credible intervals. The black dashed lines denote the lower 10th percentile and lower quartile observed in the data. Note that the estimated black smoke trajectories from the pseudo–sites are not included in the mean calculations to form the red band. If printed in black-and-white, the green band is initially the middle line, the red band is initially the upper line and the blue band is initially the bottom line. Notice that the GB–average BS levels are around a quarter of their sizes seen in Implementation 2.

Scotland, Mid Wales and the foot of Cornwall relative to Fig A.1 and Fig A.2. Implementations 1 and 2 estimated these regions to have average BS levels due to the lack of any additional information in these regions. Furthermore, Fig 3.7 shows that the posterior mean black smoke level averaged across GB is around a quarter of the size of that estimated from implementations 1 and 2 (see Fig 3.5 and Fig A.5). This is a direct consequence of the decreased levels estimated in the undersampled regions that make up a large percentage of the surface area of GB. This addresses objective 3 of the analysis.

Interestingly, model inferred black smoke levels in these unsampled regions have very high standard errors (i.e. large pointwise posterior standard deviations) associated with their point estimates. This can be seen in the bottom two plots of Fig A.3. Here, the upper 95% pointwise credible intervals actually cover the estimates from Implementation 1. As expected, the posterior estimates of the observed site trajectories (both operational and offline) change very little (see Fig 3.7).

To address Objective 4 refer to Fig 3.7. In agreement with Figures 3.5 and A.5, it appears that the magnitude of preferentiality increases over time. Initially, the annual averages at the locations of the offline observed sites far exceed those from the locations of the operational observed sites. The difference diminishes over time until the major network redesign in 1982, which led to a change in direction of the relative annual mean levels. Thus it appears that the magnitude of the bias in the reported annual black smoke levels from the operational network, relative to the Great British average increased over time - with a dramatic step-change seen in 1982. Of most importance however is the discovery that the observed black smoke levels from the network appears to have never been representative of the levels of GB as a whole, with a positive PS effect detected at all times. In fact, Fig A.3 shows that around 85-90% of the sites in the network were placed in regions with above $\mathcal{P}_2$–mean BS throughout the lifetime of the network.

Once again the autoregressive $\beta_1^\star(t)$ process reflecting the year-by-year changes in public and political mood towards pollution monitoring, captured a sharp decline in the average log intensity for site placement in 1982. The estimate is almost identical to that seen in Implementation 2 (see Fig 3.6)

and so we omit the plot.

Regarding the estimated parameters of the site selection process $R_{i,j}$, the α_{rep} term was detected to be positive with value 0.82 [95% CI (0.62, 1.02)]. This implies that there is additional clustering present that cannot be explained due to the levels of black smoke alone. This may be capturing some of the latent factors influencing the selection of monitoring sites such as population density.

3.5.4 Impacts of preferential sampling on estimates of population exposure levels and noncompliance

Whilst the dramatic decline in GB–average black smoke levels seen under the joint model in Implementation 3 is interesting, the monitoring network was not intended for the accurate mapping of black smoke across the whole of Great Britain but instead was established for tracking the progress achieved by the Clean Air Act in reducing the exposure levels of both black smoke and sulphur dioxide [McMillan and Murphy, 2017]. Thus, judging the monitoring network based on its ability to represent the levels of black smoke across GB as a whole is potentially misleading. Taking this into consideration, we now attempt to assess the effects of PS on estimates of population exposure, and hence the effects of PS on the ability of the network to fulfill its objectives. Over the time period of study, various EU limits and guidelines on annual black smoke levels were introduced, including the annual average guide value of $34\mu\text{gm}^{-3}$ introduced in 1980 (repealed in 2005) [Zidek et al., 2014]. We repeat the analysis of Zidek et al. [2014] and assess the changes in the estimates of noncompliance under PS.

For estimating the population exposure levels, we obtained gridded residential human population count data with a spatial resolution of 1 km x 1 km for Great Britain based on 2011 Census data and 2015 Land Cover Map data from the Natural Environment Research Council Centre for Ecology & Hydrology [Reis, 2017]. The data came in the form of a raster layer and we formulate our estimate of population density across the time period (1966 - 1996) by normalizing the count raster by dividing each cell by the total sum across all the cells. Here we assume that the relative population density

has remained stable from 1966-2011 for the estimated population density layer to be a good proxy across the years of study. We also assume that residential population density is a good proxy of where the population is situated throughout the year and hence that actual black smoke exposure levels are similar to estimated residential levels. Next, we define a projector matrix, to project the GMRF estimated in INLA on the triangulation mesh onto the centroids of the population density cells that make up the raster.

Finally, we are able to use the Monte Carlo samples from the posterior marginals from INLA and the projector matrix to estimate the posterior distribution of the black smoke field at each of the grid cells. Letting $\rho_j(\mathbf{s})$ denote the population density of Great Britian at location $\mathbf{s} \in \Omega$, in year j, such that $\int_\Omega \rho_j(\mathbf{s})\mathrm{d}\mathbf{s} = 1$, we can then estimate the population–mean exposure levels by approximating the following integral:

$$\begin{aligned}\mu_{pop,j}(\Omega) &= \int_\Omega \mu(\mathbf{s}, j)\rho_j(\mathbf{s})\mathrm{d}\mathbf{s} \\ &\approx \sum_{i=1}^{G} \bar{\hat{\mu}}_j(\mathbf{s}_i)\hat{\rho}_i = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{G} \hat{\mu}_{i,j,m}(\mathbf{s}_i)\hat{\rho}_i\end{aligned}$$

where $\mathbf{s}_i$ denotes the i^{th} raster grid cell centroid ($i = 1, ..., G$), $\bar{\hat{\mu}}_j(\mathbf{s}_i)$ denotes the Monte Carlo mean black smoke level at location $\mathbf{s}_i$ in year j and $\hat{\rho}_i$ denotes the estimated population density at the i^{th} grid cell. Approximate credible intervals for this quantity can also be formed. We can also use this method to estimate the proportion of the population exposed to annual average black smoke levels exceeding the EU guide level of $34\mu\mathrm{gm}^{-3}$ each year, by simply replacing the term $\hat{\mu}_{i,j,m}(\mathbf{s}_i)$ in the summation by the indicator variable representing the event that the value exceeds $34\mu\mathrm{gm}^{-3}$. Note here that the index m denotes the Monte Carlo sample number.

We now do this, both for the estimated black smoke levels under Implementation 2 (i.e. Population 1) and again under Implementation 3 (i.e. Population 2). Note that the results under Implementation 1 are almost identical to those from Implementation 2 so we omit them in the plots.

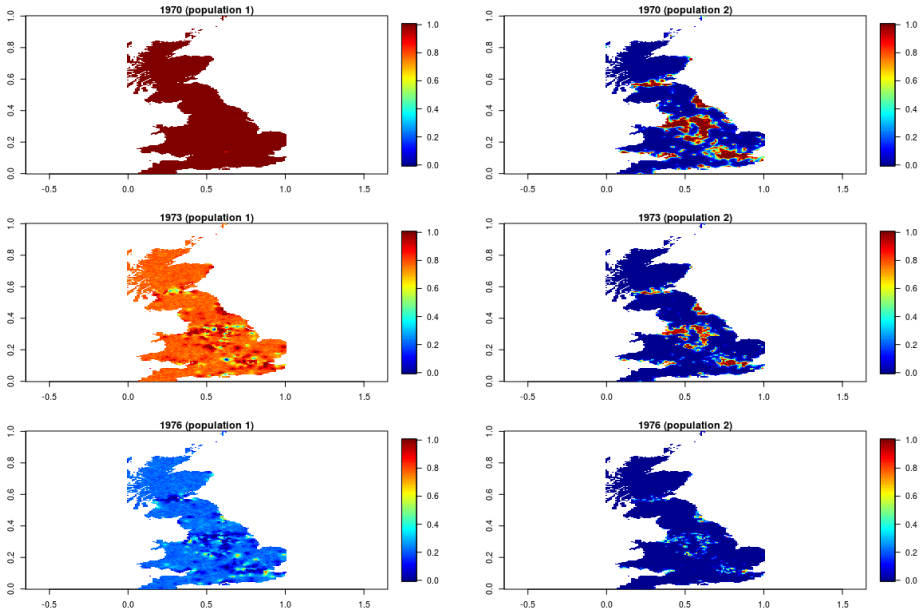

Figure 3.8: A map plot of the posterior pointwise probability of the annual average black smoke level exceeding the EU guide value of $34\mu\text{gm}^{-3}$ under Implementation 2 (left) and Implementation 3 (on the right). From top to bottom are the years 1970, 1973 and 1976. The colour scale goes from 0 to 1 for all the plots, with dark blue denoting a posterior probability of 0 and dark red denoting a posterior probability of 1. Note that the plots for Implementation 1 are almost identical to those from Implementation 2 and are omitted. Notice that almost none (the majority) of Great Britain is confidently predicted in 1970 to have BS levels below the EU guide value under Implementation 2 (3).

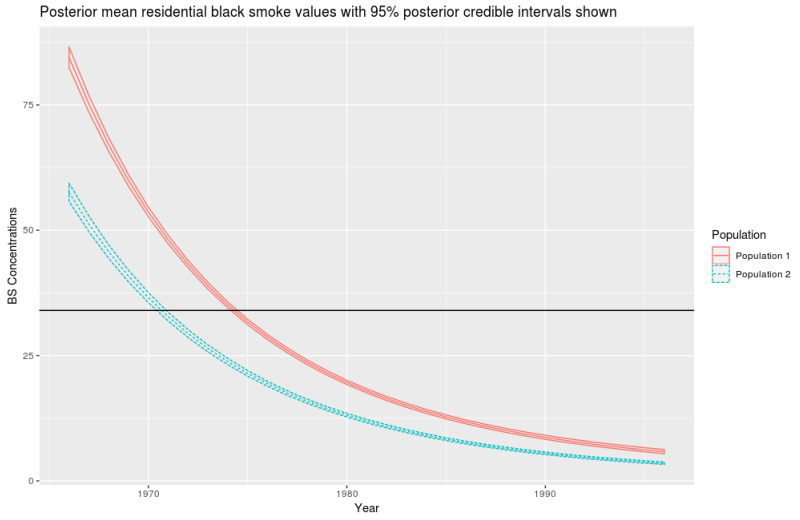

Figure 3.9: A plot showing the posterior mean and 95% credible intervals of the annual residential–average exposure levels across the years of study. Shown are the results from Implementation 2 (i.e. Population 1) and from Implementation 3 (i.e. Population 2). The horizontal line denotes the EU guide value of $34\mu\text{gm}^{-3}$.

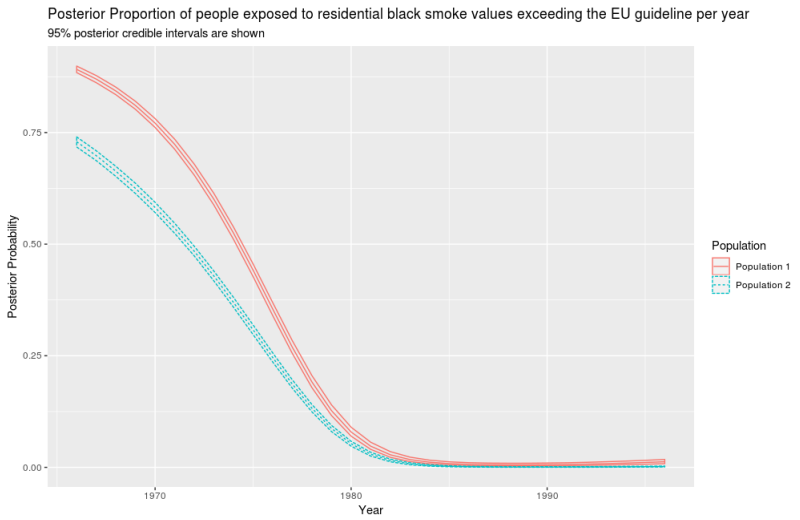

Figure 3.10: A plot showing the posterior mean and 95% credible intervals of the annual proportion of the population with black smoke exposure levels exceeding the EU guide value of $34\mu\text{gm}^{-3}$ across the years of study. Shown are the results from Implementation 2 and from Implementation 3.

Fig 3.8 shows plots of the posterior pointwise probability of exceeding the EU annual black smoke guide value of $34\mu\text{gm}^{-3}$ under implementations 2 and 3, across the years 1970, 1973 and 1976. The colour scale goes from 0 to 1 for all the plots, with dark blue denoting a posterior probability of 0 and dark red denoting a posterior probability of 1. In agreement with the plots of the pointwise posterior means (see Fig A.2 and Fig A.3), a dramatic decline in the estimates of noncompliance can be seen under Implementation 3 in the regions far from the nearest monitoring network across the years (see Fig 3.8). This has major ramifications regarding the total reported proportion Great Britain in noncompliance with the guide value. For example, under Implementation 2 almost the entirety of Great Britain is estimated to be in noncompliance with the guide value up until 1970. This figure drops to below 25% in 1970 under Implementation 3 (see Fig A.4).

However, once again the monitoring network and the guide value were intended to measure and control the population exposure to black smoke levels. Thus our maps showing the pointwise posterior probability of exceedance, whilst being dramatic, may not be a fair assessment of the network. Instead, we now focus our estimates on the estimated proportion of the population of Great Britian exposed to black smoke levels out of compliance with the air quality standard. Given that the density of monitoring sites in the network follows the large population centres of GB closely, we expect the differences between the estimates to be much lower. In fact, this is not the case. Fig 3.10 still shows a large decrease in the estimated proportion under Implementation 2, from 89% to 73% in 1966 for example. Note that the posterior credible intervals still show a large discrepancy between the estimated proportions. This is despite us including the additional short scale variability from the spatially-uncorrelated IID effects in the estimates (one pair of realized b terms per 1km grid cell, per Monte Carlo sample).

Finally, we turn our attention to the estimated population–average annual black smoke exposure levels across the two implementations (2 and 3). In agreement with Fig 3.10, Fig 3.9 shows a clear decrease in the estimated

annual averages. Given the sensitivity of health effect estimates of air pollution to the accuracy of population exposure levels, this result is especially striking.

3.6 Discussion

Importantly, a lot of the detected preferentiality effects and subsequent de–biasing effects on prediction are likely mediated by well–known covariates. For example, annual population density figures and/or industrialisation indices (in their correct functional form) would likely simultaneously explain a lot of the PS detected if included in the $R_{i,j}$ process, and be strongly positively associated with the observed levels of Y in the observation process. Sites may well be placed in regions where lots of people live and work to ensure the network captures 'typical' exposures experienced by the public, and some sites may be located in areas close to polluting industry for exceedance detection. Since the daily activities of people and industry may well be the main contributors to black smoke levels, including these covariates in the observation model Y would therefore likely lead to decreased model–estimated pollution levels in unsampled regions such as The Highlands of Scotland with low population density and industry.

In many applications, the PS may disappear upon the inclusion of such covariates and hence be reduced to a missing–at–random scenario. Given that the focus of this Chapter was to repeat previous analyses of this dataset [Shaddick and Zidek, 2014, Zidek et al., 2014] under our new framework and assess the changes, we do not consider including covariates here. Furthermore, we wanted to show that in settings where such covariates are unavailable, sensible adjustments can still be realized under a careful use of our model framework. Additionally, given that the locations of the monitoring sites are almost exclusively situated near population-dense, industrious, and urban regions, it is unclear if these locations would provide the adequate contrast required to estimate the correct the functional forms of these covariates. It would be interesting in future work to see if any PS is detected in this data after conditioning on as many such variables in both processes.

In summary, this Chapter is not attempting to bypass the need for including relevant covariates in the modelling. Rather, it is presenting a method for accounting for the effects of any residual unmeasured confounders associated with both processes by using spatio–temporal fields to act as a proxy.

This modelling framework should be considered to both detect preferential dropout within a fixed population or network $\mathcal{P}$, and to detect if the population or network $\mathcal{P}$ was preferentially placed within the domain of study Ω. Accomplishment of both of the above depends upon the choice of population of sites under consideration for the site–selection process. If PS is detected using this model, then first and foremost, the modeller should attempt to find available covariates that mediate the detected preferentiality. If, after exhausting the available mediators (e.g. population density), and after removing as many sources of variability from the site–selection process as possible, preferentiality is still detected, then this modelling framework should be used for detecting the potential consequences of this sampling scheme on the subsequent inference – either on parameters or spatio-temporal prediction.

Furthermore, different regression models can be explored for the initial site–placement and site–retention processes. For example, different covariates may be believed to affect only one of the two processes, the qualitative behaviour of certain covariates on the two processes may be different or perhaps the nature of PS could differ across the two processes. We did not explore these possibilities here, assuming only a unique intercept existed between the two processes.

Additionally, the functional form used to model PS can be as flexible as desired. Here we opted to model the direction and magnitude of PS as being constant through time. In reality this may not be suitable and the direction and magnitude of preferentiality may change through time. In Fig 3.9 we can see that initially (at $t = 1$) the operational network was established such that it gave annual readings below the $\mathcal{P}$–mean under Population 1. Then, as time progressed, the magnitude of the preferentiality decreased as the annual averages from the operational sites approached those from the population average. Thus it may make sense here to estimate a separate

preferentiality parameter d_β for times 1 and for $t > 1$. For time 1 this would likely be estimated to be smaller compared with for $t > 1$. For simplicity we opted against this approach, however such a model would help paint a more detailed picture of the dynamic nature of the PS through time.

If one wishes to adjust the estimates of the domain–average (the GB–average in our example) to the effects of PS, the population of locations $\mathcal{P}$ considered for selection should be extended to include locations in unsampled regions in the domain of study Ω. Population 2 did just that, and as a result the GB–average estimates significantly dropped under the joint model. An alternative approach would be to consider modelling the site placement events each year implicitly as realisations from a LGCP and the site retention events separately as Bernoulli trials. Two reasons for not pursuing this approach were given earlier in the Chapter.

Extensive analytic and simulation studies on jointly modelling dropout with various longitudinal clinical markers have been made in biostatistics over the past 20 years. The inspiration for this work came from the literature on the joint modelling of viral load, dropout and longitudinal clinical markers measured in HIV clinical trials [Lawrence Gould et al., 2015, Li and Su, 2018, Wu, 2009]. In fact, after transforming the data, Fig 3.2 shows black smoke trajectories that are very similar to the subject–specific dose–response trajectories seen in such longitudinal clinical data. The same philosophy behind jointly modelling informative patient dropout with the process of interest via shared random effects can be applied to spatio–temporal environmental network data with minimal alteration. The major difference with spatio-temporal data are the spatial correlations assumed on the random effects. It is this correlation which allows for the spatial extrapolation to occur.

Whilst the case study in this Chapter considered the observations to be on the same time scale as the site–selections, this need not be the case. For example, this general framework could simultaneously model high–frequency (e.g. hourly) observations with a low–frequency (e.g. annual) site–selection process. This would comprise decomposing the temporal trajectories into trend, seasonal and cyclical (e.g. daily) terms in the model. It would then

likely make most sense to include only the trend term in the linear predictor of the site–selection process.

Assuming the locations of the monitoring sites are realisations from an IPP or a LGCP, while being useful computationally, may not always be sensible in certain applications. For example, if a strict lower limit on the distances between the monitoring site locations was known, then a LGCP or a IPP would not be the most suitable model for use and alternatives such as a Matérn hard-core point process model would be more suited [Baddeley et al., 2015]. Having said that, a nice property of using our logistic regression approximation to the LGCP, is that we are able to delete pseudo–site locations in $\mathcal{P}_2$ that violate any known rule (e.g. a minimum distance/hard–core rule). Furthermore, if additional clustering is present then a cluster point process or Gibbs point process may be more desirable [Baddeley et al., 2015]. Whilst we attempt to adjust for the additional clustering seen in our dataset by constructing a covariate $I_{i,j}$, this is by no means the best way forward here.

On a closing note, it should be apparent that the modelling framework introduced in this Chapter can be applied to monitoring data that have come from static monitoring sites, mobile monitoring sites, and a combination of the two. Furthermore, the ability for the joint model framework to adjust for PS under $\mathcal{P}_1$ should be greater in applications with mobile monitors. One such study that could be revisited is the MESA Air Study (http://www.mesa-nhlbi.org/). Since this study involves the estimation of the health effects associated with exposure to various air pollutants, with pollution readings taken from a combination of static and mobile monitoring sites, this data set offers an ideal opportunity to test out this framework. Of interest may be the detection of any PS, and its resulting effects on the health effects.

3.7 Conclusion to Chapter 3

We applied our general framework to the network of air quality monitors in Great Britain between the years 1966-1996. From this, we were able

to show that the monitors were preferentially placed within Great Britain throughout the life of the network. In particular, each year the locations of the operational sites were found to have been situated in areas with black smoke levels considered much higher than the annual average level across Great Britain. Furthermore, we showed that the network was updated in a preferential manner throughout the life of the network. Monitoring sites at locations with highest black smoke levels were favoured for selection into the network each year, and monitoring sites at locations with lowest black smoke levels were favoured for removal from the network each year.

The implications for this biased network placement were then clearly demonstrated. The PS of the monitoring sites may have had a significant deleterious impact upon the ability of the network to serve its purpose as a tool for measuring the black smoke exposure levels experienced by the population of Great Britain as a whole. It appears that estimates of population exposure levels may have been overestimated (see Fig 3.9). Furthermore, estimates of noncompliance to the various air quality regulations established throughout the chosen time period of 1966 - 1996, may also have been affected by how and where the monitoring sites were situated. It appears that any estimates of noncompliance that used the observations from the air quality monitoring network may have over–estimated the true amount of noncompliance (see Figures 3.8 and 3.10). This includes historical estimates of the proportion of the population of Great Britain exposed to black smoke levels that were out of compliance.

Chapter 4

A Perceptron for Detecting the Preferential Sampling of Sites Chosen to Monitor a Spatio-temporal Process

"the priority area is the zone of highest pollution concentration within the region; one or more stations should be located in this area."
— The United States' EPA Monitoring Network Design QA Handbook Vol II Section 6.0, guidelines for selecting the number and locations of air pollution samplers

A preview

The previous chapter developed a general framework that enables PS to be detected in spatio-temporal analyses and subsequently adjusted for when making inference and predictions. The framework also allows for the selection process itself to be emulated. However, the previous Chapter relied on the precise nature of PS to be known a-priori. In particular, the method requires the selection process to be specified, with PS described in its cor-

rect functional form. In practice, the precise nature of the PS may not be known.

Furthermore, fitting the joint model framework introduced in the previous Chapter can be both computationally costly and challenging to implement. These challenges may render the method unsuitable for applied researchers. Consequently, researchers may remain unable to test for the presence of PS in their spatio-temporal datasets. This Chapter focuses on developing an easy-to-implement test for PS in spatio-temporal data.

Importantly, this Chapter relaxes the assumption that the precise nature of PS is known. It is instead replaced with a simpler assumption, namely, that the PS is monotonic. We define this to mean that the density of space-time points chosen to observe the spatio-temporal process $(\mathbf{S}, \mathbf{T}) \subset (\Omega, \times\mathcal{T})$, depends monotonically on the values of the latent spatio-temporally correlated random effects $\mathbf{Z}$ used to describe the spatio-temporal process. This simpler assumption allows for a computationally-fast and general test for PS to be developed.

We refer to the test as a perceptron as it attempts to capture the numerous factors behind the human decision-making that selected the sampled locations. Importantly, the method can also help with the discovery of a set of informative covariates that can sufficiently control for the PS. The discovery of these covariates can justify the continued use of standard methodologies. The test is applicable to both discrete-space and continuous-space settings and is developed within the STGLMMs framework introduced in Chapter 2. We demonstrate its high power across a range of settings in a thorough simulation study. We then apply it to two real-world datasets.

Throughout this Chapter, we will be using the nearest neighbour (NN) distances between points as a way to measure the local degree of clustering present. In particular, associated with each point or 'site' chosen to observe the process will be the nearest neighbour distance (or averaged K-nearest neighbour distances for some chosen integer K) to the closest point(s) or site(s). This distance value is chosen to capture the local degree of clustering that is present around each point or site. The toy examples presented earlier in sections 1 and 2 of Chapter 2, demonstrated a clear characteristic

of PS in both the discrete-space and continuous-space settings. When PS is present under a log-Gaussian Cox process, it is apparent that an increased density of space-time sampling locations (or units) is chosen to observe the spatio-temporal process in regions where the spatio-temporal process is high (or low). This suggests that when PS is present under this assumed sampling process, a correlation should exist between the NN distances and the observed response values at each of sampled locations and times.

To make this point clear, we revisit the toy example of Section 2.2. For both the PS and MCAR datasets, we compute the NN distances for each of the sampling locations and then compute the Spearman's rank correlation between the field values and the NN values. This is seen in the first two rows of Figure 4.1 and a clear negative trend between the NN distances and the field values can be seen when the data are PS. The lowest ranks of the NN distances are seen more frequently in the West of the map, precisely where the field takes the largest values, leading to a moderate negative rank correlation of -0.39 existing between the field and NN values. Conversely, when the data are not PS, no such trend is seen. The rank correlation seen in the toy example data is close to zero (-0.08). Next, we repeat the process for the PS data, but compute the average K NN distances from each point for $K = 2$ and $K = 4$. The results are seen in the bottom two rows of Figure 4.1. Stronger correlations of -0.61 and -0.79 are seen for $K = 2$ and $K = 4$ respectively, demonstrating the possible variance reduction in the ranked NN distances, and hence the possible increased power to detect PS, offered by increasing K in certain settings. The results from this example provide the intuition behind the PS test discussed next.

When the sampling process is assumed to be a different point processes (e.g. a Hardcore process), the NN distance may be a poor choice to quantify the degree of clustering. We consider the benefits of using alternative measures of clustering in this Chapter.

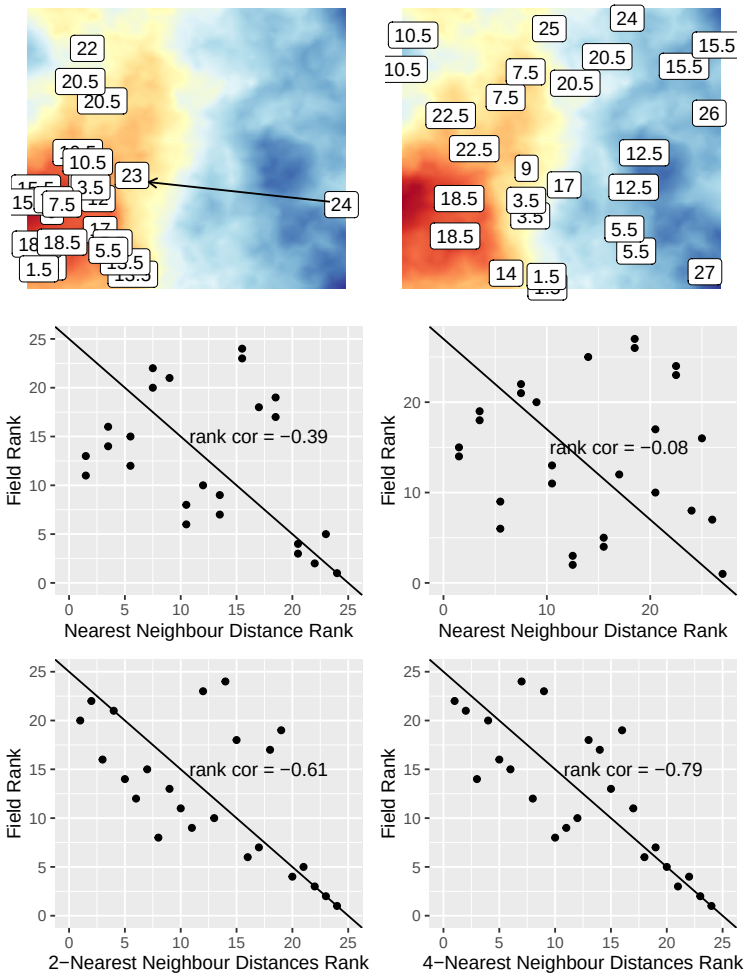

Figure 4.1: The top left plot shows the 24 PS sampling locations. The Gaussian process is shown in colour, with blue (red) representing the lowest (highest) values. Sampling locations are labeled with their ranked NN distance. Largest ranks are seen in the regions with the lowest field values. The arrow shows the NN distance from the site labelled '24'. The top right plot repeats the above for the 27 'MCAR' sampling locations. No trend can be seen. The center-left and center-right plots show the field ranks plotted against the NN distance ranks for the PS and MCAR datasets respectively. The black line demonstrates a rank correlation of -1. Rank correlations of -0.39 and -0.08 are found in the PS and MCAR data respectively. The bottom two plots show that these correlations increase for the PS data to -0.61 and -0.79 when the ranks of the average of the 2 and 4 NN distances are computed for each sampling location.

4.1 Introduction to Chapter 4

This Chapter concerns preferential sampling (PS), where the locations selected to monitor a spatio–temporal process $\mu_{st},\ s \in \Omega,\ t \in \mathcal{T}$, depend stochastically on the process they are measuring. PS is a special case of response–biased sampling. The space–time point is defined as $(\mathbf{s}, t) \in \Omega \times \mathcal{T}$. Ω denotes the spatial domain of interest and $\mathcal{T}$ denotes the temporal domain. Purely spatial processes (i.e. when $\mathcal{T}$ is a singleton) are a special case.

To gain understanding of the process $\mu_{s,t}$, a set of time points $T \subset \mathcal{T}$ at which to observe $\mu_{s,t}$ are selected. Then, for each $t \in T$, a set of n_t sampling locations $S_t \subset \Omega$ are chosen. Generally, the temporal domain $\mathcal{T}$ is a finite set, with μ a time–averaged quantity for practical reasons. Typically, $\mu_{s,t}$ is not observed directly and instead a noisy observation $Y_{s,t}$ is taken instead. The noise could be due to the presence of measurement error (i.e. the nugget effect) or other factors. S_t may represent a set of points in space (i.e. $S_t = (\mathbf{s}_i \in \Omega)_{i=1}^{n_t}$), or a set of well-defined areal units (i.e. $S_t = (\mathbf{A}_i \subset \Omega)_{i=1}^{n_t}$). In this Chapter, these two cases are referred to as the geostatistical and discrete spatial settings respectively. In the latter case, observations will generally represent spatial–averages of $\mu_{s,t}$.

Difficulties arise with the estimation of $\mu_{s,t}$ when S_t are preferentially selected. This is because most statistical methods for modelling spatio-temporal data, condition on the locations as fixed [Cressie and Wikle, 2015, Diggle et al., 2010]. Such models assume that locations were selected under complete spatial randomness. Departures from this assumption in the true sampling scheme can lead to large biases in the prediction of the process $\mu_{s,t}$ (see Chapters 2 and 3). Hereafter, models that consider the locations as fixed are referred to as 'naive'.

Additionally, a set of covariates $\mathbf{X}_{s,t}$ may exist that influence the choice of sampling locations S_t. These covariates may also be associated with the underlying process being modeled $\mu_{s,t}$. When this occurs, including the necessary $\mathbf{X}_{s,t}$ in a 'naive' regression model for $\mu_{s,t}$ may partially remove the deleterious effects of PS on the spatio-temporal prediction of $\mu_{s,t}$ [Gelfand

et al., 2012]. This becomes a regression-adjustment approach to help correct for the departure from a complete spatial randomness sampling design for S_t. Covariates common to both the sampling process and $\mu_{s,t}$, are hereafter referred to as 'informative' as in Gelfand et al. [2012].

Preferential sampling has been identified as a major concern across multiple fields. In ecology, PS may occur due to sightings data being comprised of opportunistic sightings or poorly-designed surveys (see Chapter 5). Observers frequently focus their efforts in areas where they expect to find the species, leading to PS [Fithian et al., 2015]. A consequence of this is that 'naive' estimates of the geographical distribution of a species may be severely biased. Estimates of a species' abundance have also been shown to be affected [Pennino et al., 2019]. PS should also be considered in the analysis of environmental data recorded from tagged animals. Dinsdale et al. [2019] demonstrated this with a case study using sea surface temperature recordings from tags attached to Elephant Seals in the Southern Indian ocean. The seals' preference for cooler waters led to biased 'naive' spatial estimates of sea surface temperature.

In environmental statistics, the deleterious impacts of PS have been highlighted. For example, pollution concentration levels throughout Ω and $\mathcal{T}$ are commonly estimated using noisy observations, $Y_{s,t}$, recorded from environmental monitoring networks [Shaddick et al., 2018]. Here, the locations of the monitors in a network S_t, may have been chosen in a preferential way to meet specified objectives [Schumacher and Zidek, 1993]. For example, urban air pollution monitoring sites are sometimes used for detecting noncompliance with air quality standards [EPA, 2005, Loperfido and Guttorp, 2008]. In such settings, observations $Y_{s,t}$ will likely lead to overestimates of the overall levels of the air pollutant, $\mu_{s,t}$, throughout Ω and $\mathcal{T}$. These biased 'naive' estimates $\hat{\mu}_{s,t}$ may then be unsuitable for assessing the impacts of $\mu_{s,t}$ on human health and welfare [Lee et al., 2015].

Previous PS tests have been developed for continuous spatial data, but limitations hinder their general use. Firstly, Schlather et al. [2004] developed two Monte Carlo tests. Their null hypothesis assumes that the data are a realization of a *random-field model*. They assume: the sampled point

locations S_t are a realization of a point process $\mathcal{P}$ on Ω, the recorded values (called marks) of the points are the values of a realisation of a random field $\mu_{s,t}$ on Ω, $\mathcal{P}$ and $\mu_{s,t}$ are independent processes. Independence here implies a non-preferential sampling mechanism. To detect departures from the null hypothesis, the authors define two characteristics of marked point processes, denoted $E(d)$ and $V(d)$. These represent respectively the conditional expectation and conditional variance of a mark, given that there exists another point of the process at a distance d. These are chosen since under the null hypothesis E and V should be constant. Monte Carlo tests are used to assess departures of estimates of E and V from a constant function. This approach requires the assumption of Gaussian observations and hence does not generalize to non-continuous marks.

Next, Guan and Afshartous [2007] developed an alternative simulation-free test for PS. Instead of fitting a parametric model for the marks, their approach instead divides the region Ω into non-over-lapping subregions. These are assumed to be approximately independent, generating approximately IID replicates of the test statistic. The spatial range of $\mu_{s,t}$ can be thought of as representing the inter-point distance required for two observations of $\mu_{s,t}$ to be approximately independent. Finding a suitable set of subregions required for their test may prove a challenge when the spatial range of the correlation of $\mu_{s,t}$ is large relative to the size of Ω. Furthermore, this test requires very large sample sizes; their application used a sample size of over 4000.

For modelling PS directly, it is common to take a model-based approach. Approaches often simultaneously fit a model for the observation process, $Y_{s,t}$, with a model for the sampling process, $\mathcal{P}$, within a joint-model framework [Diggle et al., 2010]. Linear combinations of any spatio-temporal latent effects used to describe $\mu_{s,t}$ are shared across the linear predictors of the two processes. This sharing of latent effects helps to capture any stochastic dependence that may exist between the two processes. A nonzero effect estimate of any of these linear combinations provides evidence that PS is present (see Chapter 3). Whilst this approach has been successfully applied to mitigate PS, the use of this approach to test for PS may be out of reach

of many researchers. These joint models are currently not implemented in many popular software packages, are computationally intensive to fit and can be difficult to design and interpret. Note that design-based approaches have also been introduced for specific scenarios [Zidek et al., 2014]. Hereafter, the collection of spatio-temporal latent effects are denoted $\mathbf{Z}_{s,t}$.

Due to the computational challenges of fitting joint models and the lack of generality of the current PS tests, PS appears to be often overlooked. Researchers may have non-Gaussian data, or have too small a sample size to perform either test. Consequently, without the ability to test for PS, researchers may then fit 'naive' models to preferentially sampled data. The potential consequences of PS on their inferences may then be ignored. Fortunately, in many situations, a sufficient set of informative covariates $\mathbf{X}_{s,t}$ may be available. The decision of where to sample is typically a human one, and hence $\mathbf{X}_{s,t}$ may include measures of accessibility (e.g. distance from the nearest road or population center). These are often associated with the underlying process being measured and thus may help to control for the PS. Verifying the existence of $\mathbf{X}_{s,t}$ would allow researchers to confidently continue to use their preferred methodologies and packages, without the need to fit joint models Fithian et al. [2015].

This Chapter presents a computationally fast method for detecting PS. The algorithm for implementing the test is both intuitive and easy to program. The method primarily requires that the researcher be able to predict the values of $\mu_{s,t}$, and any latent spatio-temporal effect $Z_{s,t}$, throughout Ω and $\mathcal{T}$. Any preferred 'naive' method can be used. The method is general in that it can test for PS in both the geostatistical and discrete spatial settings, and can be used when the responses (marks) are non-Gaussian and even non-continuous. A general algorithm is provided for all settings. The test can also be adjusted for covariates, allowing researchers to discover if a given $\mathbf{X}_{s,t}$ is sufficient for controlling the PS.

Qualitatively, PS has a clear appearance in continuous spatial data. PS often appears as a clustering of locations chosen to observe $\mu_{s,t}$ in regions where one or more $Z_{s,t}$ is either high or low. The test in this Chapter directly targets this excess clustering. In the continuous spatial setting, a suitable

point process is fit to the observed locations to capture the true sampling process under the null hypothesis of no PS. Then, Monte Carlo (MC) realisations of the point process under the null are generated. The magnitude of correlation between the degree of clustering and the estimated values of $Z_{s,t}$ is computed for both the observed data and the MC realisations. If a stronger correlation is observed in the observed data compared with the MC samples, then evidence for PS has been found. The mean of the K nearest neighbours is our default recommendation to capture the degree of clustering as this quantity may also be used in the discrete spatial setting. In the discrete spatial setting, a Bernoulli sampling process is instead fit to a population of well-defined areal units under the null. A clustering of areal units chosen to observe $\mu_{s,t}$ in regions where $Z_{s,t}$ is either high or low indicates PS.

The Chapter is organized as follows. Section 2 introduces the assumed marked point process data-generating mechanism for the geostatistical setting. Then, the perceptron algorithm and properties of the PS test are described. Section 3 repeats the above for the discrete data setting. Section 4 demonstrates the power of the test to detect PS in a thorough simulation study. The joint effects of the: sample size, spatial smoothness of $\mathbf{Z}_{s,t}$, spatio-temporal covariates $\mathbf{X}_{s,t}$ and the magnitude of PS on the power of the test are discussed. Section 5 applies the test to two real datasets previously analysed in the literature. The PS test can be performed using the R package *PStestR*, now available on GitHub, that we developed.

4.2 Preferential sampling in geostatistical data

In continuous spatio-temporal settings, observations $Y_{s,t}$ are taken at a set of point locations S_t within the study region Ω at each time step $t \in T \subset \mathcal{T}$. Standard approaches for modelling $\mu_{s,t}$ from a set of observations $Y_{s,t}$ include variogram analysis and kriging-based methods [Diggle and Ribeiro, 2007]. These methods fall under the umbrella term of "geostatistical methods" and require the assumption that the locations chosen to observe the process $\mu_{s,t}$ were not preferentially sampled [Diggle et al., 2010].

For modelling point-patterns in space and time, spatio-temporal point processes are the standard statistical toolbox [Baddeley et al., 2015, Illian et al., 2008]. This class of models will be used throughout our Chapter to explain the observed point-patterns S_t through time. Standard 'naive' geostatistical methods require the assumption that the sampling process $\mathcal{P}$ generating the sampled locations S_t be independent of the underlying spatio-temporal field $\mu_{s,t}$. This assumption implies no PS and simplifies the analysis greatly. Here, the point-pattern S_t and the marks $Y_{s,t}$ may be investigated separately using standard techniques. However, when this assumption is violated, the two processes must be considered together. Marked spatio-temporal point processes should be considered as a formal framework for such a data analysis [Schlather et al., 2004].

The PS test we are about to describe requires the following three assumptions. The final assumption describes the assumed characteristic behaviour of the PS.

Assumption 4.1 *The PS is driven by one or more spatio-temporal latent effect in* $\boldsymbol{Z}_{s,t}$.

Assumption 4.2 *All of the latent effects within* $\boldsymbol{Z}_{s,t}$ *that drive the PS are spatially 'smooth enough' relative to both the size of the study region* $|\Omega|$ *and the number of locations chosen to sample the process* $|S_t|$.

Assumption 4.3 *The density of points within* S_t *at space-time point* $(\boldsymbol{s}, t) \in \Omega \times T$) *depends monotonically on the values of the components of* $\boldsymbol{Z}_{s,t}$ *driving the PS.*

Assumptions 4.1 - 4.3 imply that preferentially sampled data will appear as point-patterns S_t that are clustered in space for each $t \in T$. These clusters will focus around regions where relevant elements of $\mathbf{Z}_{s,t}$ are especially high or low, depending on the direction of PS. The first inferential goal becomes the detection of monotonic associations between the degree of clustering throughout $\Omega \times T$ with the values of relevant $\mathbf{Z}_{s,t}$. If PS is detected, then the second inferential objective becomes the determination of whether or

not clustering can be explained by a set of informative covariates $\mathbf{X}_{s,t}$. That objective is achieved if such a set removes all of the PS-associations.

The ranked nearest neighbour distances between the sampling locations S_t are proposed as a default choice to measure the magnitude of clustering. Following the recommendations of Gignoux et al. [1999], edge-corrections for these distances are not considered within the Monte Carlo algorithm. Many other quantities can be chosen to capture local clustering and may be more suited for specific S_t generating mechanisms. The PS test developed in this Chapter can easily be modified to use another quantity. The ranked nearest neighbour quantity is chosen for its generalisability across both discrete and continuous spatial settings.

4.2.1 Assumed model for preferential sampling

Many spatio-temporal point processes have been developed, with each possessing fundamentally different properties. An appropriate choice for a given analysis depends upon the sampling protocols that generated S_t. For example, Gibbs point processes allow for second-order effects such as inter-point attraction and repulsion to exist between points. A limiting case is seen in the Hard Core process. This process does not allow for points to exist within a distance R, called the 'range of interaction'. Cluster processes provide a class of point processes that describe the locations of 'parent' points with a separate process from their 'daughter' points [Baddeley et al., 2015]. Many more processes exist and may prove useful in applications.

The simplest class of spatio-temporally varying point processes is the inhomogeneous Poisson process (IPP hereafter) [Illian et al., 2008]. The IPP is completely defined by its intensity function $\lambda(\mathbf{s},t)$. This is defined as the expected number of points per unit area and time immediately around $(\mathbf{s},t) \in \Omega \times \mathcal{T}$. More formally, the intensity can be defined as:

$$\lambda(\mathbf{s},t) = \lim_{\epsilon \to 0} E\left[N(B_\epsilon(\mathbf{s},t))\right] / \left|B_\epsilon(\mathbf{s},t)\right|, \tag{4.1}$$

where the cube $B_\epsilon(\mathbf{s},t) = (\mathbf{s}, \mathbf{s}+\epsilon\mathbf{1}) \times (t, t+\epsilon)$.

Let $\Omega \subset \mathbb{R}^2$. Define two disjoint space-time volumes $(A_1,T_1),(A_2,T_2) \subset$

$(\Omega \times \mathcal{T})$. Then the numbers of points that fall within the two space-time volumes $N(A_i, T_i)$ are independently Poisson distributed random variables with means:

$$\Lambda(A_i, T_i) = \int_{A_i} \int_{T_i} \lambda(\mathbf{s}, t) \mathrm{d}t \mathrm{d}\mathbf{s}. \tag{4.2}$$

Locally-integrable random fields $Z_{\mathbf{s},t}$ can be added to any linear predictor used to model the natural logarithm of $\lambda(\mathbf{s}, t)$. The point process is then said to be a Cox process driven by Z. If $Z_{\mathbf{s},t}$ is a Gaussian process, then $\lambda(\mathbf{s}, t)$ becomes a log-Gaussian random field and the process becomes known as a log-Gaussian Cox process (LGCP hereafter) [Simpson et al., 2016]. LGCP models are especially useful for modelling point-patterns when residual spatio-temporal correlations are expected to remain in the intensity, even after including any available covariates. In this case, the Gaussian process $Z_{\mathbf{s},t}$ is given a spatio-temporal correlation structure. Linear combinations of multiple Gaussian processes may also be included within a LGCP. We denote a set of Gaussian processes as $\mathbf{Z}_{\mathbf{s},t}$.

A base Cox process model is now introduced for describing the sampling process of S_t in many geostatistical settings. This model is very general. To ease the notational burden, only one latent effect is considered and denoted $Z_{s,t}$. Furthermore, for the remainder of the Chapter $Z_{s,t}$ is assumed to be a Gaussian process, although this constraint can be relaxed. Note that when the sampling process generating S_t deviates from the assumptions underlying the conditional Poisson process, other point processes should be considered. Details of other processes are found in Baddeley et al. [2015], Illian et al. [2008].

With a slight change of notation, removing the subscripts to improve readability, let $Y(\mathbf{s}, t)$ denote the observation process at location $\mathbf{s} \in \Omega$ and time $t \in \mathcal{T}$. This may be of any type (e.g continuous, count, binary, etc.). Let $\mu(\mathbf{s}, t)$ denote the target spatio-temporal process and let $Z(\mathbf{s}, t)$ denote a spatio-temporal latent Gaussian random field. As before, let S_t denote the collection of sampled points at time $t \in T \subset \mathcal{T}$. The following data-

generating mechanism is now assumed:

$$[Y(\mathbf{s},t)|\mathbf{s} \in S_t, Z(\mathbf{s},t)] \sim f(\mu(\mathbf{s},t), \boldsymbol{\theta}) \tag{4.3}$$

$$[S_t|Z(\mathbf{s},t)] \sim \text{IPP}(\lambda(\mathbf{s},t)) \tag{4.4}$$

$$g(\mu(\mathbf{s},t)) = \beta_0 + \boldsymbol{\beta}^T\mathbf{x}(\mathbf{s},t) + Z(\mathbf{s},t) \tag{4.5}$$

$$\log(\lambda(\mathbf{s},t)) = \boldsymbol{\alpha}^T\mathbf{w}(\mathbf{s},t) + \boldsymbol{\delta}(\mathbf{x}(\mathbf{s},t)) + h(Z(\mathbf{s},t)) \tag{4.6}$$

$$[Z(\mathbf{s},t)] \sim \text{GP}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{4.7}$$

Square brackets denote random variables. In equation (4.3), f represents the conditional probability distribution $Y(\mathbf{s},t)$, given the target latent spatio-temporal effect $Z(\mathbf{s},t)$, and given the location was sampled at time t (i.e. $\mathbf{s} \in S_t$). Values of $Y(\mathbf{s},t)$ are missing at all non-sampled locations. The link function g describes the relationship between the linear predictor and the target spatio-temporal process $\mu(\mathbf{s},t)$. Thus, the model contains the popular class of STGLMMs [Diggle and Ribeiro, 2007]. The regression equation for $\mu(\mathbf{s},t)$ is specified in (4.5), with fixed covariates $\mathbf{x}(\mathbf{s},t)$.

In equation (4.4), the sampling process $\mathcal{P}$ is modeled as a Cox process. When h is linear, the process is a LGCP with mean function $m(\mathbf{s},t) = \boldsymbol{\alpha}^T\mathbf{w}(\mathbf{s},t) + \boldsymbol{\delta}(\mathbf{x}(\mathbf{s},t))$. Note that, conditional on $Z(\mathbf{s},t)$, S_t is assumed to be a realisation from an IPP. Unique fixed covariates $\mathbf{w}(\mathbf{s},t)$ and shared covariates $\mathbf{x}(\mathbf{s},t)$ both describe the intensity of the conditional IPP. The shared covariates are transformed by functions $\boldsymbol{\delta}$ that may be nonlinear. The h function need not be linear, however when this is the case $\mathcal{P}$ is no longer a LGCP. In any case, h specifies the nature of PS.

The PS test developed in our Chapter is highly general. When h is strictly monotonic, the primary goal of the test is to detect the monotonicity of h. The precise form of h does not require specification. Suppose $h \equiv 0$ and at least one element of $\boldsymbol{\delta}$ is non-zero. The subset of covariates $\mathbf{x}(\mathbf{s},t)$ corresponding to these nonzero elements provide a sufficient set of informative covariates required to control for PS. The second goal of the test is to correctly identify this subset of covariates. Note that the covariance matrix

of the vector of $Z(\mathbf{s},t)$ values evaluated at S_t is denoted Σ. This also requires estimation. Finally, $\boldsymbol{\theta}$ are hyperparameters to be estimated in the model. Both Σ and $\boldsymbol{\theta}$ may be estimated using a maximum likelihood approach, or, given prior distributions, and then estimated under a Bayesian approach.

Including a sufficient set of informative covariates in a model for $\mu(\mathbf{s},t)$ should help to improve the prediction of $\mu(\mathbf{s},t)$ across Ω and $\mathcal{T}$, by reducing the deleterious impacts of PS on spatial prediction. However, it must be stressed that this approach is not a silver bullet; if the data were instead collected via a complete spatial randomness sampling design, then the predictive accuracy of the fitted model would likely be improved [Gelfand et al., 2012]. Thus, these methods should be viewed only as a partial remedy for badly sampled data rather than as a justification for ignoring the need for good spatial design of networks and surveys.

Finally, the conditional likelihood of the IPP given $Z(\mathbf{s},t)$ is:

$$\pi\left(S_t|Z(\mathbf{s},t)\right) = \exp\left\{|\Omega||\mathrm{T}| - \int_{\Omega}\int_{\mathrm{T}} \lambda(\mathbf{s},\mathrm{t})\mathrm{dtd}\mathbf{s}\right\} \prod_{\mathbf{s}_\mathrm{i}\in\mathrm{S_t},\mathrm{t}\in\mathrm{T}} \lambda(\mathbf{s}_\mathrm{i},\mathrm{t}), \qquad (4.8)$$

with $|\Omega|$ being the area of the domain Ω and $|T|$ being the length of the time set [Simpson et al., 2016].

4.2.2 Perceptron algorithm

Assume the above data-generating mechanism. A Monte Carlo algorithm is now designed for testing the null hypothesis that $h \equiv 0$, versus the alternative hypothesis that h is a monotonic function of Z. Under the null hypothesis $h \equiv 0$, the observation and sampling processes are conditionally independent given $\mathbf{x}(\mathbf{s},t)$. Thus, given $\mathbf{x}(\mathbf{s},t)$, no associations are expected to exist between computable quantities from the fitted (null) IPP and estimates of $Z(\mathbf{s},t)$.

Conversely, suppose that the null hypothesis is false. Specifically, let h be a monotonic increasing function of Z. Point-patterns S_t from this data-generating mechanism are expected to exhibit an excess of clustering

in regions of high $Z(\mathbf{s},t)$, relative to that explained by the null model. This phenomenon is referred to as positive PS. Here, a positive association between the localized amount of clustering and estimated $Z(\mathbf{s},t)$ values would be expected. The converse holds when h is a decreasing function.

The primary challenge is defining what constitutes a 'strong' association between estimates of $Z(\mathbf{s},t)$ and the computed quantities used to capture excess localized clustering. Positive spatio-temporal correlations are present in $\mu(\mathbf{s},t)$ due to $Z(\mathbf{s},t)$. This leads to non-standard sampling distributions for test statistics computed to capture association. Standard hypothesis tests of association (e.g. t-tests, rank-correlation tests, etc.) will have a type 1 error above the specified level due to the positive correlations.

This is why Monte Carlo methods are used. An empirical p-value associated with any desired test statistic can be computed by sampling realisations from the assumed IPP under the null hypothesis (i.e. fixing $h \equiv 0$). The application generalizes to any given dataset. Crucially, this procedure accounts for the nonstandard sampling distribution of the chosen test statistic in a natural way. The mean of the K nearest neighbour distances from each observed point is our default choice of computable quantity. Small values of this quantity within a region, indicates the presence of clustering there. When $K = 1$, this reduces to the nearest neighbour distance. For the default choice of test statistic, the Spearman's rank correlation coefficient between estimates of $Z(\mathbf{s},t)$ at locations $\mathbf{s} \in S_t$ and the mean nearest neighbour distances is proposed. This is specifically chosen to capture the degree of monotonicity of h.

We now state the probability distribution function of the (spatial) distance from any point of S_t to its nearest neighbouring point for the IPP data model (when $h \equiv 0$). Let $(\mathbf{s},t) \in \Omega \times \mathcal{T}$ define a reference space time point and let T be the time interval (t, t_T) of interest respectively. Next, define $b(\mathbf{s},r)$ as the ball of radius r centered at $\mathbf{s}$. Let $\lambda_{\text{IPP}}(\mathbf{s},t)$ once again denote the intensity function for the assumed IPP model under the null hypothesis $h \equiv 0$.

Theorem 4.1 *Assuming $h \equiv 0$ and the above data-generating mechanism,*

the probability that the nearest point of S_τ from $(\boldsymbol{s},t)$, lies within a spatial distance r, at some time $\tau \in T$ is [Matern, 1971]:

$$1-\exp\left(-\int_{b(\boldsymbol{s},r)}\int_{T}\lambda_{\text{IPP}}(\boldsymbol{\omega},\tau)\mathrm{d}\tau\mathrm{d}\boldsymbol{\omega}\right)=1-\exp\{-\Lambda(b(\boldsymbol{s},r),T)\}. \tag{4.9}$$

Equation (4.9) gives us the following intuitive result when $h \equiv 0$. The expected nearest neighbour distances are lower in regions of high intensity (i.e where $\lambda_{\text{IPP}}(\mathbf{s},t)$ is high). This is to be expected - the intensity function at $(\mathbf{s},t)$ precisely defines the expected density of points immediately around $(\mathbf{s},t)$.

The result can also been derived under the alternative hypothesis when h is linear. When $h(Z) = cZ : c \neq 0$ then the point process is a LGCP. Coeurjolly et al. [2017] derived the Palm distribution for LGCPs. The authors showed the remarkable result that conditional on a single point in S_t lying at location $(\mathbf{s},t) \in \Omega \times \mathcal{T}$, the remaining points of S_t are also a LGCP. Using the authors' result, this conditional process differs only in the mean function, with the covariance function remaining the same. The updated mean function describing this process is denoted $\mu_{(\mathbf{s},t)}(\cdot,\cdot)$. This can be thought of as representing the mean function of the LGCP, conditional on $(\mathbf{s},t) \in S_t$.

Theorem 4.2 *Assume that h is linear and assume the above data-generating mechanism. With $\Sigma(\cdot,\cdot)$ denoting the covariance function, the mean function of the LGCP conditioned on $(\boldsymbol{s},t) \in S_t$ $\mu_{(\boldsymbol{s},t)}(\cdot,\cdot)$ at $(\boldsymbol{\omega},\tau)$ is:*

$$\begin{aligned}\mu_{(\boldsymbol{s},t)}(\boldsymbol{\omega},\tau) &= \boldsymbol{\alpha}^T\boldsymbol{w}(\boldsymbol{\omega},\tau)+\boldsymbol{\delta}(\boldsymbol{x}(\boldsymbol{\omega},\tau))+c^2\Sigma(\boldsymbol{s}-\boldsymbol{\omega},t-\tau) \\ &= \mu(\boldsymbol{\omega},\tau)+c^2\Sigma(\boldsymbol{s}-\boldsymbol{\omega},t-\tau).\end{aligned} \tag{4.10}$$

Then, using the law of total expectation, the probability that the nearest point from $(\boldsymbol{s},t)$, lies within distance r, at some time $\tau \in T$ is:

$$1 - E_Z\left[\exp\left(-\int_{b(\boldsymbol{s},r)}\int_T \lambda_{\text{IPP}}(\boldsymbol{\omega},\tau)\left\{\exp\left[cZ(\boldsymbol{\omega},\tau)+c^2\Sigma(\boldsymbol{s}-\boldsymbol{\omega},t-\tau)\right]\right\}\mathrm{d}\tau\mathrm{d}\boldsymbol{\omega}\right)\right]. \tag{4.11}$$

Thus, by conditioning on a point of the process at location $(\mathbf{s},t)$, the only change to the mean function of the LGCP is the addition of the term $c^2\Sigma(\mathbf{s}-\boldsymbol{\omega},t-\tau)$. This is the covariance function between the conditioning point $(\mathbf{s},t)$ and the space-time point $(\boldsymbol{\omega},\tau)$, scaled by the squared linear coefficient c^2. If we make the following additional assumptions, then the interpretation of the (4.11) simplifies greatly.

Assumption 4.4 *The covariance is strictly non-negative, i.e.* $\Sigma(\cdot,\cdot)\geq 0$

Assumption 4.5 *Z is stationary and isotropic:* $\Sigma(\boldsymbol{s}-\boldsymbol{\omega},t-\tau)=\sigma_Z^2R(||\boldsymbol{s}-\boldsymbol{\omega}||,||t-\tau||)$

Assumption 4.6 *The correlation function decays monotonically as the distance from the conditioning point* $\boldsymbol{s}$ *increases:* $\sigma_Z^2R(||\boldsymbol{s}-\boldsymbol{\omega}||,||t-\tau||)\geq\sigma_Z^2R(||\boldsymbol{s}-\boldsymbol{\omega}||+\delta,||t-\tau||)\ \forall\delta>0$

Assumption 4.7 *The correlation function decays monotonically as the distance from the conditioning time t increases:* $\sigma_Z^2R(||\boldsymbol{s}-\boldsymbol{\omega}||,||t-\tau||)\geq\sigma_Z^2R(||\boldsymbol{s}-\boldsymbol{\omega}||,||t-\tau||+\delta)\ \forall\delta>0$

Under assumption 4.4, conditioning on point $(\mathbf{s},t)\in S_t$ will always increase the mean function and hence the intensity immediately around the $(\mathbf{s},t)$. Contrast this with the IPP. Here, the knowledge of a point of S_t existing at $(\mathbf{s},t)$ does not affect the intensity immediately around $(\mathbf{s},t)$.

Assumptions 4.4 - 4.7 are commonly made in practice. They imply that the latent process will be expected to be more similar at two space-time locations that are 'close together' than two space-time locations that are 'far apart'. Popular choices of correlation functions include the Matern correlation function across space [Diggle and Ribeiro, 2007] and autoregressive

correlation functions across time. Despite being unrealistic for most environmental processes [Cressie and Huang, 1999], spatio-temporal correlation functions are often defined to be the products of these spatial and temporal functions for computational simplicity [Blangiardo and Cameletti, 2015]. Under these models, the correlation often becomes negligible at spatial (temporal) distances greater than some value, often called the spatial (temporal) range.

Under Assumptions 4.4 - 4.7 and assuming $h(Z) = cZ : c > 0$, equation (4.11) now helps to explain the suitability of our choice of nearest neighbour distance to capture the excess clustering. Firstly, suppose the conditioning point $(\mathbf{s}, t)$ is in a region where the latent effect is above average (i.e. $Z(\mathbf{s}, t) > 0$). Here, we see that the expected nearest neighbour distance from $(\mathbf{s}, t)$ decreases monotonically relative to that expected from the IPP as: i) Z increases, ii) the correlation function $R(\cdot, \cdot)$ increases, and iii) c increases. When $c \equiv 0$, (4.11) equals (4.9) as expected.

Thus we have shown the result we wanted. Under Assumptions 4.4 - 4.7, with h a linear function with either a positive or negative slope, a monotonic association is expected between the nearest neighbour distances between the observed points $(\mathbf{s}, t) \in S_t$ and the values of Z at S_t. This monotonic association should be captured with our rank correlation test statistic when $K = 1$. Conversely, when $h \equiv 0$, no association is expected, so long as the fitted null IPP is correctly specified. In this case, the observed test statistic should be no more extreme than the Monte Carlo realisations. The test will generalize to nonlinear monotonic functions h, although the nearest neighbour probability function will not take the form (4.11). This does not cause problems for the perceptron algorithm, since the Monte Carlo simulations can still be performed.

To define the test performed by the perceptron algorithm requires some additional notation. Let T be a finite set of time intervals. Let n_t denote the observed number of points $\mathbf{s}_{i,t} \in S_t \subset \Omega$ at time $t \in T$. Let $N_{i,t}(K)$ denote the set of K nearest indices from each point $\mathbf{s}_{i,t} : i \in \{1, ..., n_t\}$. Let $\hat{Z}(\mathbf{s}, t)$ denote the estimate of $Z(\mathbf{s}, t)$. Define the superscript above each of these quantities m as the index of the Monte Carlo sample $m \in \{1, ..., M\}$.

Thus $N_{i,t}^m(K) : i \in \{1,...n_t^m\}, m \in \{1,...,M\}$ denotes the set of K nearest indices from point $\mathbf{s}_{i,t}^m$ in the m^{th} Monte Carlo sample S_t^m. Finally, let $\bar{D}_{i,t}(K), \bar{D}_{i,t}^m(K)$ denote the mean of the distances to the K nearest points from point i at time t in the original data and the m^{th} Monte Carlo sampled point-pattern respectively. Thus:

$$\bar{D}_{i,t}^m(K) = \frac{1}{K} \sum_{j \in N_{i,t}^m(K)} ||\mathbf{s}_{i,t}^m - \mathbf{s}_{j,t}^m||. \tag{4.12}$$

The terms $NN_{k,t}$ and $NN_{k,t}^m$ are defined to be the vectors of length n_t and n_t^m containing the values of $\bar{D}_{i,t}(K)$ and $\bar{D}_{i,t}^m(K)$ respectively. When calculating (4.12) in the original dataset, simply drop the m superscripts.

Algorithm 1: Perceptron NN test for PS in geostatistical data

Data:

Observations $\mathbf{y}(\mathbf{s},t)$ for $(\mathbf{s},t) \in (S_t \times T) \subset (\Omega \times \mathcal{T})$

Covariates $\{\mathbf{w}(\mathbf{s},t), \mathbf{x}(\mathbf{s},t)\}$ for $(\mathbf{s},t) \in (\Omega \times \mathcal{T})$

Result:

Empirical p-value for the test $h \equiv 0$ vs. h monotonic

begin

- Fit a model for (4.3) using a preferred method
- Produce estimates $\hat{Z}(\mathbf{s},t)$ throughout Ω, T
- Compute the $NN_{k,t}$ values $\bar{D}_{i,t}(K)$
- Evaluate $\hat{Z}(\mathbf{s},t)$ at locations S_t
- Compute the rank correlations ρ_t between $\hat{Z}(\mathbf{s}_i,t)$ and $\bar{D}_{i,t}(K)$
- Fit the chosen point process model with $h \equiv 0$ in (4.6)
- Fix $m = 1$
- **while** $m \leq M$ **do**
  - Sample n_t^m locations S_t^m from the fitted model for $t \in T$
  - Compute the $NN_{k,t}$ values $\bar{D}_{i,t}^m(K)$
  - Compute $\hat{Z}(\mathbf{s},t)$ at locations S_t^m
  - Compute the rank correlations ρ_t^m between $\hat{Z}(\mathbf{s}_i^m,t)$ and $\bar{D}_{i,t}^m(K)$
  - **if** $m = M$ **then**
    - return the empirical p-values of either pointwise or rank envelope tests using ρ_t and ρ_t^m.
  - **else**
    - $m \leftarrow m+1$
  - **end**
- **end**

end

The perceptron NN algorithm, referred to as the NN test hereafter, is defined above in Algorithm 1. It can now be summarized as follows. First, fit the assumed models (4.3) and (4.4) for both $Y(\mathbf{s},t)$ and S_t. Next, estimate $\hat{Z}(\mathbf{s},t)$ throughout $\Omega \times \mathcal{T}$ and compute the averaged K nearest neighbour distances $NN_{k,t}$. Using the estimates $\hat{Z}(\mathbf{s},t)$ at S_t and $NN_{k,t}$, compute the Spearman's rank correlation coefficient ρ_t between them for each $t \in T$. This is the observed test statistic.

Next, sample M realisations, $S_t^m : m \in \{1,...,M\}$, from the fitted point process model (4.4). For each of the M realisations, repeat the procedure. Compute the distances $NN_{k,t}^m$ and estimate the values $\hat{Z}(\mathbf{s},t)$ at S_t^m to obtain $\rho_t^m : m \in \{1,...,M\}$. Finally, compute the desired empirical p-value. For the pointwise tests, simply evaluate the proportion of the Monte Carlo-sampled ρ_t^m that are more extreme than ρ_t. For Monte Carlo envelope tests that do not suffer from the problems of multiple testing, refer to Mrkvička et al. [2017], Myllymäki et al. [2017].

4.2.3 Discussion

The values of the latent field Z are not known and must be estimated. The power of the test to detect PS may depend upon the suitability of the method used to produce estimates $\hat{Z}(\mathbf{s},t)$. Likelihood-based approaches for fitting the above observation model (i.e. components (4.3), (4.4) and (4.7)) have been shown to have many nice properties. Asymptotic consistency has been proven under the null hypothesis that $h \equiv 0$ for certain choices of f. These include Bernoulli [Ghosal et al., 2006] and Gaussian [Choi and Schervish, 2007] likelihoods. Such results help to justify the suitability of the method. Furthermore, under the null hypothesis $h \equiv 0$, some 'naive' approaches even produce unbiased estimates of $Z(\mathbf{s},t)$. For example, under the above data-generating mechanism, with f the normal distribution and g the identity function, the Gaussian process regression is the best linear unbiased predictor for $Z(\mathbf{s},t)$ [Cressie, 1992].

Other choices of a computable quantity for capturing spatial clustering can be made. These may be more suitable for certain data-generating

mechanisms of S_t. However, few choices are as generalisable across both continuous and discrete spatial data. For continuous spatial data, smoothed estimates of the residual measure from the fitted (null) point process, evaluated at the points $\mathbf{s} \in S_t$ may be suitable. However, this depends upon two tuning parameters: the details of the discretisation method chosen to approximate the likelihood and the choice of the bandwidth used to smooth the estimated values.

The nearest neighbour method does not suffer these drawbacks, and has many desirable properties in addition to its generalisability across both continuous and discrete spatial settings. For example, different choices of K can lead to improved powers to detect PS under different sampling processes. When the spatial scales of clusters within S_t are very small, and hence when clustering is very localized to only a few points per cluster, smaller K may lead to improvements in power. This is because larger values of K may 'smooth over' any clustering. Conversely, when clusters are large in spatial scale, with each cluster being comprised of several points, the power may be improved with larger choices of K. Here, the additional smoothing can reduce the variance of the computed test statistic (see Figure 4.1). Another beneficial property of the nearest neighbour quantity is that the distances can be computed exactly, with values not dependent upon any choice of computational approximation.

In some applications it may be suitable to fix the sample size across the Monte Carlo samples (i.e. $n_t^m \equiv n_t$). For example, regulatory standards may dictate the required number of samples n_t. The assumption of conditional independence between the sampled locations under the null IPP makes enforcing this condition easy.

Strictly speaking, since in practice the values of $Z(\mathbf{s}, t)$ and the parameters in (4.5) are not known and are only estimates, the plug-in test of Algorithm 1 will be invalid. This is because the null hypothesis is composite [Baddeley et al., 2017]. However, tests that ignore the effects of parameter estimation will tend to be conservative in most cases. A loss of power is typically the price to pay [Dao and Genton, 2014]. Whilst the method introduced by Dao and Genton [2014] can be used to ensure the test attains

nominal type 1 error, the required nested Monte Carlo simulations dramatically slows down the implementation. In the simulation studies of Section 4, the test defined in Algorithm 1 is found to be conservative across all tested simulation settings. Thus we do not consider this matter further.

P-values have come under increasing criticism recently (see Wasserstein and Lazar [2016] and references within). Indeed, the computation of an empirical p-value alone to identify the binary presence/absence of PS within a dataset has its flaws. For example, it does not help to quantify the potential magnitude of the biasing effects that the PS may have on the spatial prediction of $\mu(\mathbf{s},t)$. Furthermore, as with all p-values, it is easy to fall victim to the p-value fallacy. A given p-value does not provide much information on its own. A value close to 0.05 neither provides strong evidence in favour of the alternative hypothesis vs. the null hypothesis, nor implies that the frequentist error probability is close to 0.05 [Sellke et al., 2001]. Additional steps must be taken to make such inferences, such as the use of the calibrations introduced by Sellke et al. [2001]. In summary, as with all p-values, p-values reported from the tests outlined in algorithms 1 and 2 should be used with care.

Assuming the correct data-generating mechanism is specified and the true parameters are known, the test will be exact regardless of how small M is chosen. For testing at the 5% significance level, M could be chosen as low as 19. However, this comes at a cost of power, with the loss of power proportional to $1/M$ [Davidson and MacKinnon, 2000]. Furthermore, a small M implies a high standard error of the empirical p-value. This leads to a test whose outcome is heavily dependent on the precise sequence of random numbers used to implement the algorithm. To alleviate these concerns, M should be chosen as large as is computationally feasible.

4.3 Preferential sampling in the discrete spatial data setting

In the discrete spatial data setting, observations are taken across a set of areal units S_t within the study region Ω. Examples of areal units include

electoral districts and large survey transects. The sizes of these areal units may be irregular, and are assumed known. It is also assumed that the full population of all possible areal units that were available for sampling at each $t \in T$ is known. This population is denoted P_t. A binary process is fit to emulate the true sampling process. The choice between a Bernoulli and Binomial model depends on whether or not the constraint $n_t^m = n_t$ is implemented. Once again, a Monte Carlo approach is taken for testing for PS.

4.3.1 Assumed model for preferential sampling

Given the population of n_t areal units available at time t, denoted $P_t = \{\mathbf{A}_{i,t} \subset \mathcal{S} : i \in \{1, ..., n_t\}\}$, define the site-selection indicator variables as follows. Let $R_i(t)$ denote the indicator random variable that the i^{th} areal unit in P_t is selected at time t. Then the collection of sampled areal units $S_t \subset P_t$ at each time $t \in T$, is simply the subset of the population of areal unit whose indicator variables take value 1 (i.e. $S_t = \{\mathbf{A}_{i,t} \in P_t : R_i(t) = 1\}$).

Next, define $\bar{\mathbf{w}}(\mathbf{A}, t)$ to be the fixed spatio-temporal covariates for the indicator selection process at areal unit $\mathbf{A}$. These will typically be areal-aggregate or areal-count values. Similarly, $\bar{Z}(\mathbf{A}, t)$ and $\bar{\mathbf{x}}(\mathbf{A}_i, t)$ will typically be areal-aggregates of the underlying spatio-temporal process $Z(\mathbf{s}, t)$ and the spatio-temporal covariates $\mathbf{x}(\mathbf{s}, t)$ respectively. In applications, $\bar{Z}(\mathbf{A}, t)$ will typically be modeled as a discrete spatio-temporal process on the areal unit scale instead of as a continuous process. Examples include the conditional autoregressive process and its spatio-temporal extensions [Besag, 1974a, Blangiardo and Cameletti, 2015].

The same model form is assumed for the observation process $Y(\mathbf{A}, t)$ in (4.1). The only change made is the spatial scale. Thus, the new sampling process $\mathcal{P}$ is defined as:

$$[R_i(t)|\bar{Z}(\mathbf{A}_i, t)] \sim \text{Bernoulli}(p(\mathbf{A}_i, t)) \tag{4.13}$$

$$\text{logit}(p(\mathbf{A}_i, t)) = \boldsymbol{\alpha}^T \bar{\mathbf{w}}(\mathbf{A}_i, t) + \boldsymbol{\delta}^T \bar{\mathbf{x}}(\mathbf{A}_i, t) + h(\bar{Z}(\mathbf{A}_i, t)). \tag{4.14}$$

For each time step $t \in T$, it is assumed that each of the areal units $\mathbf{A}_i$ within the population P_t has values $Y(\mathbf{A}_i, t)$ sampled or not according to the outcomes of the independent Bernoulli trials defined in (4.12).

4.3.2 Perceptron algorithm

Algorithm 2: Perceptron test for PS in discrete spatial data

Data:
Observations $\mathbf{y}(\mathbf{A}_i, t)$ for $(\mathbf{A}_i, t) \in (S_t \times T)$
Covariates $\bar{\mathbf{w}}(\mathbf{A}_i, t)$, $\bar{\mathbf{x}}(\mathbf{A}_i, t)$ for $(\mathbf{A}_i, t) \in (P_t \times T)$
Result:
Empirical p-value for the test $h \equiv 0$ vs. h monotonic
begin
  Fit a model for (4.3) using a preferred method
  Produce estimates $\hat{\bar{Z}}(\mathbf{A}_i, t)$ across P_t and T
  Compute the $NN_{k,t}$ values $\bar{D}_{i,t}(K)$
  Compute $\hat{\bar{Z}}(\mathbf{A}_i, t)$ at areal units S_t
  Compute the rank correlations ρ_t
  Fit the chosen Bernoulli model with $h \equiv 0$ in (4.14)
  Fix $m = 1$. **while** $m \leq M$ **do**
    Sample n_t^m areal units S_t^m from the fitted model for $t \in T$
    Compute the NN distance measure $\bar{D}_{i,t}^m(K)$
    Compute $\hat{\bar{Z}}(\mathbf{A}_i, t)$ at areal units S_t^m
    Compute the rank correlations ρ_t^m
    if $m = M$ **then**
      return the empirical p-values of either pointwise or rank envelope tests using ρ_t and ρ_t^m.
    else
      $m \leftarrow m + 1$
    end
  end
end

All of the same assumptions and issues outlined earlier carry over to the discrete spatial setting. Once again, $Z(\mathbf{A}_i, t)$ must be spatially smooth across the areal units and estimates of $\bar{Z}(\mathbf{A}_i, t)$ must be available at each

of the areal units $\mathbf{A}_i$ in the population S_t at each time $t \in T$. Nearest neighbour distances between areal units within S_t can once again be used. Such distances can be defined relative to the areal unit-centroids or otherwise. Strictly speaking, the PS is no longer seen as a clustering of the point-pattern around high (or low) values of $\bar{Z}(\mathbf{A}_i, t)$. Instead, the PS takes the form of a clustering of the areal units with complete data around high (or low) values of $\bar{Z}(\mathbf{A}_i, t)$. The procedure is defined in Algorithm 2 below.

4.4 Simulation study

This section summarizes the key results of an investigation into the performance of the NN test. The power of the test is demonstrated across a range of simulated data settings. A more thorough treatment of the simulation study is provided in the supporting material found in the Appendix. All computations involving point processes were performed using the *spatstat* package [Baddeley et al., 2015].

The following data-generating mechanism is chosen for the Gaussian response simulation study:

$$[Y(\mathbf{s})|Z(\mathbf{s})] = Z(\mathbf{s}) \tag{4.15}$$

$$[S|Z(\mathbf{s})] \sim \text{IPP}(\lambda(\mathbf{s})) \tag{4.16}$$

$$\log(\lambda(\mathbf{s})) = \alpha_0 + \alpha_1 w(\mathbf{s}) + \gamma Z(\mathbf{s}) \tag{4.17}$$

$$[Z(\mathbf{s})] \sim \text{GP}(\mathbf{0}, \Sigma). \tag{4.18}$$

The simulated data are in the purely spatial setting (i.e. $\mathcal{T}$ is a singleton), with the $Y(\mathbf{s})$ specified as noise-free observations of $Z(\mathbf{s})$. $Z(\mathbf{s})$ is a realisation of a mean-zero Gaussian process with Matern covariance matrix Σ. The Matern roughness parameter ν is set to 1 and the standard deviation of $Z(\mathbf{s})$ is fixed at 1. The spatial range ρ_Z of the process is adjusted. The spatial range is defined here to be the distance at which the spatial correlation drops below 0.1. A larger ρ_Z implies the process has a greater spatial smoothness (i.e. a lower frequency).

The sampled locations are generated from a LGCP with a single (non-informative) covariate $w(\mathbf{s})$. The number of points is fixed equal to n, and thus the true process is a Binomial point process. The parameter γ determines the magnitude of PS, with $\gamma = 0$ corresponding to the null IPP model of no PS. Again, $w(\mathbf{s})$ is an independent realisation of another Gaussian process with Matern covariance function. Both the roughness ν and the standard deviation are again fixed at 1, but the range parameter ρ_w is varied independently from ρ_Z. The values of $w(\mathbf{s})$ are assumed known throughout Ω. The parameter α_1 determines the effect of the covariate on the intensity $\lambda(\mathbf{s})$.

The NN test is performed at the 5% significance level using 19 Monte Carlo samples (i.e. $M = 19$). $M = 19$ is chosen to allow for an exact 5% significance level to be attained, however this small value of M implies that these results provide a lower bound on the power of the test [Davidson and MacKinnon, 2000]; a higher power would have been attained had a larger M value been chosen. All tests are performed with the two-sided alternative hypothesis that h is a monotonic function of Z. Each experimental setting is repeated 200 times. In the study, all combinations of the following parameters are evaluated:

- Sample size $n \in \{50, 100, 250\}$,
- PS magnitude $\gamma \in \{0, 1, 2\}$,
- Covariate effect $\alpha_1 \in \{0, 1\}$,
- Spatial range of Z, $\rho_Z \in \{1.00, 0.20, 0.02\}$,
- Spatial range of w, $\rho_w \in \{1.00, 0.02\}$,
- Number of nearest neighbour distances $K \in \{1, ..., 15\}$.

Along with the NN test outlined in Algorithm 1, a Monte Carlo test using estimates of the raw residuals of the assumed IPP under the null hypothesis is also computed. This time the rank correlation between the

estimates of $Z(\mathbf{s})$ and the estimated residuals is the test statistic. As before, both sets of estimates are only evaluated at the point locations. The estimated residual values are simply kernel smoothed raw-residuals. An edge-corrected Gaussian kernel is used with bandwidth selected using leave-one-out cross-validation. We refer to this perceptron residual test as the residual test hereafter. It is interesting to assess the relative performance of the NN test, given its generality across all point processes and to the discrete spatial setting. We now summarize the results of the study.

First, our investigation strongly suggests that the type 1 error is bounded above by the nominal level. This is seen across all simulation settings when the null hypothesis is true. Thus, it appears that the computationally costly nested Monte Carlo approach of Dao and Genton [2014] need not be used, except in the interest of improving the power of the test. Fig. 4.2 shows the results for $n = 50$ in the simplest setting without covariate effects (i.e. $\alpha_1 = 0$) or PS effects (i.e. $\gamma = 0$). The spatial range ρ_Z is changed and the test is performed across different values of K. Both Monte Carlo tests attain Type 1 error at or below the 5% level. For comparison, the two standard simulation-free rank correlation tests attain a Type 1 error well above the 5% level. The error of these tests increases dramatically with ρ_Z because they ignore the spatial correlation and hence the non-standard sampling distribution of the test statistic. We omit the simulation-free results hereafter.

Next, the power is assessed. Across all the simulation settings, the power improved with increasing spatial range ρ_Z. This is in agreement with the earlier analytic results. $Z(\mathbf{s}, t)$ must be spatially-smooth to achieve high power. Furthermore, the power of the NN test is found to be sensitive to the choice of K value. Optimal choice of K depends upon both the spatial range of Z and the sample size. Larger values of both implies that a larger value of K should be chosen. Fig. 4.3 shows the results for the setting where no covariate effects exist (i.e. $\alpha_1 = 0$), but where moderate positive preferential sampling occurs (i.e. $\gamma = 1$). For $n = 50$, the NN test has a slightly lower power than the residual test. This difference diminishes as the sample size increases. Conversely, the NN test outperforms when both

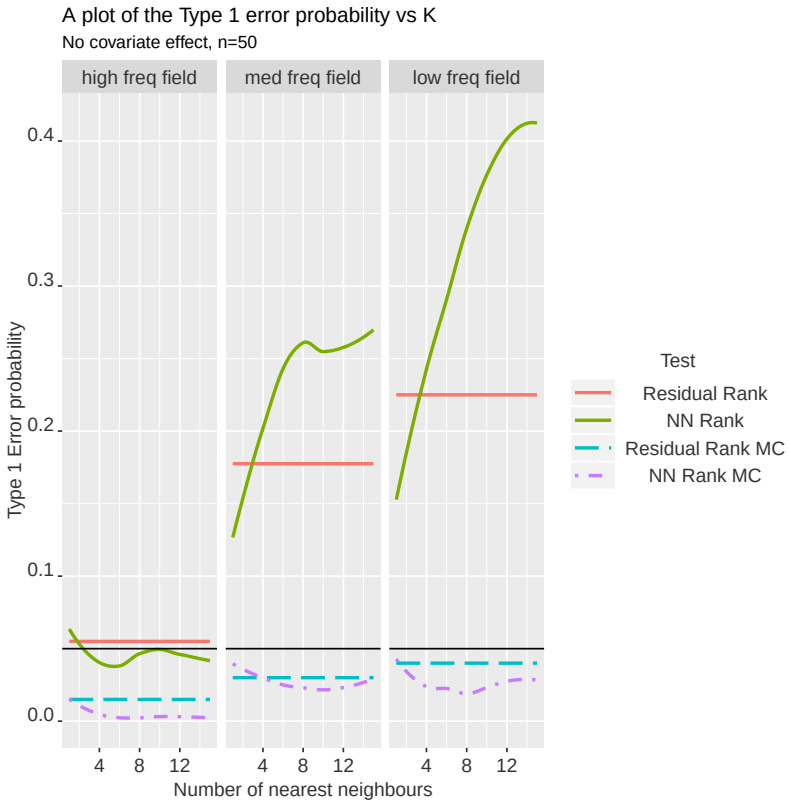

Figure 4.2: A plot of the Type 1 error for four tests. The three boxes show the results for $\rho_Z \in \{0.02, 0.2, 1\}$, from left to right respectively for a sample size of 50. The two 'Residual' tests are computed using the kernel-density smoothed values of the residuals from the fitted homogeneous Poisson processes. Leave-one-out cross-validation was used to select the bandwidth. The 'NN' tests are those based on the K nearest neighbour values. The suffix 'MC' denotes the test has been computed from Monte Carlo realisations of the fitted point process. A black line is plotted at the type 1 error level of 0.05 to indicate the target value. Notice that the simulation-free tests can have severely inflated type 1 error.

the spatial range is very small ($\rho_Z = 0.02$) and when the magnitude of PS is very high ($\gamma = 2$). Fig. A.17 in the Appendix demonstrates this. Under these conditions, very small clusters form. However, a different choice of

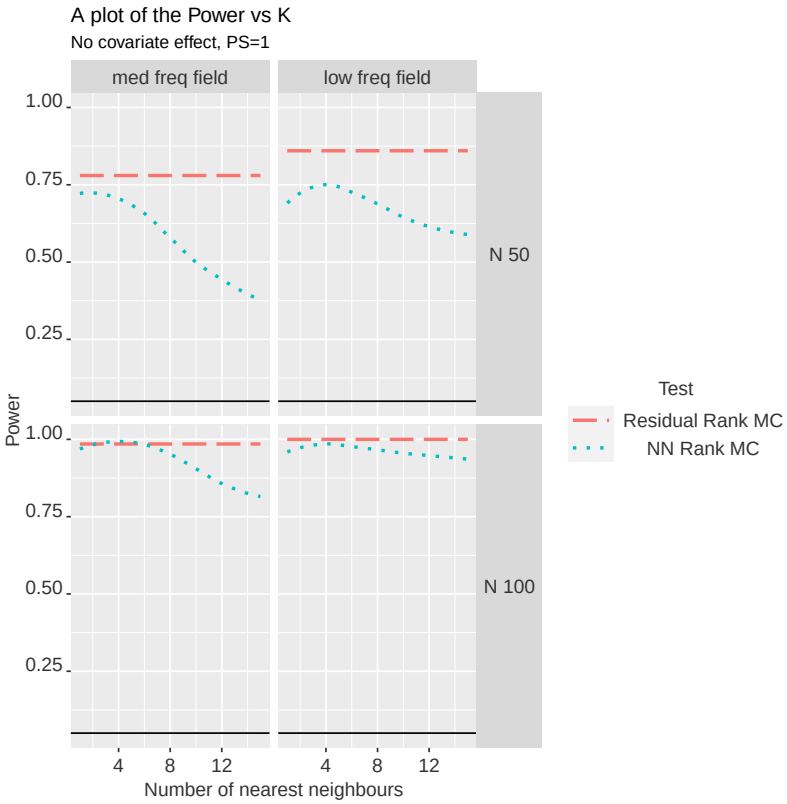

Figure 4.3: A plot of the Power for two tests when the PS parameter γ equals 1. The two columns show the results for $\rho_Z \in 0.2, 1$, from left to right respectively. The two rows show results for a sample size of 50 and 100 respectively. The 'Residual' tests are computed using the kernel-density smoothed values of the residuals from the fitted homogeneous Poisson processes. Leave-one-out cross-validation was used to select the bandwidth. The 'NN' test is based on the K nearest neighbour values. The suffix 'MC' denotes the test has been computed from Monte Carlo realisations of the fitted point process.

bandwidth-selection method may improve the power of the residual test.

Strong (non-informative) covariate effects in the sampling process hurts the power of all the tests. This can be seen in Fig A.18 in the Appendix. Interestingly, the NN test is shown to be competitive across all settings

tested, except one. When the spatial ranges of both the covariate $\mathbf{w}(\mathbf{s})$ and the field $Z(\mathbf{s})$ are large and similar, the power of the NN test is very low. Here, the residual test, with residuals computed from the fitted IPP, performs much better. The power is almost doubled that of the NN test. In these settings, despite the $\mathbf{w}(\mathbf{s})$ and $Z(\mathbf{s})$ arising from independent distributions, the empirical correlations between them in any given realisation may be large. Consequently, the NN test may be unable to distinguish between the clustering due to $Z(\mathbf{s})$, and the clustering due to the measured covariate $\mathbf{w}(\mathbf{s})$. This is because the NN test, unlike the residual test, does not directly adjust for $\mathbf{w}(\mathbf{s})$. However, the IPP residual test is not always superior. When the ρ_w is very low, the NN test has higher power when $\rho_Z = 0.2$.

The performance of the tests are also assessed in settings where the response is non-Gaussian, and when the true sampling process $\mathcal{P}$ is not an IPP. $Y(\mathbf{s})$ take the form of Poisson counts and the true sampling process is set equal to a Hardcore process. Different radii of interactions are compared. The Hardcore process is purposefully chosen. Here, the use of nearest-neighbour distances to capture additional clustering is poor. Since the nearest neighbour distances are lower bounded, the contrast in their observed values will decrease as the radius of interaction increases. Conversely, estimates of the smoothed residuals will not be directly affected. Figure A.21 in the Appendix clearly shows that the test based on the smoothed residuals far outperforms the NN test when the radius of interaction is high. This demonstrates a clear need for the researcher to choose a measure of clustering that is suitable for the true sampling process. The power remains high for Poisson f.

4.5 Case studies

The ability of the test to detect PS in two real case studies is now demonstrated. These two datasets are chosen since the presence of PS within them has previously been shown in published work. The first example shows how researchers can easily detect positive PS and then search for a sufficient

set of informative covariates $\mathbf{x}(\mathbf{s}, t)$ using the test. In the latter example, negative PS is detected.

4.5.1 Great Britain's black smoke monitoring network

Annual concentrations of black smoke were obtained from the UK National Air Quality Information Archive (airquality.co.uk). Site locations and annual average concentrations of black smoke (μgm^{-3}) were obtained from the monitoring sites. Previous analyses and the results presented in Chapter 3 demonstrated that the network had been preferentially sampled, with h a strictly increasing function across the years of 1966-1996 [Shaddick and Zidek, 2014].

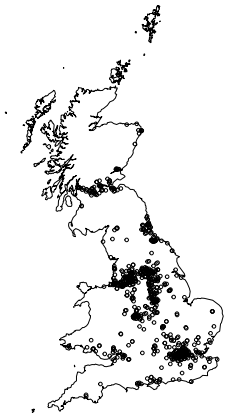

Figure 4.4: A plot of the locations of the black smoke monitoring sites in 1966. Observe the clustering of sites around the populous cities of London, Manchester, and Glasgow.

The analysis is restricted to the 1966 concentrations. For reasons outlined in [Shaddick and Zidek, 2014], only sites that gave readings for at least 273 days of the year (75% data capture) are considered. Fig. 4.4 shows that it is apparent that sites were located in the industrial cities around London, the Midlands and the North West of England, with almost no sites present in

the industry-free Scottish Highlands. Thus the point-pattern displays clear evidence of clustering around regions expected to have high concentrations of black smoke.

The *R-INLA* package (see Chapter 2) with the SPDE approach is used to fit a standard geostatistical model [Lindgren et al., 2011b, 2015, Rue et al., 2009]. Other methodologies, Bayesian or frequentist, could be used. As in Chapter 3, the log-ratios of the concentrations are the choice of response. PC priors are placed on the approximate Matern field [Fuglstad et al., 2018]. A prior probability that the spatial range is below 5km is set to 0.1, and a prior probability that the standard deviation of the field is above 3 is set as 0.1. The (mean-centered) posterior means of the log-transformed black smoke levels across Ω were then used as the 'naive' $\hat{Z}(\mathbf{s})$ values.

Gridded residential human population count data with a spatial resolution of 1 km x 1 km were obtained for Great Britain. These counts came from the 2011 Census data and 2015 Land Cover Map data from the Natural Environment Research Council Centre for Ecology & Hydrology [Reis, 2017]. As in Chapter 3, it is assumed that the relative population density across Great Britain remained approximately stable from 1966-2011. These normalized counts are used as an informative covariate $x(\mathbf{s})$ for both the null IPP sampling process and the observation process. Initially, the presence of PS is tested for without the population density covariate included in either process. This test is referred to as V1. Next, population density is controlled for as an informative covariate in both the observation and the null IPP sampling processes (it is found to be strongly associated with both). It is then investigated if PS remains. This test is referred to as V2.

Estimates of $\hat{Z}(\mathbf{s})$ in the V2 test are thus corrected for population density. Population density is included as a Bayesian spline to capture any nonlinear effects of population density on the observed black smoke $Y(\mathbf{s})$. Mechanistically, including population density as an informative covariate in a model for black smoke concentration is reasonable. Localized sources of black smoke include the combustion of carbon-based fuels, with expected levels of combustion expected to increase with population density. If PS is no longer detected after this adjustment, then the population density has

explained away the PS. Conversely, if PS is still detected, then standard 'naive' methods may be biased even after controlling for population density.

Table 4.1 shows the empirical pointwise p-values of the tests with changing K, under both of the assumed sampling mechanisms. Results are shown for both the V1 and V2 tests. The empirical p-values are the proportion of rank correlations in the 1000 Monte Carlo samples, that were more negative than observed in the data. Thus this test is 1-sided. For $K > 1$, strong evidence is found against the null in favour of a positive monotonic h under the V1 HPP test. This remains under the V2 IPP sampling process. Note that the p-values have not been adjusted for multiple testing (over K).

Table 4.1: A table of empirical p-values for the UK black smoke dataset for both the assumed homogeneous and inhomogeneous Poisson point process models.

K	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
HPP P value V1	0.28	0.01	0.01	0.01	0.00	0.00	0.01	0.01
IPP P value V2	0.20	0.01	0.01	0.01	0.00	0.00	0.00	0.01

The insights gained from these tests are as follows. A 'naive' model fit to the black smoke data without adjusting for population density may be biased due to PS being present. Furthermore, population density fails to explain the observed PS seen in the data. Residual PS remains after controlling for population density in a model for black smoke (IPP V2 result). Thus a sufficient set of PS-removing covariates has not yet been identified. Either this iterative process of finding a sufficient set of covariates must be continued, or a joint model should be considered as in Chapter 3.

4.5.2 Galicia lead concentrations

The second real-world dataset consists of the concentrations of lead in moss samples collected across Galicia, northern Spain, in 1997 [Fernández et al., 2000]. The concentration is measured in micrograms per gram of dry moss. The 1997 locations were previously shown to have been preferentially sampled in the landmark preferential sampling paper by Diggle et al. [2010]. In

fact, in this example, a significant linear h effect with negative slope was found. Thus h was found to be monotonically decreasing.

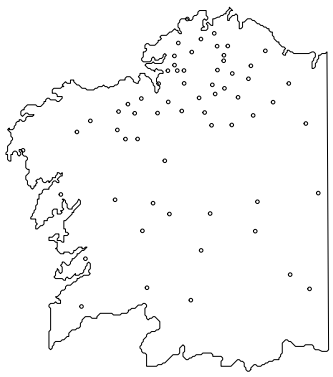

Figure 4.5: A plot of the 1997 sampled locations of lead concentrations in Galicia, northern Spain. Observe the clustering of sites in Northern Galicia.

Fig. 4.5 shows a clear increase in the density of sampled locations in the northern half of Galicia. This half was found to have the lowest concentrations of lead. PC priors were again placed on the approximate Matern field. A prior probability that the spatial range is below 10km was set to 0.1, and the prior probability that the standard deviation of the field is above 3 was set to 0.1. The posterior mean log-transformed lead concentrations were used as the $\hat{Z}(\mathbf{s})$ values. Empirical p-values were computed using 1000 Monte Carlo samples. The direction of the inequality was reversed this time, to test for negative PS.

Table 4.2: A table of empirical p-values for the Galicia dataset.

K	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
P value	0.01	0.03	0.06	0.06	0.07	0.08	0.08	0.09

The empirical pointwise p-values of this test are shown in Table 4.2. Even after considering Monte Carlo error, moderately strong evidence of negative PS exists, with the strongest evidence occurring at $K = 1$. A researcher would now have to decide whether to pursue a sufficient set of covariates or fit a joint model as in Diggle et al. [2010].

4.6 Concluding remarks

A fast and intuitive test has been presented for detecting preferential sampling (PS) in both continuous-space and discrete-space spatio-temporal data. The test is highly general; any preferred methodology for estimating a latent spatio-temporal process $Z(\mathbf{s}, t)$ may be used. This includes both Bayesian and frequentist methods. In many situations, detected PS may be adequately described by a set of available informative covariates that are associated with both the sampling process and the observation process being measured. The test presented in this Chapter can identify such an informative covariate set. In cases where residual PS remains, even after controlling for a set of informative covariates, researchers can either seek a sufficient set of informative covariates, or seek a method that directly models the PS (e.g. a joint model).

In this Chapter, the properties and validity of the test were demonstrated through an extensive simulation study. The suitability of the test for real-world applications was then confirmed through the re-analysis of two previously published case studies. Both case studies had previously detected PS, and the test successfully replicated the findings of both. The power of the test to detect PS was shown to increase with: the spatial range (i.e. the inter-point distance required for observations to be approximately independent), the sample size, and the degree of PS. The power decreased dramatically with the inclusion of covariates, independent from the observation process, that had a strong effect on the sampling process.

The general framework introduced in Chapter 3 could also have been used to identify PS in both of the two case studies. However, we recommend that the test developed in this Chapter be used instead for the purposes of

testing. Unlike the general framework, the test requires only that the PS be monotonic. This could potentially have large implications for inference. An incorrectly specified functional form of PS within the general framework could lead to it failing to detect PS altogether. Unfortunately, the high computational cost of fitting the framework prohibited its comparison within the simulation study. Furthermore, the three tests previously developed in the literature would be unsuitable for use in the two case studies. Whilst the two case studies both involved continuous responses, the two tests developed by Schlather et al. [2004] would only be able to accept or reject the assumption that the data were a realization of a random-field model. The direction of PS would not be discernible. Additionally, population density could not have been integrated within their test. The test by Guan and Afshartous [2007] would not be suitable since the sample sizes seen in the two case studies were orders of magnitude too small for its use.

The biasing effects of PS on spatial prediction has been demonstrated across a wide range of fields and can be severe in magnitude [Diggle et al., 2010, Dinsdale et al., 2019, Pennino et al., 2019]. Thus, PS should not be ignored in spatial analyses. The perceptron test proposed in this Chapter is suitable for assessing the presence of PS in many spatio-temporal analyses. Therefore, in cases where the sampling protocol is either unknown or known to be preferential, reporting the results from a PS test alongside any publication of spatio-temporal analyses should become standard practice. We have made a user-friendly R package, *PStestR*, available on GitHub for implementing the algorithm. No extra work or computation time is required. The package works seamlessly with many of the commonly used data types (e.g. from the *sp*, *sf* and *spatstat* libraries [Baddeley et al., 2015, Bivand et al., 2013, Pebesma, 2018, Pebesma and Bivand, 2005]). The test can perform both pointwise and rank envelope tests, to alleviate the multiple testing problem. The package can also fit numerous point processes, including Hardcore and Cluster point processes.

Three avenues of research should be pursued in the future. First, how to best capture the localized clustering in specific sampling settings should be explored. For example, in this Chapter it was shown that the nearest neigh-

bour distance may be a poor choice for measuring the degree of clustering under certain sampling processes. How to optimize the power of the test in specific settings should be pursued. Next, eliminating the need for generating Monte Carlo samples from the null point process may be possible [Acosta et al., 2018]. We found that 'effective-sample size'-adjusted rank correlation tests showed very poor performance. Convergence rates of the test could be as low as 10% for specific simulation settings. Thus, we omitted the results. Further investigation here is warranted. Finally, adjustments are required for using this test in spatio-temporal applications where sampling locations are retained from one time-step to the next. Environmental monitoring networks are an example. Here, the chosen network locations from one time step to the next are not independent. Additional work should be pursued to generalize the methods in this Chapter to such settings.

Chapter 5

Estimating Animal Utilization Distributions from Multiple Data Types: a Joint Spatio-temporal Point Process Framework

"In our lust for measurement, we frequently measure that which we can rather than that which we wish to measure... and forget that there is a difference."
— George Udny Yule

A preview

The previous two Chapters developed methods for accounting for preferential sampling (PS) in the statistical analysis of both discrete-space, and continuous-space, spatio-temporal data. In Chapter 2, we introduced a third type of spatio-temporal data, which we referred to as spatio-temporal point-pattern data. Accounting for PS in the statistical analysis of spatio-temporal point-pattern data is the focus of this Chapter.

Accounting for PS in spatio-temporal point-pattern data offers additional challenges compared with the other two data types. No longer can the observed locations themselves be used directly to inform a model about the nature of PS through the sharing of latent effects between processes. Instead, additional strong assumptions are required. We assume the scenario where a point-pattern is observed in $\Omega \times \mathcal{T}$. We assume that the point-pattern represents a censored collection of events that are recorded by a set of observers. We assume that numerous factors caused the true point-pattern to be only partially observed throughout $\Omega \times \mathcal{T}$. By borrowing ideas from thinned point processes, we multiplicatively decompose the intensity of the observed (censored) point process into three terms. The first term describes the true spatio-temporal intensity process responsible for generating the latent, uncensored point-pattern. The next two terms reflect the processes that lead to some of the events (points) to become censored. The first of these terms describes the behaviours of the observers through space and time. We refer to this term as an effort surface and it reflects the observers' efforts spent searching for the events. The final term then reflects the abilities of the observers to detect an event at location $\mathbf{s}$ and time t, conditional on their being one there. We refer to this term as a detectability surface.

For modelling, we consider the setting where researchers have additional knowledge, either in the form of a set of informative covariates or an external emulator, that can spatio-temporally describe both the behaviours of the observers, and the detectability of the events, sufficiently well. The application we consider is the modelling of the spatio-temporal distributions of animals. In particular, we develop a framework for estimating the utilisation distribution (UD) of animals that can adjust for both heterogeneous observer effort and heterogeneous detectability. These issues are commonplace in ecological applications. We then apply our framework to a real case study: the estimation of the spatio-temporal distribution of an endangered ecotype of killer whales using opportunistic sightings made by citizen scientists. As beautifully stated in the above quote, we must adjust for the opportunism of these citizen scientists if we wish to draw sensible inference

from their sightings.

5.1 Introduction to Chapter 5

Accurate knowledge of the spatio-temporal distribution of animals is vital for understanding the effects of climate and habitat changes on species and for governments to implement successful management policies. In particular, ecologists are interested in estimating the space use of individual animals to inform a wide range of questions. For example, such estimates can be used to quantify the location and extent of habitats required for a species' conservation strategy [Fleming et al., 2015]. Estimates of space use can be linked to environmental covariates within an occupancy model to explain resource use [Mordecai et al., 2011]. These estimates are also used within a spatial capture-recapture framework for estimating home range centers and population abundance [Royle et al., 2011]. The space use of an individual is often referred to as the utilization distribution (UD). The UD of an individual defines the probability density that an individual is found at a given point in space and time [Lele et al., 2013]. Estimating individual UDs is often complicated by complex animal movement and observer processes, both of which must be considered within an analysis [Royle et al., 2011]. Accounting for observers with unknown, but estimable, effort remains an open problem for UDs.

Ecologists have a plethora of methods available to them for estimating UDs from different types of individually-identified sightings. In particular, many methods exist for animal tracking datasets where the locations of individuals are followed through time with devices such as GPS tags. Simple methods include geometric techniques such as the minimum convex polygon [Fieberg and Börger, 2012], and statistical smoothers such as kernel density estimators [Worton, 1989]. Methods have also been developed to account for the autocorrelations due to animal movement [Fleming et al., 2015], but these assume that the sightings were made without observer bias. For tracking data collected at high temporal frequency, resource selection function models can be used to estimate UDs [Johnson et al., 2013]. Fewer

methods exist for sightings (or captures) data made at a set of discrete locations (e.g. camera traps) or by a group of mobile observers with recorded locations. Such datasets are commonly collected [Hussey et al., 2015, Whoriskey et al., 2019]. In these settings, spatial capture-recapture models can be used to estimate UDs throughout the study region [Royle et al., 2011]. However, these methods typically assume that the UD of each individual follows a bivariate normal distribution centered at a unique, latent, home range center. This assumed form of the UD may be overly simplistic when large quantities of data are available for each individual. We propose a framework to model complex individual UDs that accounts for the observer efforts from both mobile and static observers and models the relationships between the individuals' space use and environmental characteristics.

A similar but distinct objective to estimating the UDs of individuals is estimating the distribution of a species. Species distribution models (SDM's hereafter) are tools for predicting the density of a species and for relating it to measured environmental covariates. SDMs have been the focus of statistical research for decades [Elith and Leathwick, 2007, Fithian et al., 2015]. In particular, much work has been done on how to use point process methods to develop SDMs that can combine data of varying type (see Miller et al. [2019] for a recent review), account for observer processes and biases [Koshkina et al., 2017], and include spatio-temporal random effects to capture additional autocorrelations [Yuan et al., 2017]. Furthermore, point processes are scale-invariant and do not encounter the 'pseudo-absence' problem faced by competing methods [Warton et al., 2010]. Thus, point processes have emerged as the most promising integrative model framework for jointly modelling data of varying types and quality [e.g. Giraud et al., 2016, Renner et al., 2015]. By sharing parameters and latent effects between the likelihoods of each data type within a joint model [e.g. Bedriñana-Romano et al., 2018, Giraud et al., 2016], strength can be borrowed and hence a greater precision in conclusions and improved management policies can be attained [Fithian et al., 2015]. This approach can be made more robust for datasets of especially poor quality, by allowing only a correlative relationship to exist between their models and the latent effects defining the SDM [Pacifici et al.,

2017].

In this Chapter, we show how to adapt these recent point process frameworks for use with UDs. In particular, we assume that the individuals' UDs are stationary within well-defined time periods and that we can subset the sightings data to obtain approximately independent snapshots of the UDs. We then argue that these recent developments make point processes an ideal basis to create a framework for modelling individual UDs in data-rich settings. In particular, using point processes for UDs allows us to combine numerous datasets following different complex data collection protocols [Miller et al., 2019]. For example, presence-only sightings that lack any records of locations where sightings were not made (i.e. absences) may be combined with high quality presence-absence data (see Hefley and Hooten [2016], Miller et al. [2019] and references within). Accounting for differences in protocols is crucial. Some protocols may focus observer efforts in regions where the density of the species under study is highest. This preferential sampling can lead to positively biased estimates of species density, abundance, and UDs [Pennino et al., 2019].

With the data appropriately subsetted, any autocorrelation between the sightings that could bias inference is removed [Johnson et al., 2013]. Yet, there remain two hurdles that we must overcome to apply point process methods to UD estimation. First, we must adapt the point process framework to the setting where individuals can be identified at numerous locations through time. Typical SDM analyses often assume that the locations of individuals remain fixed throughout the sampling time [Giraud et al., 2016, Koshkina et al., 2017]. Second, we must generalize the point process framework to allow for combinations of mobile and static observers with highly complex, and potentially unknown (but estimable) effort. Well-defined discrete sampling 'sites' are often assumed to exist [Giraud et al., 2016]. However, in many cases, observer effort is continuous in both space and time. In continuous-time, unless strict distance sampling protocols with high observer speed are followed, animal movement can bias estimates of absolute intensity [Glennie et al., 2015, Yuan et al., 2017]. Explicitly modelling the animal movement model within the observer's sampling process to elimi-

nate the bias caused by animal movement can prove challenging [Glennie et al., 2020]. This task is made especially difficult when information on the sampling protocols is unavailable, as is often the case.

Taking the above concerns into account, we create a flexible framework for estimating the UDs of individuals in data-rich settings. We relax many of the stringent assumptions and requirements made previously, creating a general approach for estimating the spatio-temporal distributions of individual animals. Data from combinations of mobile and static observers, with differing skill levels, and following differing protocols, may be combined. Locations of observers may be known (e.g GPS), or unknown. In the latter, either a highly informative set of covariates or an emulator must be available that can adequately describe the observers' efforts. In either case, observer effort can be controlled for, with any uncertainties in effort propagated through to the resulting inference. Our approach circumvents the modelling of an animal's movement process by focusing on relative intensity values when estimating UDs. This approach largely avoids the bias discussed by Glennie et al. [2020] for estimating absolute intensity values. We demonstrate this claim in a thorough simulation study and show that large improvements in the predictions of UDs are possible using our approach in settings where the degree of spatial heterogeneity in observer effort is high.

Using our approach, statistical inference can be made at both an individual level, and/or a population level. We show that population inference is trivial in settings where sightings data of each individual in the population are available. When only a subset of the population is observed, the sampled subpopulation must be representative of the population for extrapolation to be accurate. To adapt the point process framework for use with repeated sightings of individuals, we redefine the intensity surface being modeled as an encounter rate instead of as an expected number of individuals per unit area. Including random fields and random effects can model the additional spatio-temporal correlation induced by unmeasured spatially-smooth covariates and/or biological processes [Pacifici et al., 2017, Yuan et al., 2017]. Since this approach is subsumed under a log-Gaussian Cox process (LGCP) framework [Chakraborty et al., 2011], implementing this idea is made especially

easy using the R package inlabru [Bachl et al., 2019, R Core Team, 2019], enabling researchers to adapt this framework for use in their applications.

The paper is structured as follows. First, we present our motivating problem: estimating the spatio-temporal distribution of the Southern Resident Killer Whale between May and October. This species is of special conservation concern. Next, we define the types of encounter events assumed to generate the data. Then, we introduce marked LGCPs as the preferred tool for modelling the encounters. We discuss their properties and define the new intensity surface to be modeled in terms of encounter rates and effort intensities. After presenting our modelling framework, we demonstrate its properties in a thorough simulation study, before adapting it to our motivating problem. Finally, we design easily-interpretable maps to display the UDs of the whales across the months. We then aggregate these to provide inference on both a pod (i.e. group) level and a population level. Computer code using the inlabru package [Bachl et al., 2019] is provided to enable researchers to adapt this framework for their applications.

5.2 Motivating problem

5.2.1 An introduction to the problem

The Southern Resident Killer Whale (SRKW) population is listed as Endangered under both the Canadian Species at Risk Act (DFO) and the United States Endangered Species Act (NOAA) because of their small population size, low reproductive rate, the existence of a variety of anthropogenic threats, and prey availability. The range of the SRKW extends from southeastern Alaska to central California; however, between May - September, all three pods of these whales frequent the waters of both Canada and the United States, concentrating in the Salish Sea and Swiftsure bank (DFO). We extend our study region beyond these areas (see Fig 5.1).

The development of successful and effective policies to help protect the SRKW requires accurate, high-resolution knowledge on how their use of space evolves across the calendar year. The SRKW are highly social ani-

mals, spending the majority of their time in three well-defined groups called the J, K and L pods [Ford et al., 1996]. Hereafter, we consider pods as individuals for the purposes of modelling. Inter-pod variation in the summer space-use exists [Hauser et al., 2007]. Due to the differing characteristics of the three pods, additional knowledge of space-use on a pod-level may help to improve the effectiveness of management decisions. While SRKWs are known to favour the inshore waters of Washington State and British Columbia in the summer months [Ford et al., 2017, Hauser et al., 2006], precise knowledge surrounding their space use across the months is lacking, as is precise knowledge surrounding the differences between the pods [Hauser et al., 2007].

5.2.2 The data available

Multiple sources of SRKW sightings are available, including GPS-tracked focal follows of targeted individuals by citizen scientists, presence-only sightings from commercial whale-watching vessels, and opportunistic sightings reported by the public. We judge two of these data sources to be most suited for use in an effort-corrected analysis as we are able to estimate its observer effort and we are confident that the source could accurately differentiate between the three SRKW pods.

The first source of data we use are presence-absence data collected through a project funded by the Department of Fisheries and Oceans Canada. The data were collected from a mobile vessel on the west side of the Strait of Juan de Fuca fitted with a GPS tracker between 2009 and 2016. The GPS coordinates of the vessel were recorded at a high frequency, and all sightings of SRKWs were reported, along with the pod (see Fig 5.1). The marine mammal observer on this vessel is an expert on the SRKW, and thus, the reported pod classification can be considered accurate. The motivations of the Captain's data collection varied from year to year, and hence, the spatial distribution of their observer effort varied. We refer to this dataset as DFO hereafter.

The second reliable data source we use are the presence-only SRKW

sightings reported by the whale-watch industry between 2009 - 2016 and collected by two organisations. The B.C. Cetacean Sightings Network (BCCSN) [Vancouver Aquarium] and The OrcaMaster (OM) [The Whale Museum] [Olson et al., 2018] datasets both contain sightings from a vast range of observer types, but we exclusively model the whale-watch sightings for three reasons. First, the whale-watch operators have a high degree of expertise on the SRKW, with vessels typically having a biologist or other expert onboard. This results in accurate pod classifications. Second, the whale-watch companies are known to share the locations of the sighted SRKW with each other. Thus, our dataset likely contains the majority of whale-watch sightings that were made, not just the subset of those made by the operators who report to the databases. Third, a vast amount of data has been collected on the activities of the whale-watching industry operating in the area. This enables us to estimate the observer effort from these companies with a high degree of accuracy and precision. We refer to this combined dataset as WW hereafter.

The majority of the data we obtained on the activities of the whale-watch industry came from the Soundwatch Boater Education Program (Soundwatch hereafter). Soundwatch is a vessel monitoring and public education outreach program that systematically monitors vessel activities around cetaceans during the whale-watch season (May - September) in the Haro Strait Region of the Salish Sea [Seely et al., 2017]. Since 2004, Soundwatch has been using data collection protocols established in partnership with NOAA, DFO, and the Canadian Straitwatch Program. This includes detailed accounts of vessel types, whale-watch vessel numbers, and whale-watch vessel activities.

5.2.3 Previous work estimating the space use of SRKW

Previous work estimating the summer space use of SRKW has greatly assisted the development of critical habitat regions and has helped inform successful management initiatives. In the past 15 years, two pieces of published research tackled the problem from different angles. Hauser et al. [2007]

estimated the summer space use of the SRKW within the Salish Sea's inshore waters of Washington and British Columbia. The authors assumed a constant observer effort across the months from the whale-watch operators within their study region, based on results from a field study [Hauser et al., 2006]. Pod-specific core areas were identified and SRKW hotspots were clearly displayed. However, their study region was smaller than ours.

Most recently, Olson et al. [2018] estimated an effort-corrected map of SRKW summer space use. This expanded on the work of Hauser et al. [2007], by estimating the space use across a larger area than theirs, as well as by incorporating the heterogeneous observer effort in their modelling directly. Regions of 'high' effort-adjusted whale density were identified and clearly presented in detailed plots. To reduce the impact of autocorrelation from the whales' movements on the analysis, they defined 'whale days' as their target metric. They defined a whale day to be any day where a SRKW was reported, regardless of the number of times they were reported on that day. A smaller study region relative to ours was studied, and no environmental covariates were used. Estimation of observer effort followed previous unpublished research from the Vancouver Aquarium (Erin U. Rechsteiner [2013], *pers comm*).

5.2.4 Goals of the analysis

We aim to build upon the previous research and build a model for estimating high-resolution, effort-corrected, and temporally-changing SRKW space use across the summer months (May - October). For each pod, their UD will provide the probability that they occupy a specific region at any given instant in time within each month. Under the assumption that the study region properly captures the full spatial extent of the UDs, we can then estimate the spatial distribution of the SRKW for each month. All estimates will be conditioned upon detailed estimates of observer effort from the whale-watch companies, combined with GPS tracks from the DFO dataset. Unlike previous attempts, our model will use multiple environmental covariates such as sea-surface temperature and various measures of primary productivity (e.g.

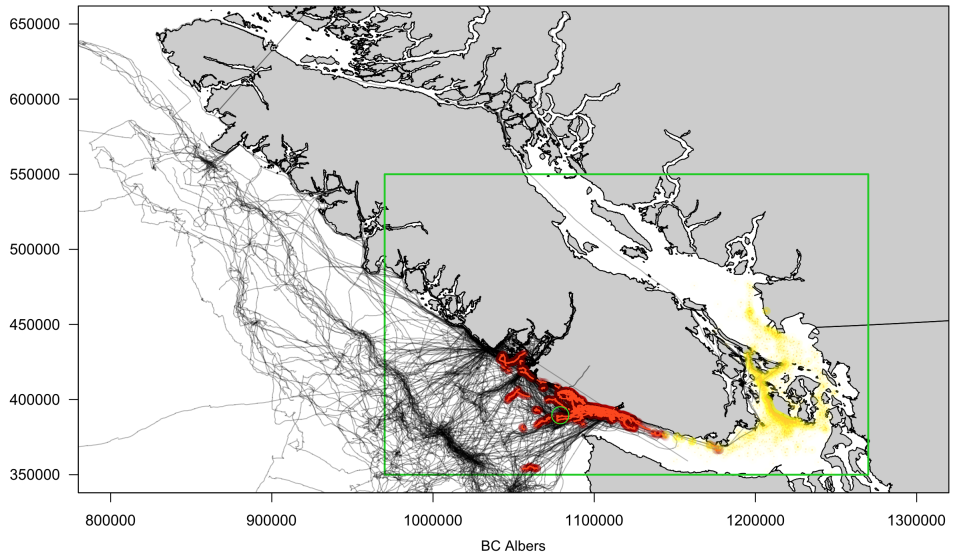

Figure 5.1: A plot showing our area of interest Ω in green, with the GPS tracklines of the DFO survey effort displayed as black lines. All DFO survey sightings are shown as a red overlay on top of the effort. All sightings from the OM and BCCSN datasets are shown in yellow. Shown are sightings and tracklines from May - October 2009 - 2016. The green circle roughly locates the Swiftsure Bank, with the waters due East and North of this representing the Salish Sea. The BC Albers projection shown is in units of metres.

chlorophyll-A) to improve the accuracy of predicted maps.

By turning to statistical/probabilistic modelling, we will attempt to account for all sources of uncertainties, including the uncertainties associated with our estimates of the observer effort from the whale-watch vessels. This has not previously been done. Finally, we will demonstrate how the methodology allows for the creation of maps that simultaneously display regions of high SRKW intensity for each month, along with their corresponding uncertainties. We display these for the month of May for exhibition. Identifying critical habitat plays a major role in the protection plans for any endangered species, thus we hope our work can assist with future policy decisions

surrounding the protection and management of the SRKW population.

5.3 Building the modelling framework

Defining observer-animal encounters

We define the movement trajectory through time of an individual animal m as $\xi_m(t)$. This denotes the spatial coordinate of the individual at time t with respect to the coordinate reference system used. We assume that there exists sufficiently large time windows $T_l \subset \mathbb{R} : l \in L$ such that the movement process driving the trajectories of the individuals $\xi_m(t) : t \in T_l$ have stationary invariant densities for each T_l, denoted $\pi_m(\mathbf{s}, T_l)$. These densities define the UDs we aim to estimate and are assumed a-priori to be arbitrarily complex and represent the long-run density of locations at which the individual visits during T_l. An example time window could be a calendar month.

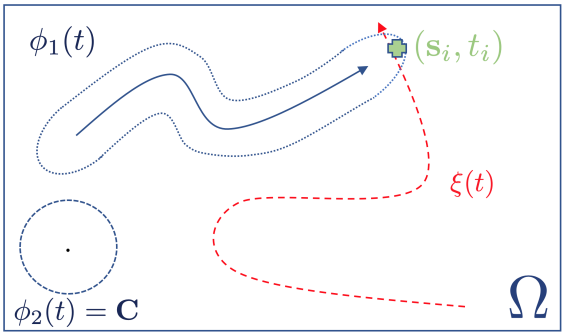

Figure 5.2: A diagram showing an example of an 'encounter'. Two observers, one mobile and one static, with circular space-time fields-of-view $\phi_1(t)$ and $\phi_2(t)$ respectively are plotted with their fields-of-view through time shown as blue dotted lines. Both observers search for an individual moving with space-time trajectory $\xi(t)$ throughout the study region Ω shown as a red dashed line. The arrows denote the direction of travel. At time $t^\star$, the individual crosses into the mobile observers field-of-view at location $\mathbf{s}^\star$. Thus, at time $t^\star$ the individual is encountered at location $\mathbf{s}^\star$. Formally, at time $t^\star$, $\xi(t^\star) \in \phi_1(t^\star) \subset \Omega$.

Observers fall into two categories: static observers (e.g. hydrophones and camera traps) and mobile observers that may move through continuous-space through time (e.g. vehicle-based observers). For each observer $o \in O$, we denote their position and field-of-view at time t as $\xi_o(t)$ and $\phi_o(t)$ respectively. Unlike with $\xi_o(t)$, which defines a unique point in space at each time t, $\phi_o(t)$ defines a unique region or line. Thus $\phi_o(t) \subset \Omega \; \forall t$ is a subset of the study region Ω. Both $\xi_o(t)$ and $\phi_o(t)$ are assumed to arise from arbitrarily complex processes. We discuss these processes and our proposed approximation method (Definition 5.1) later. Note that the fields-of-view from two observers $o, \tilde{o} \in O$ overlap in T_l if $\phi_o(t) \bigcap \phi_{\tilde{o}}(t) \neq \emptyset$ for some $t \in T_l$. We now describe the assumed data generating mechanism that generate the realisation of encounter locations and times used for inference.

Under the assumption of perfect detectability, and assuming the observers are searching continuously throughout T_l, we say that an encounter of individual m occurs during time window T_l when m's movement trajectory $\xi_m(t)$ intersects with one or more observer's field-of-view function $\phi_o(t)$ for some time $t \in T_l$. That is $\xi_m(t) \in \phi_o(t)$ for some $t \in T_l$ and $o \in O$ during the study. When observers only search during a discrete set of times $t_j \in T_l : j \in \{1, ..., J\}$, we say that a sighting occurs at time t_j if $\xi_m(t_j) \in \phi_o(t_j)$. It is these encounters that we perform inference on. The assumption of perfect detectability by an observer within their field-of-view may be unrealistic and we allow for this assumption to be relaxed later. Fig 5.2 presents an example diagram displaying a setting with one moving individual, with two observers (one static, one mobile) searching for it. The individual intersects observer 1's field-of-view at time $t^\star$ and encountered at $\mathbf{s}^\star$. Thus, $(\mathbf{s}^\star, t^\star)$ becomes available for inference.

Target point process model

We assume that there exists an effort surface, denoted $\lambda_{eff}(\mathbf{s}, T_l)$, such that, by conditioning on $\lambda_{eff}(\mathbf{s}, T_l)$ within an inhomogeneous Poisson process (IPP) model, the UD $\pi_m(\mathbf{s}, T_l)$ can be recovered with any observer bias removed. Thus, $\lambda_{eff}(\mathbf{s}, T_l)$ is assumed to fully control for the bias caused

by the heterogeneous observer effort that is driven by the complex processes that generate the observers' fields-of-views and efforts. We use the following inhomogeneous Poisson process (IPP) model to describe the true data generating mechanism. We assume that the number of encounters of m within any subregion $A \subset \Omega$ and time window $T_l \subset \mathcal{T}$ is Poisson-distributed with mean $\Lambda_{obs}(A, T_l) = \int_A \lambda_{obs}(\mathbf{s}, T_l) \mathrm{d}\mathbf{s} = \int_A \lambda_{true}(\mathbf{s}, T_l) \lambda_{eff}(\mathbf{s}, T_l) \mathrm{d}\mathbf{s}$. We refer to $\lambda_{true}(\mathbf{s}, T_l)$ as individual m's true intensity surface. With $\lambda_{eff}(\mathbf{s}, T_l)$ able to fully capture the effort, $\pi_m(\mathbf{s}, T_l) \propto \lambda_{true}(\mathbf{s}, T_l)$. At location $\mathbf{s}$ during time window T_l, it is defined as:

$$\lambda_{true}(\mathbf{s}, T_l) = \lim_{\epsilon \to 0} E\left[N(B_\epsilon(\mathbf{s}), T_l)\right] / \left|B_\epsilon(\mathbf{s})\right|,$$

where $B_\epsilon(\mathbf{s})$ denotes a circle with centre $\mathbf{s}$, radius ϵ, and area $|B_\epsilon(\mathbf{s})|$. $N(B_\epsilon(\mathbf{s}), T_l)$ denotes the number of encounters with m per unit effort within the circle during time window T_l. Given the above assumptions, $\lambda_{true}(\mathbf{s}, T_l)$ loosely represents the expected number of encounters with individual m's trajectory within an infinitesimally small region around point $\mathbf{s}$, per unit effort. A flat $\lambda_{true}(\mathbf{s}, T_l)$ throughout Ω implies that m exhibits complete spatial randomness throughout Ω. Conversely, regions of high $\lambda_{true}(\mathbf{s}, T_l)$ indicate 'hotspots' of m.

Conditioned upon knowing $\lambda_{obs}(\mathbf{s}, T_l)$ and observing a point-pattern consisting of n_{T_l} points (i.e. a collection of n_{T_l} encounter locations) within the time set $T_l \in \mathcal{T}$, $\mathbf{Y}_{T_l} = \{\mathbf{s}_i : i \in \{1, ..., n_{T_l}\}, \mathbf{s}_i \in \Omega\}$, the likelihood of the spatio-temporal IPP is [Simpson et al., 2016]:

$$\pi\left(\mathbf{Y}_{T_l} | \lambda_{obs}\right) = \exp\left\{ |\Omega| \int_\Omega \lambda_{obs}(\mathbf{s}, T_l) \mathrm{d}\mathbf{s} \right\} \prod_{\mathbf{s}_i \in \mathbf{Y}_{T_l}} \lambda_{obs}(\mathbf{s}_i, T_l) \tag{5.1}$$

where $|\Omega|$ denotes the area of the domain Ω.

In practice, we will not know the effort intensity surface $\lambda_{eff}(\mathbf{s}, T_l)$ and we will need to estimate it. When encounters are made in continuous-time, and when overlap exists between the observers' fields-of-view, this quantity may be very complex. In this setting, accurately modelling this term would require the explicit modelling of the animal movement model within

the observer's sampling process [Glennie et al., 2020]. In this Chapter, we show that even crude approximations of $\lambda_{eff}(\mathbf{s}, T_l)$ can improve the statistical inference of $\pi_m(\mathbf{s}, T_l)$. The crude approximation we use is based on the estimated path integrals of the observers' fields-of-view $\phi_o(t)$. When the locations of the observers are known (e.g. GPS positions), this can be computed by estimating $\phi_o(t)$ around each recorded location and then summing through time. When GPS positions are unavailable, the path integral can still prove a useful target quantity for building an effort emulator, or an effort model from a set of informative covariates.

Let $|\cdot|$ denote the area or length function. Assuming no overlap exists between observers, we define the path integral approximation to $\lambda_{eff}(\mathbf{s}, T_l)$ as follows:

Definition 5.1 *the path integral approximation to* $\lambda_{eff}(\boldsymbol{s}, T_l)$ *within a region* $A_i \subset \Omega$ *is* $\int_{T_l} \sum_{o \in O} |\phi_o(t) \bigcap A_i| dt$.

When the entirety of A_i is observed throughout T_l by an observer o, the estimated cumulative effort in A_i from o becomes $|A||T_l|$. The degree of approximation error will depend on many factors including the size of the fields-of-view relative to the size of Ω, the number of observers, the accuracy in estimates of $\phi_o(t)$, and whether or not encounters were made in continuous-time or discrete-time. Modelling $\lambda_{eff}(\mathbf{s}, T_l)$ in discrete-time is easier.

Log-Gaussian Cox processes as a suitable base model for estimating UDs

Linking an animal's space use to a set of covariates (e.g. sea surface temperature) is often a key component of an ecological analysis. The estimated relationships between encounter rate and environment allow researchers to predict variation in space use across space and time, possibly extrapolating into areas beyond the study area Ω, and into time windows beyond the temporal domain $\mathcal{T}$. As with many popular regression-based SDM methods (linear models, GLMs, GAMs, etc.), $\lambda_{true}(\mathbf{s}, T_l)$ may be modeled with a collection of nonlinear transformations of covariates, interactions, and splines

within a log linear model:

$$\log \lambda_{true}(\mathbf{s}, T_l) = \boldsymbol{\beta}^T \mathbf{x}(\mathbf{s}, T_l), \tag{5.2}$$

where $\mathbf{x}(\mathbf{s}, T_l)$ denote the set of measured covariates at location $\mathbf{s} \in \Omega$ assumed constant throughout time window $T_l \subset \mathcal{T}$. This IPP may be inadequate for use in ecological settings [Pacifici et al., 2017]. The Poisson distribution assumed on the counts inside any subregion $A \subset \Omega$, implies the variance of the counts is equal to the mean. If the amount of environmental variability not captured by the modeled covariates is high, and the overdispersion is not controlled for, model-based confidence intervals can become overly-narrow and suffer from poor frequentist coverage [Baddeley et al., 2015]. Spurious 'significance' between the associations of covariates and the intensity may then be reported, unless computationally-intensive resampling methods, such as block-bootstrap, are performed [Fithian et al., 2015].

Cox process models extend the IPP by treating the intensity surface as a realisation of a random field [Baddeley et al., 2015]. This enables the variance-mean relationships of the point process models to be more flexible. The random fields from the Cox process models can be specified to capture spatial, temporal, and/or spatio-temporal correlations, helping to control for any unmeasured covariates and biological processes driving the true intensities of the studied individuals [Yuan et al., 2017].

A popular class of flexible random fields chosen are Gaussian (Markov) random fields, letting $\lambda_{true}(\mathbf{s}, T_l)$ be a realisation of a log-Gaussian process [Simpson et al., 2016]. These models are called log-Gaussian Cox processes (LGCPs). Specification of a LGCP is achieved by adding a Gaussian process, denoted as $Z(\mathbf{s}, T_l)$, to the linear predictor in (2):

$$\log \lambda_{true}(\mathbf{s}, T_l) = \boldsymbol{\beta}^T \mathbf{x}(\mathbf{s}, T_l) + Z(\mathbf{s}, T_l), \tag{5.3}$$

$$\mathbf{Z}(\mathbf{S}, \mathbf{T}) = \left[Z(\mathbf{s}_1, T_{l(1)}), ..., Z(\mathbf{s}_n, T_{l(n)}) \right]^T \backsim N(\mathbf{0}, \Sigma), \tag{5.4}$$

where Σ denotes the variance-covariance matrix of the Gaussian process

evaluated at all of the $n = \sum_{l \in L} n_l$ locations and time windows $(\mathbf{s}_i, T_{l(i)})$. The function $l(i)$ maps each observation to its corresponding time window. Different choices of covariance structures, lead to Gaussian processes with fundamentally different properties and uses. R packages such as spatstat and inlabru can fit such models [Bachl et al., 2019, Baddeley and Turner, 2014]. We choose the LGCP as the base model for our framework due to its flexibility.

Covariates to account for detection probability and observer effort

In practice, we can rarely satisfy the previous assumption of perfect detectability. Thus, we relax this assumption now and assume the existence of a 'detection probability surface', $p_{det}(\mathbf{s}, T_l)$ that can describe the heterogeneous detectability within the earlier point process model. Additional covariates may be available that can model both the heterogeneous detectability [e.g. visibility indices, distance from the observer, etc., Fithian et al., 2015] and/or the heterogeneous cumulative effort of each observer (e.g. distance from the nearest road). If these covariates are included in their correct functional forms, this regression adjustment approach may help to capture some of the heterogeneity in the observer effort and partially remove the associated biases [Dorazio, 2014]. For applications with strongly informative covariates, such approaches have been shown to significantly improve predictive performance in SDMs [Elith and Leathwick, 2007, Fithian et al., 2015].

We model $p_{det}(\mathbf{s}, T_l)$ with a set of covariates, $\mathbf{w}_1(\mathbf{s}, T_l)$, that are believed to affect only the observer abilities and not influence the true intensity of the individual of interest. These covariates are assumed constant throughout each time window T_l. By definition, the values of $p_{det}(\mathbf{s}, t)$ are constrained to lie between 0 and 1. A value of 1 implies that all instances where the individual's trajectory intersects an observer's field-of-view leads to a recorded encounter. This could reflect a scenario where an easily-detected individual is in the immediate proximity to an observer under perfect weather conditions. A value less than 1 implies that some encounters are missed

or not recorded. Thus, this helps to capture the processes that drive the under-reporting seen in many ecological studies. In the context of point processes, $p_{det}(\mathbf{s},t)$ is known as a thinning function. T_l-average sea state and T_l-average visibility are two example covariates that could be included in $\mathbf{w}_1(\mathbf{s},T_l)$. Distance sampling functions can also be included following the approach of Yuan et al. [2017].

Similarly, we can model $\lambda_{eff}(\mathbf{s},T_l)$ with a set of covariates, $\mathbf{w}_2(\mathbf{s},T_l)$, within a loglinear model. As before, the covariates used to explain observer effort are assumed to not directly impact the true intensity of the animals. This time, these covariates are believed to explain both the spatial distributions of observers and explain any differences in their efficiencies. When differences in observer efficiency exist, the earlier approximation to $\lambda_{eff}(\mathbf{s},T_l)$ seen in Definition 5.1 becomes $\int_{T_l}\sum_{o\in O}\omega_o|\phi_o(t)\bigcap A_i|\mathrm{d}t$, with the relative efficiencies captured in their weights ω_o. This can help with the selection of relevant covariates and with model formulation. We advise modelling the observer efficiencies ω_o within the loglinear model and not the earlier probability surface to avoid an upper bound being placed on the efficiency of an observer. This allows for the existence of observers with higher skill levels than those who collected the data used to fit the model. Distance from road and observer type are two example covariates.

Our joint model for true species' intensity, detection probability and observer effort is:

$$\lambda_{obs}(\mathbf{s},T_l) = \lambda_{true}(\mathbf{s},T_l)p_{det}(\mathbf{s},T_l)\lambda_{eff}(\mathbf{s},T_l) \tag{5.5}$$

$$g^{-1}\left(p_{det}\right)(\mathbf{s},T_l) = \boldsymbol{\gamma}_1^T\mathbf{w}_1(\mathbf{s},T_l) \tag{5.6}$$

$$\log\,\lambda_{eff}(\mathbf{s},T_l) = \boldsymbol{\gamma}_2^T\mathbf{w}_2(\mathbf{s},T_l) \tag{5.7}$$

$$\log\,\lambda_{true}(\mathbf{s},T_l) = \boldsymbol{\beta}^T\mathbf{x}(\mathbf{s},T_l) + Z(\mathbf{s},T_l), \tag{5.8}$$

with g, a suitable link function (e.g. the logistic function), mapping the linear predictor of the detection probability surface to the unit interval. By assumption, $\pi(\mathbf{s},T_l) = \lambda_{true}(\mathbf{s},T_l)/\int_\Omega \lambda_{true}(\mathbf{x},T_l)d\mathbf{x}$.

Suppose the encounters are made by a collection of observers $o \in O$ with their unique observer efficiencies modeled with unique intercepts with respect to a baseline observer type. Then, the $\lambda_{true}(\mathbf{s}, T_l)$ being modeled is interpreted as the expected encounter rate at location $\mathbf{s} \in \Omega$ during T_l of the individual by the chosen baseline observer. The log linear model then ensures that any differences in the observer efficiencies are modeled multiplicatively. Crucially, the interpretation of $\lambda_{true}(\mathbf{s}_1, T_l) = 2\lambda_{true}(\mathbf{s}_2, T_l)$ is that the individual will occupy the area immediately around $\mathbf{s}_1$ twice as often as around $\mathbf{s}_2$ in the long run. Finally, if a series of encounters/detections are made where the entirety of Ω is perfectly observable, then both $p_{det}(\mathbf{s}, T_l)$ and $\lambda_{eff}(\mathbf{s}, T_l)$ should be fixed equal to a constant. Two examples are when an ultra high-resolution satellite image containing the entirety of Ω is taken and where telemetry data is available. In the latter case, care must be taken to properly subset the telemetry data to ensure that any autocorrelations from the individual's movement removed.

Estimation of the non-intercept terms within the detection $\boldsymbol{\gamma}_1$, effort $\boldsymbol{\gamma}_2$ and the environmental $\boldsymbol{\beta}$ parameters is possible, so long as all the corresponding covariates $\mathbf{x}$, $\mathbf{w}_1$, and $\mathbf{w}_2$ are not linearly dependent or interact [Dorazio, 2014]. Non-perfect correlation between the three sets of covariates, whilst making estimation more difficult, does not affect the identifiability of these parameters [Fithian et al., 2015]. This is a desirable property in the context of UDs. The intercept parameters are estimable if either there is independence between $\mathbf{x}$, $\mathbf{w}_1$ and $\mathbf{w}_2$, or if at least one accurate control dataset is included in the joint model [Fithian et al., 2015]. A control could be a survey with known observer effort. When the intercept is desired, the animal movement process should be considered to reduce bias [Glennie et al., 2020]. Furthermore, the estimability of observer-specific intercepts (i.e. relative observer efficiencies) will require significant spatial-overlap to exist between the cumulative observer efforts of the different observers. This is especially true when spatially correlated $Z(\mathbf{s}, T_l)$ terms are included, since any differences in observer efficiencies may be erroneously captured by the $Z(\mathbf{s}, T_l)$ terms. Of course, the assumption of non-overlapping fields-of-view at all times $t \in T_l$ is still required. If the above conditions hold, then af-

ter model-fitting, fixing both sets of covariates $\mathbf{w}_1(\mathbf{s},T_l), \mathbf{w}_2(\mathbf{s},T_l)$ equal to a constant allows the effects of variable detection probability and observer effort to be removed from predictions throughout Ω.

Approximating effort from GPS-tagged observers

In many situations, $\int_{T_l} |\phi_o(t) \bigcap A_i| \mathrm{d}t$ may be known or directly estimable for a set of pixels $A_i \subset \Omega$ used to approximate (5.1). For example, mobile observers may record their GPS coordinates, or may be known to travel along a strict network of known routes (e.g. shipping lanes). Similarly, the locations of static observers (e.g. camera traps) and their detection ranges may also be known. In both cases, a simple, yet effective approach for approximating $\lambda_{eff}(\mathbf{s},T_l)$ can be implemented. Given $\phi_o(t)$, we can include $|A_i|^{-1} \int_{T_l} |\phi_o(t) \bigcap A_i| \mathrm{d}t$ as a fixed covariate within $\mathbf{w}_2$ to represent the average effort within A_i. Then, only the corresponding slope term ω_o in $\boldsymbol{\gamma}_2^T$ needs to be estimated. When only one observer type is available, including the logarithm of $|A_i|^{-1} \int_{T_l} |\phi(t) \bigcap A_i| \mathrm{d}t$ as an offset (i.e. fixing $\omega \equiv 1$) is all that is required.

The performance of the above approach may deteriorate as the degree of overlap between the fields-of-view of the observers increases. One solution is to remove the effort and encounters from overlapping observers. Alternatively, it may be possible to model the inter-observer autocorrelations directly [Clare et al., 2017]. We demonstrate the utility of the path integral approximation in a simulation study in Section 4.

Generalising the base model with the addition of marks

Often, the utilization distributions of multiple individuals of a species or population and their changes through time are desired [Elith and Leathwick, 2007, Fithian et al., 2015]. Furthermore, understanding the factors driving the individuals to use the space may also be of importance to researchers. Spatio-temporal point processes can be further generalized to marked spatio-temporal point processes to allow for a greater range of research questions to be tackled [e.g. Chakraborty et al., 2011]. The main idea of marked

point processes is that for each point, we observe attributes in addition to its location and time. These attributes are called marks. Marks might be categorical variables such as whale pod or an indicator of foraging behaviour, count variables such as group size, or continuous variables such as travel speed.

Formally, we associate a random variable m_y (a mark) to each location and time window of the random set $\mathbf{y} \in Y_{T_l}$. We place a probability distribution on each of the marks, and model the joint distribution of the locations and marks. Let M denote the support of a distribution of marks m_y. The mark distribution is allowed to depend upon space and time (i.e. depend upon $\mathbf{y}$ and T_l), but is not allowed to depend on other points in Y_{T_l}. Thus, the m_y for different $\mathbf{y} \in Y_{T_l}$ are independent. Now the pair (Y_{T_l}, m_Y) may be viewed as a random variable $Y^{\star}_{T_l}$ in the product space $\Omega \times M$. There is no limit to the number of marks that can be associated with each point. We simply need to include a joint probability distribution for the J marks $m^j_{Y_{T_l}} : j \in \{1, ..., J\}$.

When the mark distribution of one of the J marks is discrete (i.e. when $m^j_{Y_{T_l}} \in \{1, ..., K\}$), as in the case of individual ID, we can estimate the probability that the presence of an individual at a given location $\mathbf{s} \in \Omega$ within $T_l \subset \mathcal{T}$ has j^{th} mark equal to $k \in \{1, ..., K\}$. For notational simplicity, let $J = 1$ and define $\lambda_{true}(\mathbf{s}, T_l, k)$ to be the true intensity for the mark category k during T_l (i.e. $\{\lambda_{true}(\mathbf{s}, T_l, m) : m_{\mathbf{s},T_l} = k\}$). The probability at location $\mathbf{s}$ is then:

$$\begin{aligned} p_{true}(\mathbf{s}, T_l, k) &= \frac{\lambda_{true}(\mathbf{s}, T_l, k)}{\sum_{\kappa=1}^{K} \lambda_{true}(\mathbf{s}, T_l, \kappa)} \\ &= \frac{\exp\Big(Z(\mathbf{s}, T_l, k) + \boldsymbol{\beta}^T \mathbf{x}(\mathbf{s}, T_l, k)\Big)}{\sum_{\kappa=1}^{K} \exp\Big(Z(\mathbf{s}, T_l, \kappa) + \boldsymbol{\beta}^T \mathbf{x}(\mathbf{s}, T_l, \kappa)\Big)} \end{aligned} \tag{5.9}$$

In many cases, computing and plotting estimates of $p_{true}(\mathbf{s}, T_l, k)$, the true mark-specific probabilities will be the inferential target. When k denotes the individual ID, estimates of $p_{true}(\mathbf{s}, T_l, k)$ can help to establish spa-

tial niches specific to the k^{th} individual, whilst removing any observer and detectability biases. Note that when K denotes the number of individuals of a target population and when Ω is sufficiently large, then the normalized denominator of (5.9) reflects the distribution of the whole population. When some or all of the parameters and/or covariates are shared between the mark-specific intensities, cancellations will occur in (5.9).

The proposed model framework

Let Ω, $\mathcal{T}$, and $T_l \subset \mathcal{T} : l \in L$ be defined as before. Let the support of the marks be M. Suppose for each T_l we have a collection of encounter locations and marks $(\mathbf{Y}_{T_l}, \mathbf{M}_{Y_{T_l}}) = \{(\mathbf{s}_i, \mathbf{m}_i) : (\mathbf{s}_i, \mathbf{m}_i) \in \Omega \times M\}$. The proposed model is:

$$\lambda_{obs}(\mathbf{s}, T_l, \mathbf{m}) = \lambda_{true}(\mathbf{s}, T_l, \mathbf{m}) p_{det}(\mathbf{s}, T_l, \mathbf{m}) \lambda_{eff}(\mathbf{s}, T_l, \mathbf{m}) \tag{5.10}$$

$$g^{-1}\left(p_{det}(\mathbf{s}, T_l, \mathbf{m})\right) = \boldsymbol{\gamma}_1^T \mathbf{w}_1(\mathbf{s}, T_l, \mathbf{m}) \tag{5.11}$$

$$\log\, \lambda_{eff}(\mathbf{s}, T_l, \mathbf{m}) = \boldsymbol{\gamma}_2^T \mathbf{w}_2(\mathbf{s}, T_l, \mathbf{m}) \tag{5.12}$$

$$\log\, \lambda_{true}(\mathbf{s}, T_l, \mathbf{m}) = \boldsymbol{\beta}^T \mathbf{x}(\mathbf{s}, T_l, \mathbf{m}) + Z(\mathbf{s}, T_l, \mathbf{m}). \tag{5.13}$$

For estimates of $\lambda_{true}(\mathbf{s}, T_l, \mathbf{m})$ under the above framework to be free from confounding by effort and for estimates of individual-environment relationships to be accurate, many assumptions are required in addition to those highlighted earlier. As shown in the causal directed acyclic graph (DAG) in Fig 5.3 [Hernan and Robins, 2020], one of the fundamental assumptions required for estimates of $\lambda_{true}(\mathbf{s}, T_l)$ to be free of confounding, is that $\lambda_{eff}(\mathbf{s}, T_l, \mathbf{m})$ fully describes the efforts of the observers through space and time across the marks. This is achieved when either the covariates $\mathbf{w}_2(\mathbf{s}, T_l, \mathbf{m})$ completely explain the efforts of the observers, or when known or accurate estimates of observer effort are included in $\mathbf{w}_2(\mathbf{s}, T_l, \mathbf{m})$. For estimates of the UD to be free of confounding, only the relative efforts of the observers need be known or estimable. In either case, with unknown effort, the existence of unobserved covariates of effort can confound estimates of

$\lambda_{true}(\mathbf{s}, T_l, \mathbf{m})$ in two ways.

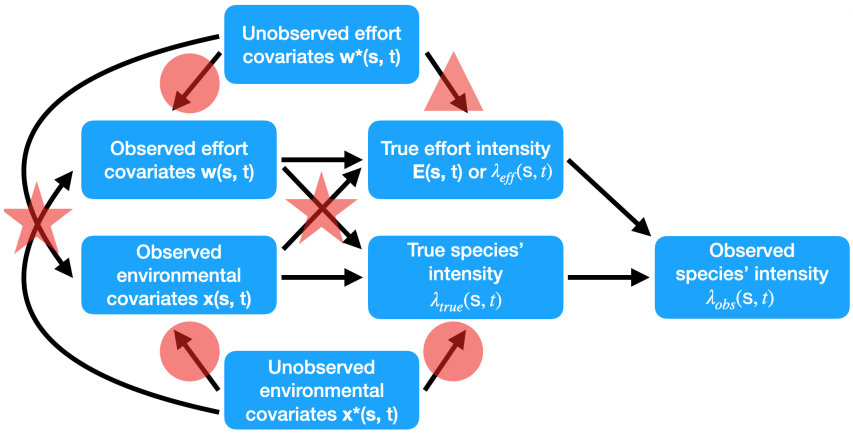

Figure 5.3: A plot showing the assumed causal DAG for the proposed framework with the detection probability assumed constant. An arrow between a variable set A and a variable set B indicates that at least one variable exists in both sets with a direct causal effect between them. The causal Markov assumption is made such that a variable is independent of its non-descendants, when conditioned on its parents [Hernan and Robins, 2020]. If any of the causal effects found within the red shapes exist, then problematic confounding may follow. This is explained at depth in the supporting material found in the Appendix.

Firstly, if the unobserved effort covariates affect one or more observed environmental covariates $\mathbf{x}(\mathbf{s}, T_l, \mathbf{m})$, then the corresponding effect estimate $\boldsymbol{\beta}$ and hence the individual's intensity $\lambda_{true}(\mathbf{s}, T_l, \mathbf{m})$ may remain confounded by the effort. For example, suppose effort is unknown for a dataset containing encounters with marine-based individuals. Suppose that distance-to-shore is not included as a covariate despite strongly impacting search effort. Furthermore, suppose chlorophyll-A, which has high values closer to shore, is included in $\mathbf{x}(\mathbf{s}, T_l, \mathbf{m})$. Estimates of the effects of chlorophyll-A in $\boldsymbol{\beta}^T$ will likely be confounded by effort, which in turn will bias the estimates of $\lambda_{true}(\mathbf{s}, T_l, \mathbf{m})$.

Secondly, even if the unobserved effort covariates are independent of

$\mathbf{x}(\mathbf{s}, T_l, \mathbf{m})$, residual spatio-temporal correlations in the sightings data driven by the unobserved effort covariates may be erroneously captured by the Gaussian process $Z(\mathbf{s}, T_l, \mathbf{m})$. Consequently, estimates of $\lambda_{true}(\mathbf{s}, T_l, \mathbf{m})$ may therefore remain confounded by the heterogeneity in the observer effort. At this point, one may be tempted to simply include a unique Gaussian process for λ_{eff} to capture these missed covariates. This cannot be done. Without additional knowledge available that can adequately constrain the additional Gaussian process, it will be non-identifiable. Similar problems occur if a detection probability surface $p_{det}(\mathbf{s}, T_l, \mathbf{m})$ is estimated. Once again, the true detectability of the species must be fully captured by $\mathbf{w}_1(\mathbf{s}, T_l, \mathbf{m})$.

Removing the confounding of $\lambda_{true}(\mathbf{s}, T_l, \mathbf{m})$ by effort will be challenging for many ecological applications and it highlights the need for increased collection of effort information along with sightings data. However, Fig 5.3 shows that if effort is known, then conditioning on it can remove all the problematic confounding at location $\mathbf{s}$. Furthermore, in the simulation study in Section 4, we show that even crude estimates of observer effort and detectability can dramatically improve estimates of $\lambda_{true}(\mathbf{s}, T_l, \mathbf{m})$ compared with simply ignoring effort altogether. The issues of confounding are not exclusive to our framework and are present across all methods for estimating both UDs and SDMs [Fithian et al., 2015, Koshkina et al., 2017]. Thus, these concerns should not be seen as a weakness of our framework, but instead as a weakness inherent to biased data collection-protocols.

If observer effort is limited to a small subregion $\Omega_0 \subset \Omega$, such that $\lambda_{eff}(\mathbf{s}, T_l, \mathbf{m}) = 0 \; \forall \mathbf{s} \in \Omega \cap \Omega_0^C$, then additional assumptions must be placed on how Ω_0 was selected. For example, if search effort was focused in regions where the species' intensity was expected to be highest, then extrapolated estimates of $\lambda_{true}(\mathbf{s}, T_l, \mathbf{m})$ into the regions of zero effort $\Omega \cap \Omega_0^C$ may remain biased. Estimates of the true intensity $\lambda_{true}(\mathbf{s}, T_l, \mathbf{m})$ into these regions may be too high, with estimates of the intercept positively biased. This issue is known as preferential sampling [Pennino et al., 2019]. This highlights the benefits of conducting high-quality surveys with randomized effort. By choosing Ω_0 at random, no systematic bias is expected in predictions into $\Omega \cap \Omega_0^C$, and extrapolation of the intensity throughout Ω can be performed

with greater confidence. When there is little-to-no confidence that Ω_0 was selected in a random manner, predictions should be constrained to lie within Ω_0.

An important advantage of the framework is that data from different observers and of differing type can be combined to jointly estimate one intensity surface [Koshkina et al., 2017]. This is achievable as the intensity given by (5.10), can be linked to the likelihoods of several common data types, including aggregated forms. For example, the logistic regression likelihood for binary site presence-absence data, the Poisson likelihood for site count data and the LGCP likelihood for presence-only data can all be derived from the intensity (see Hefley and Hooten [2016] and the supporting material found in the Appendix). Distance sampling methods have also been fit using LGCPs (see Yuan et al. [2017] for details). All that is required for the suitable combination is that encounters with individuals are approximately independent snapshots of their UDs and that any heterogeneous effort or detectability can be suitably controlled for. This is a clear demonstration of the unifying potential of the LGCP for ecological data, and the causal DAG of Fig 5.3 is a useful tool to assess whether the assumptions needed to apply model (5.10) are satisfied. A short summary of how to approximate the LGCP likelihood and additional details on the DAG are provided in the Appendix.

5.4 Simulation study

We now present a simulation study to demonstrate the ability of the framework to combine encounter data from mobile and static observers and predict an animal's UD with both minimal bias and high precision. We first simulate the movements of observers and an animal and generate encounters following the earlier data generating mechanism. Next, we use the recorded locations of the observers to compute the approximation to effort that was introduced earlier in Definition 5.1. We ignore issues of overlap. We then plug this estimate of effort into a point process model and attempt to recover the UD from the generated encounters. Note that we do not attempt to explicitly

model the movements of the animal within the sampling process model as in [Glennie et al., 2020]. Despite the use of a crude approximation for effort, we show that the framework offers improvements in prediction performance, even when the analyst incorrectly specifies $\phi_o(t)$ throughout time. Based on these results, we provide a list of recommendations for analysts.

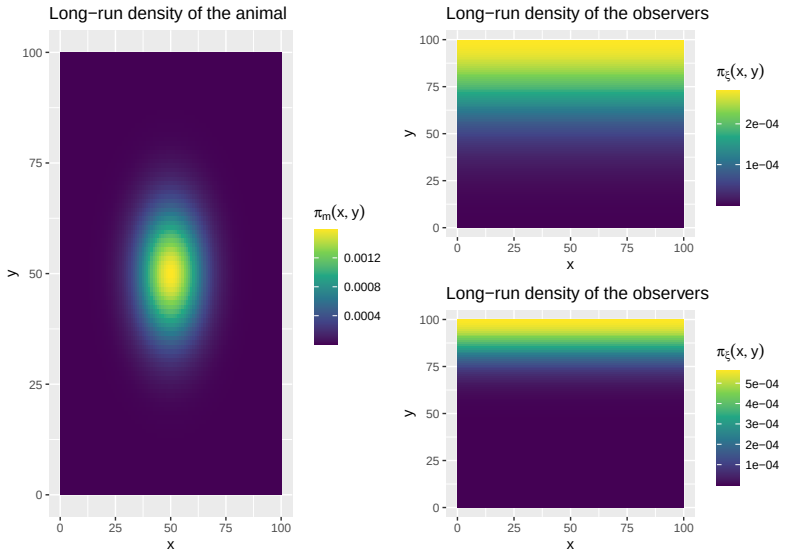

Figure 5.4: A plot showing the long run densities of the animal and the observers. The top-right and bottom-right plots show the low and high observer bias settings respectively. The smaller Brownian motion variance leads to a higher concentration of effort in the North of the study region and hence a larger degree of observer bias.

We simulate the movements of an animal and a set of observers using the model of Brillinger et al. [2012]. In particular, we use a stochastic differential equation (SDE) with potential functions chosen to ensure a desired long-run behaviour. The animal's potential function is chosen to be the logarithm of a symmetric bivariate normal distribution centered at $(\mu_x, \mu_y) = (50, 50)$, with variance 100. The observers' potential function is specified as the logarithm of a univariate half-normal distribution centered at $m_y = 100$, with variance 200. The variance of the Brownian motion terms driving the movements

is fixed at 2 for the animal, and fixed at either 2 or 8 for the observers. Thus, the animal's UD is a symmetric bivariate normal distribution, with the observers' UD a univariate normal distribution which focuses their efforts in the North of the study region (Fig 5.4). The study region is a square with side lengths equal to 100 arbitrary units. The observers are given circular fields-of-view with maximum range of 10 units. These settings imply that the study region is very small.

We discretize time and use a first-order approximation to generate paths from the continuous-time SDE. Thus, both the simulated movements and potential encounter events occur across the discrete time-steps. The average distance travelled at each time step is roughly 1.75 units for the animal, and either 1.75 or 3.5 units for the mobile observers depending on whether the variance of the Brownian motion is 2 or 8. At each time step, if the animal is closer than 10 units of distance from an observer, it is encountered with a probability that decays linearly from 1 to 0 as the distance from the observer increases from 0 units to 10 units. If the animal is detected within 500 time-steps, representing a single 'trip', then the encounter location is recorded along with the observers' tracks. If no encounter occurs during the trip, then only the observers' tracks are recorded. For each trip, we randomly sample the initial locations of the animal and the observers from their respective UDs. Subsequent locations are restricted from leaving the study region. For static observers, we simply hold their initial values fixed through time. The fields-of-view of all observers may overlap.

For each simulation iteration, we then repeat the above steps 150 or 300 times to generate 150 or 300 trips. We fit a (IPP) point process model with correctly specified parametric form to the encounter locations with and without effort adjustment. We compute a crude approximation of effort as follows. The observers' paths are mapped to a coarse 100×100 grid of pixels. Next, estimates of their fields-of-view, $\hat{\phi}_o(t)$, are computed, and then path integrals of $\hat{\phi}_o(t)$ are taken over the grid. Log-values of these path integral approximations are then included as an offset within the IPP. Here, estimates of effort are summed across the observers, ignoring any overlap in their fields-of-view. We also present results from a method that controls for

overlapping fields-of-view in the supporting material found in the Appendix, but only minor improvements in predictive performance are seen. For model comparison, we compute at each simulation iteration both the mean squared prediction error (MSPE) of the animal's UD across the grid of pixels, and the bias of the estimated y-axis center of the UDs $\hat{\mu}_y$. The MSPE is computed with respect to the true UD.

To understand how the method performs in practice, we change both the data-generating mechanism (DGM) and the assumptions made by the analyst when formulating estimates of observer effort. For the DGM, we adjust the degree of observer bias from 'high' to 'low' by changing the variance of the Brownian motion driving the observers' motions from 2 to 8 respectively. We also change the number and type of observers (mobile and/or static) and the number of trips made (150 or 300). For estimating effort, we either assume perfect detectability across $\hat{\phi}_o(t)$, or we model the linearly decaying distance sampling function. Next, we either underestimate, correctly specify, or overestimate the detection range of $\hat{\phi}_o(t)$ at 2, 10, and 50 units respectively. We perform 100 replications of each setting.

5.4.1 Effects of observer effort and detection range misspecification

Fig 5.5 and Fig A.23 in the supporting material found in the Appendix demonstrate that improvements in both prediction variance and bias can be attained with the approximate effort-correction approach. These benefits are seen across all the observer types (i.e. static, mobile, and combinations) and the typical performance of the bias-corrected method is seemingly insensitive to the degree of the observer bias. In contrast, the performance of the uncorrected model is greatly affected by both the level of observer bias and observer type. In particular, the uncorrected model performs poorly when one ignores large observer bias. The results from both the 150 and 300 trip settings are similar and so we aggregate the results when forming the plots.

The performance of the bias-correction approach is sensitive to the analyst using the correct detection ranges of the observers to define the observers' fields-of-view $\phi_o(t)$ (Figs A.24 and A.25). However, improvements

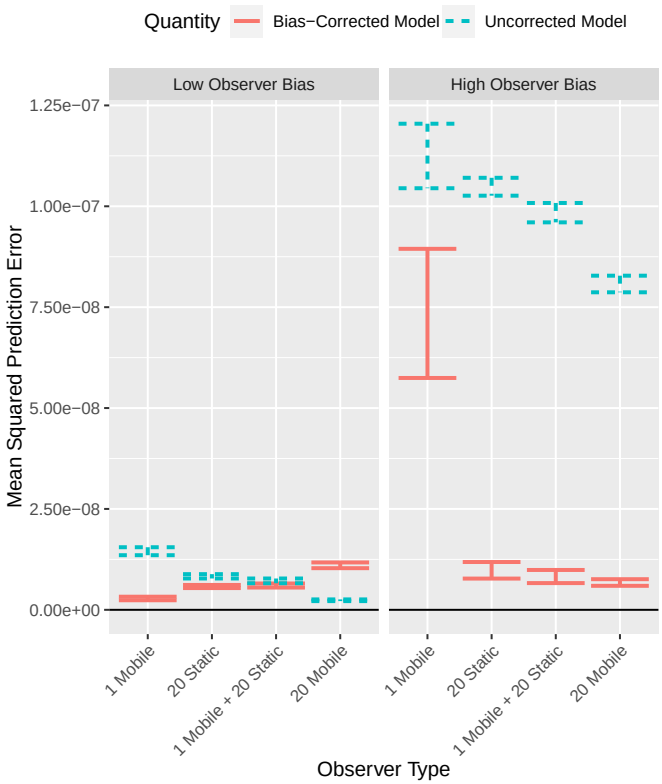

Figure 5.5: A plot showing the mean squared prediction error (MSPE) of the animals UD under the bias-corrected and bias-uncorrected models vs the types of observers. From left to right are the results from one mobile observer, twenty static observers, twenty static with one mobile observer, and twenty mobile observers. The degree of observer bias is changed from low to high in the columns. The red solid lines and the blue dashed lines show the median MSPE along with robust intervals computed as $\pm 2c$MAD from the Bias-corrected and uncorrected models across the 100 simulation replicates respectively. The median absolute deviations about the medians (MAD) have been scaled by $c = 1.48$. This ensures that the intervals are asymptotically equivalent to the 95% confidence intervals that would be computed if the MSPE values were normally distributed. Note that here all the analyst's assumptions correctly match the true data-generating mechanism, albeit with any overlap in the observers' efforts ignored.

in both the prediction variance and the bias of UD center-estimates can still be seen, even with badly misspecified detection ranges. Underestimating the observers' detection range leads to an over-correction of observer bias and leads to estimates of the animal's UD center to be negatively biased. The converse is true when the observers' ranges are overestimated. Both lead to increases in MSPE. Interestingly, the bias-correction method appears insensitive to whether or not a distance sampling function is used.

The MSPE is a measure of predictive performance that is driven by both the squared prediction biases and the variances of the predictions. For the uncorrected model, the heterogeneous observer effort is the major cause of prediction bias. This is expected to decrease as the number of observers increases, due to the study region becoming increasingly explored. For the effort-corrected model, two major causes of prediction bias remain in our crude approach for approximating effort. The first such cause is due to the approximation error of using the path integrals of the fields-of-view to represent the cumulative effort. Even if $\phi_o(t)$ is correctly specified at all times, approximation error will remain due to the statistical dependence between the encounter/non-encounter events through time. To understand this, suppose an observer has failed to record an encounter for a significant period of time. Conditional upon this information, the current location of the animal is unlikely to be situated within the immediate proximity of the observer. Accurate estimates of the true effort $\lambda_{eff}(\mathbf{s}, t)$ would need to be adjusted to account for this fact. This would require explicitly modelling the animal's movement process jointly within the sampling model [Glennie et al., 2020]. The second cause of prediction bias is due to overlap in the observers' fields-of-view. This gets worse as the density of observers increases. Multiple factors impact the variance of the predictions. For both models, the variance of the predictions decreases as the number of encounters increases. For the effort-corrected model, longer observer paths further reduce the variance. Both encounter frequency and the cumulative observer path length increase with the number of observers.

For the uncorrected model, we indeed see that the mean squared prediction error (MSPE) decreases as the number of observers is increased. The

largest improvements are seen with the addition of mobile observers due to the study region being increasingly explored. With twenty mobile observers in the low bias setting, the impact of the observer bias is negligible on the prediction performance of the uncorrected model and it outperforms the effort-corrected model (Fig 5.5). Conversely, in the low observer bias setting, the MSPE from the effort-corrected approach is found to increase with the number of observers. Here, increases in the squared prediction bias dominate any possible reductions in the variance of predictions. In the high bias setting, the reverse relationship is seen. Here, reductions in prediction variance offered by the increased number of observers offsets any increases in the squared prediction bias. Fig A.23 shows that in this setting, the variability in the estimates of the UD center decreases substantially as the number of observers increases.

We demonstrate our claims made above in an additional simulation study explained in depth in the supporting material found in the Appendix. In it, we change two simulation settings. First, we increase the speed of the movements across each time step to reduce the autocorrelation between the encounters. This moves the simulation from the pseudo continuous-time encounter setting to a more discrete-time setting. Second, we fit an additional effort-corrected model that directly accounts for the overlap between the observers' fields-of-view. This 'overlap-corrected model' is found to completely eliminate the estimation bias of the UD center (Fig SA.28). Interestingly however, no change in the MSPE is witnessed relative to the previous effort-corrected model (Fig SA.29). Thus it appears that the bias reduction from the overlap-correction approach comes at a cost of an increased variance of the UD predictions. Both effort-corrected models outperform the uncorrected model with respect to MSPE across all levels of observer bias.

In summary, it appears that the benefits of effort-correction can be attained when little is known about the precise nature of the observer effort. As long as the animal's UD remains reasonably constant throughout the trips, crude attempts at effort correction appear to be better than ignoring effort in most settings. Furthermore, the path integrals of observers' fields-of-view appears control for effort reasonably well. When the degree of

observer bias is expected to be high, as is expected in our case study, it appears that this form of effort-correction can lead to dramatic improvements in predictive performance, without the need to consider observer overlap or explicitly model the animal movement.

5.5 Application to empirical data

Special considerations required for our motivating problem

To demonstrate the utility of our modelling framework, we apply it to the southern resident killer whale (SRKW) data. We partition the temporal domain (May - October) $\mathcal{T}$ into months $T_l : l \in \{\text{May}, ..., \text{October}\}$ and assume the intensities and hence the UDs of the pods are constant within each month. We denote the day as $d \in \{1, ..., N_{T_l}\}$ and the year as $y \in \{2009, ..., 2016\}$. We assume that no changes to the UDs occur between 2009-2016. Our motivating dataset contains several special features that require careful consideration.

As mentioned in Section 5.2.2, the pod identities (J, K, or L) of the sightings can be considered known. We denote the pod identity for a sighting with the discrete mark m and we consider the pods as our 'individuals'. Pods are often found swimming together in 'super-pods'. We break up sightings of super-pods into their individual components. For example, if a sighting of super-pod JK is made (i.e. J and K are found together), then we record this as a sighting of pod J and a sighting of pod K and ignore the potential interaction.

The data are heavily autocorrelated. Sightings are often made of the same pod in quick succession, and the locations of whale pod sightings are shared between whale-watch operators. In fact, once a pod has been sighted, it is rarely lost by the tour operators for the remainder of the day. To remove the autocorrelations, we consider only the first sightings per day of each pod, discarding all repeated sightings made within a day. Importantly, for each day and for each pod, we also discard all predicted effort that occurs after the initial sighting. Because whales move quickly relative to $|\Omega|$, an overnight

window between sightings is sufficient to remove the autocorrelation between sightings. Next, we estimate the cumulative monthly observer effort from all observers. The effort is summed across the 8 years.

Incorporating the observer effort from the DFO data

The daily GPS tracklines of the DFO vessel prior to each initial SRKW sighting are used to approximate the DFO's observer effort. The GPS data are irregular, with a typical resolution of around 15 seconds. We predict the locations at regular 30 second intervals using a continuous-time correlated random walk model fit to each trip using the crawl package [Johnson and London, 2018, Johnson et al., 2008]. We denote the approximate locations and effort as $\xi_{DFO}(t, y, d)$ and $E^{obs}_{DFO}(\mathbf{s}, y, T_l, m)$ respectively. Next, we count up the number of predicted points that fall into a set of polygonal regions A_i used to approximate (5.1). Thus, we assume that at each 30 second interval, the observer's field-of-view $\phi_{DFO}(t, y, d)$ is uniform throughout the A_i that contains the vessel. Thus, $\int_{A_i} E^{obs}_{DFO}(\mathbf{s}, y, T_l, m) d\mathbf{s} \approx \sum_d \int_{T_l} \mathbb{I}\{\xi_{DFO}(t, y, d) \in A_i\} dt$. The A_i are approximately circular with radius 2.6km (see Fig A.31). Note that we only have the location of the vessel during encounters. Results from the simulation study suggest that these steps are unlikely to significantly impact the analysis, given the large number of boat tracks available, the large Ω, and given that our assumed maximum detection range of 2.6km for $\phi_{DFO}(t)$ is likely not orders of magnitude form the truth. Effort is scaled into units of hours.

Estimating the whale-watch observer effort

To incorporate the observer effort from the whale-watch vessels, we build a stochastic emulator of the cumulative 'boat-hours' spent in each of the integration points A_i by the whale-watch companies for each day, month, and year under study. We refer to the cumulative pod-specific monthly whale-watch observer effort intensity as $E^{obs}_{WW}(\mathbf{s}, y, T_l, m)$. Because the whale-watch sightings are not linked to a specific vessel, we assume throughout that the observer efficiencies across the whale-watch vessels are constant. We do not

adjust for overlap between the fields-of-view of the vessels. The density of boats within the study region is expected to be far smaller than it was in the simulation study with twenty vessels and the degree of observer bias is very high, suggesting that the results should be accurate. Note that the assumptions made on $\phi_{WW}(t, y, d)$ match those of $\phi_{DFO}(t, y, d)$.

For each day and for each pod, we first record the number of hours into the operational day at which the initial discoveries were made. We denote this τ. We assume that the daily operational period for the whale-watch companies is 9am - 6pm [Seely et al., 2017], thus $\tau \in [0, 9]$. As an example, suppose that on a given day, pods J and K were both sighted at 12pm and pod L was never sighted. Then τ would be recorded as 3 hours for pods J and K and 9 hours for pod L. To account for the changing effort throughout the day, we use the numbers of vessels reported by Soundwatch to be in close proximity with whales by hour of day as our proxy for whale-watch effort intensity. We then use these reported values to estimate a cumulative distribution function $F_E(\tau)$ for the proportion of total whale-watch effort spent τ hours into the day. Thus, for an initial sighting of a pod made τ hours into the day, $F_E(\tau)$ represents the fraction of total whale-watch effort spent prior to that sighting.

Let $(\tau_{m,y,T_l,d})_{d=1}^{N_{T_l}}$ denote the number of hours after 9am when the first sighting of pod m, in year y, in month T_l and on day d occurs. Under the assumption that an overnight window removes the autocorrelation between the SRKW locations, the fraction of total WW observer effort spent prior to the initial sightings of pod m in a given month/year is:

$$\frac{1}{N_{T_l}} \sum_{d=1}^{N_{T_l}} F_E(\tau_{m,y,T_l,d}). \tag{5.14}$$

Next, we need to estimate the maximum possible number of boat hours of observer effort for each year, month, and day. We denote it $E_{WW}(y, T_l, d)$. We will then multiply the year, month sums $E_{WW}(y, T_l) = \sum_{d=1}^{N_{T_l}} E_{WW}(y, T_l, d)$ by the fraction (5.14). The result will be an estimate of the observer effort associated with the initial sightings. This requires some strong assumptions

that are detailed in the supporting material found in the Appendix, including that the average spatial distribution of the whale-watch boat observer effort is constant throughout the day.

Soundwatch reports on: the number of active whale-watch ports per year, the maximum number of trips departing each day from each port, the changing number of daily trips across the months, and the duration (in hours) of the trips from each port. We also download wind-speed data and ask various operators for their operational guidelines on cancellations due to poor weather/sea state. We then remove days considered 'dangerous'. Given the large sources of uncertainties associated with estimating the above quantities, we formulate probability distributions to appropriately express the uncertainties with each of our estimates. These probability distributions form the backbone of our stochastic emulator of $E_{WW}(y, T_l, d)$.

To estimate the spatial distribution of the observer effort, we estimate how many boat hours could fall in each of the integration points A_i per month and year. Estimates of maximum travel ranges from each port are obtained, considering land as a barrier. Typical vessel routes from the whale-watching companies are established through: private communications with the operators, Soundwatch reports, the operators' flyers and websites. Combining these together, we then formulate plausible effort fields from each port by hand using GIS tools.

We denote $E_{WW}(\mathbf{s}, y, T_l)$, the maximum possible observer effort intensity for year y, month T_l and at location $\mathbf{s} \in \Omega$. It is subject to the following constraint:

$$\begin{aligned}\int_{\Omega} E_{WW}(\mathbf{s}, y, T_l)\mathrm{d}\mathbf{s} &= \text{Total possible WW boat hours in year } y, \text{ month } T_l \\ &= E_{WW}(y, T_l).\end{aligned}$$

Our estimate of the pod-specific monthly whale-watch observer effort surface associated with our initial daily sightings is:

$$E^{obs}_{WW}(\mathbf{s}, T_l, m) = \sum_{y=2009}^{2016} E^{obs}_{WW}(\mathbf{s}, y, T_l, m) \tag{5.15}$$

$$E^{obs}_{WW}(\mathbf{s}, y, T_l, m) = E_{WW}(\mathbf{s}, y, T_l) \times \frac{1}{N_{T_l}} \sum_{d=1}^{N_{T_l}} F_E(\tau_{m,y,T_l,d}).$$

Combining effort surfaces

Due to their spatially disjoint observer efforts (Fig 5.1), almost no spatial overlap exists between the two sources of sightings: presence-only sightings (reported by the whale-watch operators) and the presence-absence data (recorded from the DFO boat survey). Consequently, under our LGCP framework, any intercept term added to $\mathbf{w}_1(\mathbf{s}, T_l, m)$ for capturing the relative observer efficiencies between the two observer types will not be estimable due to confounding with the spatial field $Z(\mathbf{s}, T_l, m)$.

Since both observer types involve similarly-sized vessels, we make the assumption that the efficiencies across the two observer types are identical. Thus, we simply sum the two observer effort layers to get the total observer effort:

$$E^{obs}_{Total}(\mathbf{s}, T_l, m) = E^{obs}_{WW}(\mathbf{s}, T_l, m) + E^{obs}_{DFO}(\mathbf{s}, T_l, m) \tag{5.16}$$

Large uncertainties surround our estimates of the whale-watch observer effort, with the coefficient of variation exceeding 0.25 for the estimates from some of the smaller ports. Failing to account for these uncertainties could lead to over-confident inference. We produce G Monte Carlo samples of the effort field $E^{obs}_{WW,g}(\mathbf{s}, T_l, m) : g \in \{1, ..., G\}$. For each sampled observer effort field, we then fit the LGCP model and sample once from the posterior distributions of all the parameters and random effects. These new posterior distributions will help account for the uncertainty in observer effort, so long as G is chosen sufficiently large to reduce the Monte Carlo error. We choose

$G = 1000$.

Model selection

We propose and fit several candidate models of increasing complexity for analysis. We fit the models using the R-INLA package with the SPDE approach [Lindgren et al., 2011b, 2015, R Core Team, 2019, Rue et al., 2009]. All models use the estimated observer effort field E^{obs}_{Total}, with no detectability or observer effort covariates used (i.e. $p_{det} \equiv 1$ and $\lambda_{eff} \equiv E^{obs}_{Total}$). Model candidates start from the simplest complete spatial randomness model. This assumes that conditioned on observer effort, encounter locations for each pod and month arise from a homogeneous Poisson process. They finish with models for $\lambda_{true}(\mathbf{s}, T_l, m)$ which include: covariates, temporal splines, and Gaussian (Markov) random fields with separable spatio-temporal covariance structures within (5.13). To avoid excessive computation time, we perform model selection on a single realisation of our observer effort field. Then, for our 'best' model, we propagate the uncertainties with observer effort through to the results via the Monte Carlo approach.

We explore two space-time covariates and one spatial covariate: sea-surface temperature (SST), chlorophyll-A (chl-A), and depth. Covariates were downloaded from the ERDDAP database [Simons, 2019], with the monthly composite SST and chl-A rasters (Fig A.34) extracted from satellite level 3 images from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor onboard the Aqua satellite (Data set ID's: erdMH1sstdmday and erdMH1chlamday respectively). We compare the two types of hierachical space-time centering seen in Yuan et al. [2017].

We also explore the addition of spatial random fields, spatio-temporal random fields, and temporal splines within (5.13). Including these adds a substantial amount of complexity to the model. To avoid over-fitting the data, we start with the simplest models without random effects, and iteratively increase the complexity of the model in a stepwise manner. To choose the 'best' model, we use the Deviance Information Criterion (DIC). This trades-off the goodness-of-fit of the model with a penalty for the model's

complexity [see Spiegelhalter et al., 2002]. The candidate models that do not contain random fields or splines are equivalent to MAXENT models, a commonly used SDM method [Renner and Warton, 2013]. Thus the DIC values of the models allow for comparisons to be made between the commonly used MAXENT models and our proposed LGCP model.

We also conduct posterior predictive checks on the candidate models [Gelman et al., 1996]. In particular, we assess the ability of the models to accurately estimate the total number of first sightings of each pod, per month. We also assess the models' abilities to suitably capture the spatial trend by comparing the observed number of sightings falling within each region A_i with their model-estimated credible intervals.

The final selected model

The final 'best' model, as judged by DIC and posterior predictive check assessments, includes a spatial random field shared across the three pods, a spatial field unique to pod L, pod-specific temporal effects (as captured by second-order random walk processes), SST, and chl-A. Both covariates were space-time centered. Depth was omitted as its inclusion led to numerical instabilities due to high multicollinearity. Details of all the models are in the supporting material found in the Appendix (see Table A.2).

The importance of including a random field unique to pod L implies that pod L exhibits different space use compared with J and K. This result is in agreement with [Hauser et al., 2007]. No unique spatial field for pod J or K was found to significantly improve the model. The pod-specific random-walks reflect the different times the pods arrive and leave the area of interest. For example pod J is found to remain in the area of interest across the months, whereas pods K and L are found to have lower intensities in May relative to September (Fig A.35). This is in agreement with Ford et al. [1996] 104 pp.

Finally, we use the the causal DAG shown in Fig 5.3 to display our assumptions about the 'best' model. The first assumption is that we can accurately emulate observer effort $\lambda_{eff}(\mathbf{s}, T_l, m)$ with $E^{obs}_{Total}(\mathbf{s}, T_l, m)$ and that

no unmeasured strong predictors of effort $\mathbf{w}^{\star}(\mathbf{s}, T_l, m)$ exist. Large residual spatio-temporal correlations caused by $\mathbf{w}^{\star}(\mathbf{s}, T_l, m)$ would be erroneously captured in the spatial fields for $\lambda_{true}(\mathbf{s}, T_l, m)$, leading to estimates of pod intensity to remain confounded by effort. The next assumption is that no path in the DAG exists between the environmental covariates, measured or unmeasured (i.e. $\mathbf{x}(\mathbf{s}, T_l, m)$ nor $\mathbf{x}^{\star}(\mathbf{s}, T_l, m)$), and the effort $\lambda_{eff}(\mathbf{s}, T_l, m)$. This would also lead to pod intensity estimates to remain confounded by effort. For the estimates of the species-environment effects $\boldsymbol{\beta}$ to be accurate, we need to assume that no unmeasured environmental covariates $\mathbf{x}^{\star}(\mathbf{s}, T_l, m)$ exist. This assumption is unlikely to hold. The choices driving the movements of the SRKW are likely far more complex than explained by the two covariates alone and unmeasured environmental factors $\mathbf{x}^{\star}(\mathbf{s}, T_l, m)$ are likely to interact with both $\mathbf{x}(\mathbf{s}, T_l, m)$ and $\lambda_{true}(\mathbf{s}, T_l, m)$ causing confounding. However, the presence of $\mathbf{x}^{\star}(\mathbf{s}, T_l, m)$ should not impact our ability to predict $\lambda_{true}(\mathbf{s}, T_l, m)$, since any strong residual autocorrelations due to $\mathbf{x}^{\star}(\mathbf{s}, T_l, m)$ should be captured by $Z(\mathbf{s}, T_l, m)$.

Displaying the results

Large uncertainties surround our estimates of the SRKW intensity (i.e. encounter rate). Side-by-side maps of posterior mean and posterior standard deviation can prove challenging to interpret, making it difficult to determine regions of 'high' intensity. Instead, by using a large number of posterior samples (G) from our model, we are able to compute exceedance probabilities and then clearly display both point estimates with uncertainty in a single map, called an exceedance map.

Exceedance maps display the posterior pointwise probabilities that the value of a random surface evaluated across a regular lattice grid of points exceeds a chosen threshold. For our application, we are interested in identifying regions of high whale intensity. As such, our maps will display the posterior pointwise probabilities for month t that the pod-specific intensity $\lambda_{true}(\mathbf{s}, T_l, m)$, at location $\mathbf{s}$, lies above a chosen intensity threshold value. First, we choose the 70th percentile of that pod's intensity averaged across

all times and spatial grid pixels. Hotspots are then identified by displaying only the points that have a posterior pointwise probability above a probability threshold. We choose a probability threshold of 0.95, which represents areas where the model predicts with at least a probability of 0.95 that the posterior intensity is in the top 30% of values for that pod. Then, we repeat the above process but this time with the 70th percentile threshold fixed at a specific month. Such maps simultaneously present our point (i.e. 'best') estimates whilst also reflecting the uncertainties surrounding these estimates. For example, regions predicted to have a high encounter rate, but also a large uncertainty (e.g. regions rarely visited, but where a few encounters were made), will no longer appear in these exceedance plots.

For demonstration, we explore regions that our model confidently predicts to have a high J-pod intensity $\lambda_{true}(\mathbf{s}, T_l, m)$ during May. These correspond to hotspots of their UD. Panel A in Fig 5.6 shows the posterior probability that the J-pod intensity in May lies in the top 30% of values across all months. The plots show clear hotspots in J-pod's May intensity in the West of the region and in inshore waters. We repeat the plot, but now colour all pixels grey for which a posterior probability of exceeding the 70th percentile value is below 0.95. This helps to differentiate the regions of interest that we are most confident about (Fig 5.6 B). If we change the upper exceedance value to be the 70th percentile value for the month of May only, rather than across all months, the regions of interest are larger (Fig 5.6 C-D).

Pod-probability maps can identify the core areas within Ω associated with each pod and month. For a chosen month T_l and pod m, we define its 'core area' to be a region $D_{T_l,m} \subset \Omega$ such that if an encounter is made within $D_{T_l,m}$ during T_l, there is a 'high' probability that it is of pod m. Under our multi-type LGCP framework, because we can fix observer effort, we are able to compute the posterior probabilities that an encounter made at a given location and month contains a specific pod (see equation (5.9)). For May, we display the posterior probabilities that an encounter made at location $\mathbf{s} \in \Omega$ contain pod J, K and L respectively in panels E, F and G in Fig 5.6. It is apparent that in May, pod J is most likely to be encountered,

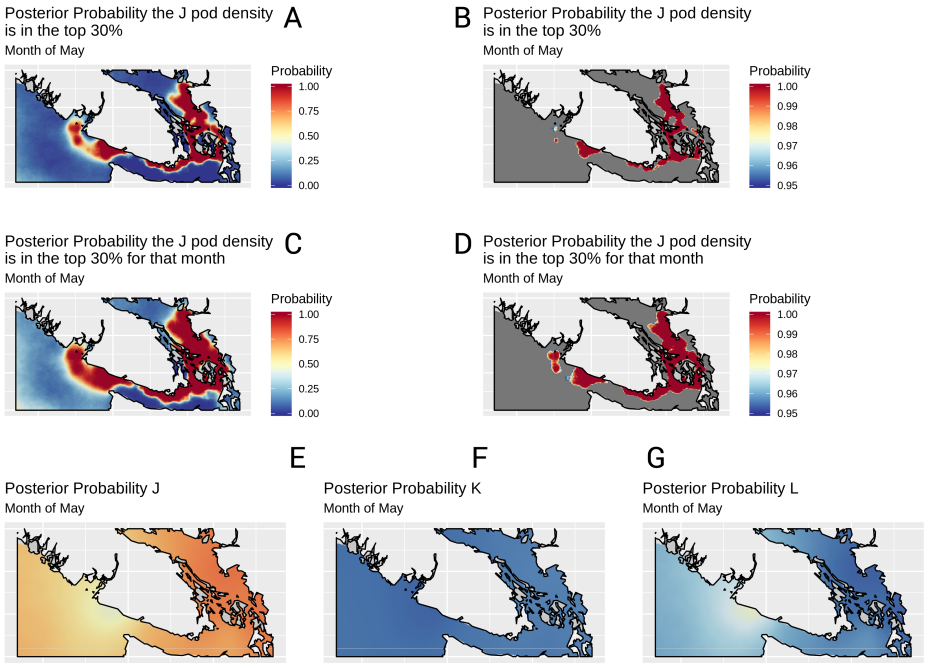

Figure 5.6: A series of plots demonstrating the different types of plots possible under our modelling framework. Panels A and B show the posterior probability that J pod's intensity across the region takes value in the upper 30% in the month of May. Panel A shows the raw probabilities, while Panel B has a minimum probability threshold of 0.95. Panels C and D are the same; however the upper 30% exceedance value is defined uniquely for the month of May instead of as an average over all the months. Panels E, F, and G show the posterior probabilities that a sighting made at a given location in May contains pods J, K and L respectively. All results are shown for the 'best' model with Monte Carlo observer effort error.

in agreement with [Ford et al., 1996].

When the sightings of every individual from the target population are available, one can sum the individuals' intensities and then normalize to create estimates of the population-level distribution. Maps of the population's distribution may be especially useful for conservation purposes. See

for example Fig A.32, where we fix the upper value to exceed as the 70th percentile value of the sum of the three pod's intensities across all months. We assume that individuals strictly swim in their pods, and that each pod is a single unit of identical size. We do not scale the pod-specific intensities by their group sizes. Thus, the intensity represents the expected number of encounters of any pod per boat hour of effort. The effort $E^{obs}_{Total}(\mathbf{s}, T_l, \mathbf{m})$ is estimated to be nonzero throughout most of the region Ω, with two exceptions. The first is the region in the very top of Ω, to the West of Vancouver. The second is in the Northwestern corner of Ω. These regions were never visited and so little can be said about the true SRKW intensity in these regions. This is reflected in the very large posterior standard deviations shown there in Fig A.36.

5.6 Discussion

We have built upon the recent developments made in the species distribution modelling literature and presented a general framework for estimating an individual's utilization distribution (UD). In addition, we have shown that these estimates can be combined to form the spatio-temporal distribution of a species or group. We demonstrated its use by identifying areas frequently used by an endangered ecotype of killer whale. Using the methodology, data from multiple observers, and data of varying quality and type, may all be combined to jointly estimate the spatio-temporal distribution. Crucially, high-quality survey data can be combined with low quality opportunistic data, including presence-only data. Data types compatible for modelling with this framework extend beyond those seen in this motivating example. Log-Gaussian Cox processes (LGCPs) have the unifying feature of providing a base model for deriving the likelihoods of many of the commonly found data, including presence-only, presence-absence, site occupancy, and site count data [Miller et al., 2019]. Such data fusion can improve the spatial resolution and statistical precision of estimates of the spatio-temporal distribution of species [Fithian et al., 2015, Koshkina et al., 2017].

However, including presence-only data requires knowledge about the ob-

server effort, either directly (e.g. GPS records) or through a set of strong predictors (e.g. distance from the nearest road). In either case, we show that approximating the observer effort by either computing or modelling the path integrals of the observers' fields-of-view can be a relatively straightforward and successful approach. Furthermore, results from our simulation study suggest that only crude estimates of the observers' fields-of-view are required, and that substantial improvements in the accuracy of UD predictions can be attained when the degree of observer bias is high. Furthermore, these improvements are still seen in settings where substantial overlap exists between the observers' fields-of-view and where the size of the study region is small. A fundamental assumption of our work was that the utilization distributions of the individuals were stationary throughout known time intervals. This greatly simplified the task of estimating the observer effort. If this stationarity assumption is unsuitable and the UDs evolve continuously through time, then observer effort needs to be known or estimated on a continuous-time scale too. Estimating unknown effort in continuous-time from a set of covariates will likely prove to be a challenge.

While the mathematical theory underpinning the LGCP may appear challenging to many researchers, the application of these models is widely applicable. Recent developments in spatial point process R packages [R Core Team, 2019], such as spatstat [Baddeley and Turner, 2014] and inlabru [Bachl et al., 2019] facilitate their computation. Inlabru requires only basic knowledge of R packages such as sp [Bivand et al., 2013, Pebesma and Bivand, 2005], rgeos [Bivand and Rundel, 2013], and rgdal [Bivand et al., 2015]. Pseudo-code is supplied in the Appendix to show how a dataset with a combination of distance sampling survey data and opportunistic presence-only data could be analysed using this modelling framework. Joint models are fit and sampled from using only 7 function calls, emphasising the applicability of the framework across a wide range of disciplines.

A biology-focused companion paper is currently underway, using the final model outputs to explore SRKW habitat use and how it varies in this region across pods and summer months. Importantly, it will compare and contrast habitat use based on traditional opportunistic sightings data

analyses and for the first time present relative SRKW habitat use across the entire extent of SRKW critical habitat in Canadian Pacific waters together with estimates of confidence. Thus the models developed in this Chapter will play an important role in planning future SRKW conservation efforts and highlighting regions of ecological significance.

Chapter 6

Summary, conclusions, and future work

"More data means more information, but it also means more false information."
— Nassim Nicholas Taleb, Antifragile: Things That Gain from Disorder

Data are being collected on an increasingly large number of phenomena, and the speed at which data are being created is increasing. The result is that enormous amounts of data are becoming available for every conceivable entity with which humans interact, hence the emergence of the term "Big Data". However, as stated in the above quote, an increase in the quantity of data collected on a phenomenon does not necessarily lead to an improved understanding of it. In this dissertation, we have shown that this disconnect can be stark when the phenomenon under study is spatio-temporal in nature because spatio-temporal data are routinely collected to meet one objective and then analyzed to meet another. We have shown that this mismatch in objectives, when ignored, can have a deleterious impact on the statistical inference of spatio-temporal data. Consequently, researchers need to question why and how a spatio-temporal dataset was collected to avoid preferential sampling (PS) biasing their understanding of the phenomenon under question.

In this dissertation, we focused on two major objectives. The first was to demonstrate that PS can have severe impacts on the statistical inference of spatio-temporal data and that it should be considered within any analysis. The second was to provide researchers with a set of tools for both detecting the presence of PS, and for subsequently adjusting for it in their analyses. Throughout the dissertation we focused our attention on real-world data, both to demonstrate that PS is prevalent in real-world data, and to show that our tools are applicable in practice.

In Chapter 2 we introduced the concepts of PS in all three types of spatio-temporal data, which we referred to as: discrete-space, continuous-space, and point-pattern data. We then focused on the popular spatio-temporal generalized linear mixed-effects (STGLMMs) class of models that are commonly used to describe the three data types in practice. Next, we provided demonstrative examples of PS in all three settings. In all the examples, we shared spatio-temporal random effects between the processes that described the PS and the target spatio-temporal process being observed. These demonstrative examples of PS in all three spatio-temporal data types, combined with the formal definition of STGLMMs, then provided us with the necessary framework required in the later Chapters for both developing tests for PS, and, for developing methods to adjust inference to its presence. We ended the Chapter with a discussion of the Integrated Nested Laplace Approximation (INLA) method that allows for models within the STGLMMs class to be efficiently fit within a Bayesian framework.

In Chapter 3, we built on the STGLMMs framework seen in Chapter 2 and developed the first general framework for modelling spatio-temporal data. The framework is applicable in both the discrete-space and continuous-space settings. We demonstrated its utility by analyzing historical air pollution data collected across a network in Great Britain. We demonstrated that PS was present throughout the lifetime of the network and that this may have led to a dramatic overestimation of black smoke levels, including estimates of population exposure.

In Chapter 4, we also considered the STGLMMs framework and used it to develop a general test for preferential sampling. The test is the first to be:

applicable in both discrete-space and continuous-space settings, applicable to non-continuous response data, and powerful in small sample size settings. We demonstrated the high power of the test across a wide range of settings in a thorough simulation study, before applying it to two previously-published real-world case-studies.

In Chapter 5, we focused on PS in point-pattern data. We turned to spatio-temporal log-Gaussian Cox processes and decomposed the spatio-temporal intensity surface into the product of a term reflecting the true intensity and an additional two terms that reflected both the spatial 'effort' exerted and the detectability of the points. We then used this approach to develop a framework for estimating the utilization distributions (UDs) of animals. UDs help ecologists to build a better understanding of how animals interact with their environment and use space. This information can then be used for informing successful management policies. We demonstrated its utility in a real-world case study to estimate the space use of an endangered ecotype of killer whales, using sightings data from observers who are known to focus their efforts in regions where the animals are expected to be present.

PS has been identified as a serious problem, with a recent surge in interest sparked by the landmark paper by Diggle et al. [2010]. Since that paper, it has been made clear that PS is commonplace across many fields of research including those related to ecology, public health, the environment, and econometrics [Gelfand and Shirota, 2019, Lee et al., 2015, Pennino et al., 2019, Shaddick et al., 2016]. Furthermore, as shown in Chapter 3 with an application to Great Britain's air pollution monitoring network, the impacts of PS on inference can be very large.

Whilst the application to Great Britain's air pollution monitoring network was ideal for demonstrating the utility of both the framework and test, the application was by no means cherrypicked. To appreciate how systemic PS may be throughout national air pollution monitoring networks, one only needs to read government guidelines for their design. Frequently, these networks are designed with the purposes of noncompliance detection and maximum concentration detection [Lee et al., 2015]. For example, consider the published guidelines for air quality monitoring network design from

the United States' EPA and Canada's CCME. In the EPA's handbook, it is suggested that monitors be placed to measure "concentrations in areas of high population density" and to determine the "highest concentration expected to occur in the area covered by the network" (von Lehmden and Nelson [1977], Section 6.1). In the CCME's guidelines, it is suggested that monitors be placed to "measure the highest representative ozone concentrations in metropolitan areas" and to "measure representative PM and ozone concentrations in populated areas across the country" (of Ministers of the Environment [2011], Executive Summary). Thus, PS may be rife in air pollution data.

Now it is important to state that this dissertation is not implying that there is a problem with the design of the air quality networks themselves. The networks are designed with a clear set of objectives in mind. The problem lies with researchers who decide to use these data for a purpose for which they were not collected. Researchers typically fail to incorporate these objectives within their statistical analyses which can severely impact their conclusions. The ramifications for our understanding of issues of great importance such as establishing the public health consequences associated with air pollution exposure, and monitoring compliance to regulatory air quality standards may be great (see Chapter 3 and Lee et al. [2015]).

Whilst existing methods for modelling PS have been developed, they have been criticized on philosophical grounds. In his discussion of the work by Diggle et al. [2010], Richard D. Wilkinson stated that the spatial-only model described "could never arise in practice as the surveyors do not know S [the value of the field]", before observing it. In addition to advancing the PS literature into the spatio-temporal setting, the general framework we introduced in Section 3 bypasses this philosophical concern. Instead of requiring that network designers had a complete understanding of the field before measurement, they are instead allowed to choose site placement based on the previous years' observations. Thus our approach gets closer to being able to emulate the real selection processes that truly governed the site selection.

Existing methods have also been criticized for being computationally

prohibitive and challenging to program [Pennino et al., 2019], requiring bespoke Monte Carlo-based maximum likelihood procedures to be written [Diggle et al., 2010]. All of the methods we introduced in Chapters 3, 4, and 5 were developed with computational efficiency and accessibility in mind. By framing the methods of Chapters 3 and 5 within the STGLMMs class of models, we ensured that the INLA approach can be used to quickly, efficiently, and accurately implement them. Importantly, the user-friendly *R-INLA* package [Lindgren et al., 2011a, Rue et al., 2009], when used in conjunction with the *inlabru* package [Bachl et al., 2019], ensures that the methodologies can be implemented by researchers with only modest computational skills required. For the test in Chapter 4, we developed a user-friendly R package *PSTestR*, now available on GitHub, for implementing the test. Crucially, the package is compatible with objects from the popular *R* packages *sp*, *sf*, and *spatstat*. Thus, we hope that the methods introduced in this dissertation will be widely used by researchers.

Statistical models used to describe spatio-temporal phenomena are increasingly being used to inform government policy across a range of issues including matters of public health and the environment. Whilst the use of evidence to inform policy is not new, there has been a recent increase in international demand for evidence-based policy (EBP) following its popularization in the UK during Tony Blair's leadership [Sutcliffe and Court, 2006]. The UK's increased interest in EBP began with the publication of a governmental white paper in 1999 titled "modernising government" which demanded a "better use of evidence and research in policy making" [Office, 1999]. Since then, the UK has been actively promoting the concept internationally [Sutcliffe and Court, 2006], and the demand for EBP has increased. In 2017, the US Commission on Evidence-Based Policymaking published a report titled "The Promise of Evidence-Based Policymaking" in which they proposed a vision of "a future in which rigorous evidence is created efficiently, as a routine part of government operations, and used to construct effective public policy" [on Evidence-Based Policymaking, 2017]. Yet, for policy-makers to implement EBP effectively, the evidence being 'created' needs to be both accurate and reliable. This dissertation has demonstrated

that the accuracy of evidence arising from standard spatio-temporal statistical methods can be strongly impacted by the protocols governing how the data were collected. The methods developed in this dissertation can help to reduce this sensitivity.

Whilst we have developed a suite of tools for accounting for PS in the statistical analysis of spatio-temporal data, a large range of research questions remain for future work. Firstly, in Chapter 3 we showed that estimates of Great Britain's population exposure to black smoke may have been wildly inaccurate. In spatial epidemiology, estimates of the health effects of exposure to air pollutants are commonly derived from ecological models of health, with exposure values typically imputed from models that ignore PS. Investigating the impacts that PS may have on these health effect estimates is an important avenue of research. We expect that PS will impact the imputed exposures derived from other air pollution monitoring networks to varying degrees. We are currently testing this hypothesis on air quality data from the United States.

Secondly, in Chapter 3, we shared linear combinations of spatio-temporally correlated latent effects between the model used to describe the environmental process and the model used to describe the PS. This joint model approach allowed us to adjust for PS. However, we never considered the possibility that the latent effects describing the spatio-temporal process being measured could influence the process governing the PS in a nonlinear way. In practice, there may often be a desire by the committees and processes responsible for choosing the sampling locations and times to seek out extreme values of the spatio-temporal process. A linear functional form would not suffice to represent this objective. It would be interesting in future work to investigate how sensitive the PS-adjusted predictions of the underlying spatio-temporal process are to the functional form specified on the PS.

Thirdly, in Chapter 4 we developed a test for PS in both the spatial and spatio-temporal settings. Yet these ideas could be carried across to the temporal setting. In bio-statistics, longitudinal studies are commonly employed to investigate the population-average temporal response to treatments. These studies frequently suffer from patient dropout and the pro-

cesses driving the decision of a patient to dropout of a study are commonly feared to be related to the underlying patient's treatment response being measured. This relation between dropout and treatment response is known as informative dropout and joint models are commonly fit to adjust for its biasing effects [Wu, 2009]. Yet this type of informative dropout falls within our definition of PS. Thus, by studying the patient dropout times as the realization of a temporal point process whose intensity may be related to the underlying treatment response, a test for informative dropout (i.e. PS) may be derived using the same principles as seen in Chapter 4. Such a test would allow researchers to quickly assess the evidence for PS within a given study, without having to fit conceptually-difficult and computationally-costly joint models.

Finally, in Chapter 5 we developed an approach for adjusting the statistical inference of point-patterns to the presence of PS. This required us to either emulate the process that governed the PS, or estimate the PS with a set of available covariates. However, in many ecological settings, aggregated sightings data is collected at discrete 'sites'. For example, site count data is frequently collected when researchers visit discrete locations ('sites') and record the number of sightings of the target species. In these settings, one could consider the locations of 'sites' as a point-pattern and jointly model the intensity governing the chosen 'sites' with the underlying species' intensity describing the count data. By sharing the latent effects across the likelihoods, this joint model would allow researchers to investigate whether or not the chosen 'site' locations were preferentially sampled. Statistical inference on the species' distributions could then be adjusted for PS accordingly, without the need for any covariates or emulators of effort.

Bibliography

Jonathan Acosta, Ronny Vallejos, et al. Effective sample size for spatial regression models. *Electronic Journal of Statistics*, 12(2):3147–3180, 2018. → page 129

B. Ainslie, C. Reuten, DG Steyn, N.D. Le, and J.V. Zidek. Application of an entropy-based bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants. *Journal of environmental management*, 90(8):2715–2729, 2009. → page 42

Fabian E. Bachl, Finn Lindgren, David L. Borchers, and Janine B. Illian. inlabru: an R package for bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10:760–766, 2019. doi:10.1111/2041-210X.13168. → pages 33, 136, 146, 172, 178, 248

Adrian Baddeley and Rolf Turner. Package 'spatstat'. *The Comprehensive R Archive Network ()*, 2014. → pages 146, 172

Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial point patterns: methodology and applications with R*. Chapman and Hall/CRC, 2015. → pages 10, 22, 23, 30, 50, 90, 101, 102, 103, 117, 128, 145, 200, 214

Adrian Baddeley, Andrew Hardegen, Thomas Lawrence, Robin K Milne, Gopalan Nair, and Suman Rakshit. On two-stage monte carlo tests of composite hypotheses. *Computational Statistics & Data Analysis*, 114: 75–87, 2017. → page 113

Haakon Bakka. Mesh creation including coastlines, Jan 2017. URL https://haakonbakka.bitbucket.io/btopic104.html#what-is-needed-for-a-good-mesh. → page 198

Haakon Bakka, Håvard Rue, Geir-Arne Fuglstad, Andrea Riebler, David Bolin, Janine Illian, Elias Krainski, Daniel Simpson, and Finn Lindgren.

Spatial modeling with r-inla: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6):e1443, 2018. → page 2

S. Banerjee, B. P. Carlin, and A.E. Gelfand. *Hierarchical modeling and analysis for spatial data. Second Edition.* CRC Press, 2015. → page 1

Luis Bedriñana-Romano, Rodrigo Hucke-Gaete, Francisco Alejandro Viddi, Juan Morales, Rob Williams, Erin Ashe, José Garcés-Vargas, Juan Pablo Torres-Florez, and Jorge Ruiz. Integrating multiple data sources for assessing blue whale abundance and distribution in chilean northern patagonia. *Diversity and Distributions*, 24(7):991–1004, 2018. → page 133

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974a. → page 115

J. Besag and C. Kooperberg. On conditional and intrinsic auto-regressions. *Biometrics*, 82:733–746, 1995. → page 11

J.E. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Socety, Series B*, 36:192–236, 1974b. → page 9

Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991. → page 28

Roger Bivand and Colin Rundel. rgeos: interface to geometry engine-open source (geos). *R package version 0.3-2*, 2013. → page 172

Roger Bivand, Tim Keitt, Barry Rowlingson, Edzer Pebesma, Michael Sumner, Robert Hijmans, Even Rouault, and Maintainer Roger Bivand. Package 'rgdal'. *Bindings for the Geospatial Data Abstraction Library. Available online: https://cran. r-project. org/web/packages/rgdal/index. html (accessed on 15 October 2017)*, 2015. → page 172

Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition.* Springer, NY, 2013. URL http://www.asdar-book.org/. → pages 128, 172, 250

Marta Blangiardo and Michela Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015. → pages 6, 9, 10, 109, 115

David R Brillinger, Haiganoush K Preisler, Alan A Ager, and JG Kie. The use of potential functions in modelling animal movement. In *Selected Works of David Brillinger*, pages 385–409. Springer, 2012. → page 155

Avishek Chakraborty, Alan E Gelfand, Adam M Wilson, Andrew M Latimer, and John A Silander. 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, 2011. → pages 23, 135, 149

H. Chang, A.Q. Fu, N.D. Le, and J.V. Zidek. Designing environmental monitoring networks to measure extremes. *Environmental and Ecological Statistics*, 14(3):301–321, 2007. → page 42

Taeryon Choi and Mark J Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98 (10):1969–1987, 2007. → page 112

John Clare, Shawn T McKinney, John E DePue, and Cynthia S Loftin. Pairing field methods to improve inference in wildlife surveys while accommodating detection covariance. *Ecological applications*, 27(7): 2031–2047, 2017. → page 149

Jean-François Coeurjolly, Jesper Møller, and Rasmus Waagepetersen. Palm distributions for log gaussian cox processes. *Scandinavian Journal of Statistics*, 44(1):192–203, 2017. → page 107

J. Colls. *Air pollution, modelling, and mitigation.* Routledge, Abingdon, Oxford, 2002. → page 53

Paul B Conn, James T Thorson, and Devin S Johnson. Confronting preferential sampling when analyzing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution*, 8:1535–1545, 2017. → page 48

N. Cressie and H.-C. (Huang. Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Am. Statist. Assoc. Statist. Assoc.*, 94:1330–40, 1999. → page 109

N. Cressie and C.K. Wikle. *Statistics for spatio-temporal data*, volume 465. Wiley, 2011. → pages 1, 6

N.A.C. Cressie. *Statistics for Spatial Data, Revised edition.* John Wiley, New York, 1993. → pages 1, 5

Noel Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992. → page 112

Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015. → page 96

Ngoc Anh Dao and Marc G Genton. A monte carlo-adjusted goodness-of-fit test for parametric models describing spatial point patterns. *Journal of Computational and Graphical Statistics*, 23(2): 497–517, 2014. → pages 113, 119

Russell Davidson and James G MacKinnon. Bootstrap tests: How many bootstraps? *Econometric Reviews*, 19(1):55–68, 2000. → pages 114, 118

DFO. Killer whale (northeast pacific southern resident population). http://www.dfo-mpo.gc.ca/species-especes/profiles-profils/killerWhalesouth-PAC-NE-epaulardsud-eng.html. Accessed: 2019-03-29. → page 136

Peter J Diggle, JA Tawn, and RA Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47 (3):299–350, 1998. → page 28

PJ Diggle and Paulo Justiniano Ribeiro. Model-based geostatistics (springer series in statistics). 2007. → pages 100, 104, 108

P.J. Diggle, P.J. Ribeiro, and SpringerLink (Service en ligne). *Model-based geostatistics*, volume 846. Springer New York, 2007. → pages 8, 10

P.J. Diggle, R. Menezes, and T. Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232, 2010. → pages 2, 4, 16, 19, 46, 48, 51, 64, 66, 68, 78, 96, 98, 100, 125, 127, 128, 176, 177, 178, 240

Daniel Dinsdale, Matias Salibian-Barrera, et al. Modelling ocean temperatures from bio-probes under preferential sampling. *The Annals of Applied Statistics*, 13(2):713–745, 2019. → pages 97, 128

Robert M Dorazio. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, 23(12):1472–1484, 2014. → pages 146, 148

Shah Ebrahim and George Davey Smith. Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Human genetics*, 123(1):15–33, 2008. → page 1

Jane Elith and John Leathwick. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and distributions*, 13 (3):265–275, 2007. → pages 133, 146, 149

EPA. Air quality criteria for ozone and related photochemical oxidants. Technical report, http://oaspub.epa.gov/eims/eimsapi.dispdetail?deid=149923, 2005. → pages 2, 42, 97

Doug Sandilands Iain U. Smith Alana. V. Phillips Lance G. Barrett-Lennard Erin U. Rechsteiner, Caitlin F. C. Birdsall. Quantifying observer effort for opportunistically-collected wildlife sightings. Unpublished - url: https://killerwhale.vanaqua.org/document.doc?id=140, 2013. Accessed: 2019-03-29. → page 139

JA Fernández, A Rey, and A Carballeira. An extended study of heavy metal deposition in galicia (nw spain) based on moss analysis. *Science of the Total Environment*, 254(1):31–44, 2000. → page 125

John Fieberg and Luca Börger. Could you please phrase "home range" as a question? *Journal of mammalogy*, 93(4):890–902, 2012. → page 132

William Fithian and Trevor Hastie. Finite-sample equivalence in statistical models for presence-only data. *The annals of applied statistics*, 7(4): 1917, 2013. → pages 44, 49, 66, 200

William Fithian, Jane Elith, Trevor Hastie, and David A Keith. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4): 424–438, 2015. → pages 2, 22, 24, 32, 33, 97, 99, 133, 145, 146, 148, 149, 153, 171

Chris H Fleming, William F Fagan, Thomas Mueller, Kirk A Olson, Peter Leimgruber, and Justin M Calabrese. Rigorous home range estimation with movement data: a new autocorrelated kernel density estimator. *Ecology*, 96(5):1182–1188, 2015. → page 132

John KB Ford, Graeme M Ellis, and Kenneth C Balcomb. *Killer whales: the natural history and genealogy of Orcinus orca in British Columbia and Washington*. UBC press, 1996. → pages 137, 167, 170

John KB Ford, James F Pilkington, M Otsuki, B Gisborne, RM Abernethy, EH Stredulinsky, JR Towers, and GM Ellis. *Habitats of special importance to Resident Killer Whales (Orcinus orca) off the west coast of Canada*. Fisheries and Oceans Canada, Ecosystems and Oceans Science, 2017. → pages 137, 244

Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, pages 1–8, 2018. → pages 29, 124, 213, 245, 250

Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, 114(525): 445–452, 2019. → pages 69, 197

Alan E Gelfand and Shinichiro Shirota. Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, 89(3):e01372, 2019. → pages 2, 176

Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010. → page 30

Alan E Gelfand, Sujit K Sahu, and David M Holland. On the effect of preferential sampling in spatial prediction. *Environmetrics*, 23(7): 565–578, 2012. → pages 17, 47, 96, 97, 105

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996. → page 167

Subhashis Ghosal, Anindya Roy, et al. Posterior consistency of gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34(5):2413–2429, 2006. → page 112

Jacques Gignoux, Camille Duby, and Sébastian Barot. Comparing the performances of diggle's tests of spatial randomness for small samples with and without edge-effect correction: application to ecological data. *Biometrics*, 55(1):156–164, 1999. → page 102

Christophe Giraud, Clément Calenge, Camille Coron, and Romain Julliard. Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics*, 72(2):649–658, 2016. → pages 133, 134

Richard Glennie, Stephen T Buckland, and Len Thomas. The effect of animal movement on line transect estimates of abundance. *PloS one*, 10 (3):e0121333, 2015. → page 134

Richard Glennie, Stephen Terrence Buckland, Roland Langrock, Tim Gerrodette, LT Ballance, SJ Chivers, and MD Scott. Incorporating animal movement into distance sampling. *Journal of the American Statistical Association*, (just-accepted):1–17, 2020. → pages 135, 144, 148, 155, 159

Virgilio Gómez-Rubio. *Bayesian inference with INLA*. CRC Press, 2020. → pages 27, 38

Virgilio Gómez-Rubio and Håvard Rue. Markov chain monte carlo with the integrated nested laplace approximation. *Statistics and Computing*, 28(5):1033–1051, 2018. → page 38

Yongtao Guan and David R Afshartous. Test for independence between marks and points of marked point processes: a subsampling approach. *Environmental and Ecological Statistics*, 14(2):101–111, 2007. → pages 98, 128

Marc Hallin, Zudi Lu, and Lanh T Tran. Kernel density estimation for spatial processes: the l1 theory. *Journal of Multivariate Analysis*, 88(1): 61–75, 2004. → page 6

Ephraim M Hanks, Erin M Schliep, Mevin B Hooten, and Jennifer A Hoeting. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26(4):243–254, 2015. → page 218

Donna DW Hauser, Glenn R Van Blaricom, Elizabeth E Holmes, and Richard W Osborne. Evaluating the use of whalewatch data in determining killer whale (orcinus orca) distribution patterns. *Journal of Cetacean Research and Management*, 8(3):273, 2006. → pages 137, 139

Donna DW Hauser, Miles G Logsdon, Elizabeth E Holmes, Glenn R VanBlaricom, and Richard W Osborne. Summer distribution patterns of southern resident killer whales orcinus orca: core areas and spatial segregation of social groups. *Marine Ecology Progress Series*, 351: 301–310, 2007. → pages 137, 138, 139, 167

James J Heckman. Selection bias and self-selection. In *Econometrics*, pages 201–224. Springer, 1990. → page 1

Trevor J Hefley and Mevin B Hooten. Hierarchical species distribution models. *Current Landscape Ecology Reports*, 1(2):87–97, 2016. → pages 22, 24, 134, 154, 239

Miguel A Hernan and James M Robins. Causal inference: What if, 2020. → pages 1, 151, 152, 237

Nigel E Hussey, Steven T Kessel, Kim Aarestrup, Steven J Cooke, Paul D Cowley, Aaron T Fisk, Robert G Harcourt, Kim N Holland, Sara J Iverson, John F Kocik, et al. Aquatic animal telemetry: a panoramic window into the underwater world. *Science*, 348(6240), 2015. → page 133

Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons, 2008. → pages 10, 101, 102, 103

EH Isaaks and R Mohan Srivastava. Spatial continuity measures for probabilistic and deterministic geostatistics. *Mathematical geology*, 20 (4):313–341, 1988. → page 46

Devin S. Johnson and Josh M. London. crawl: an r package for fitting continuous-cime correlated random walk models to animal movement data, 2018. URL https://doi.org/10.5281/zenodo.596464. → page 162

Devin S. Johnson, Josh M. London, Mary-Anne Lea, and John W. Durban. Continuous-time correlated random walk model for animal telemetry data. *Ecology*, 89(5):1208–1215, 2008. doi:10.1890/07-1032.1. URL https://doi.org/10.1890/07-1032.1. → page 162

Devin S Johnson, Mevin B Hooten, and Carey E Kuhn. Estimating animal resource selection from telemetry data using point process models. *Journal of Animal Ecology*, 82(6):1155–1164, 2013. → pages 132, 134

Olatunji Johnson, Peter Diggle, and Emanuele Giorgi. A spatially discrete approximation to log-gaussian cox processes for modelling aggregated disease count data. *Statistics in medicine*, 38(24):4871–4887, 2019. → page 22

Matthias Katzfuss, Joseph Guinness, Wenlong Gong, and Daniel Zilber. Vecchia approximations of gaussian-process predictions. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–32, 2020. → page 10

JF Kingman. Poisson processes. *Oxford Studies in Probability. Oxford: Oxford University Press*, 1994. → page 227

Vira Koshkina, Yan Wang, Ascelin Gordon, Robert M Dorazio, Matt White, and Lewi Stone. Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, 8(4):420–430, 2017. → pages 133, 134, 153, 154, 171

Elias T Krainski, Virgilio Gómez-Rubio, Haakon Bakka, Amanda Lenzi, Daniela Castro-Camilo, Daniel Simpson, Finn Lindgren, and Håvard Rue. *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. CRC Press, 2018. → pages 36, 37

Kalimuthu Krishnamoorthy. *Handbook of statistical distributions with applications*. CRC Press, 2016. → page 28

A Lawrence Gould, Mark Ernest Boye, Michael J Crowther, Joseph G Ibrahim, George Quartey, Sandrine Micallef, and Frederic Y Bois. Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. *Statistics in medicine*, 34(14):2181–2195, 2015. → page 89

N.D. Le and J.V. Zidek. *Statistical analysis of environmental space-time processes*. Springer Verlag, 2006. → pages 6, 29

A Lee, A Szpiro, SY Kim, and L Sheppard. Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics*, 26(4):255–267, 2015. → pages 97, 176, 177

Subhash R Lele, Evelyn H Merrill, Jonah Keim, and Mark S Boyce. Selection, use, choice and occupancy: clarifying concepts in resource selection studies. *Journal of Animal Ecology*, 82(6):1183–1191, 2013. → page 132

Qiuju Li and Li Su. Accommodating informative dropout and death: a joint modelling approach for longitudinal and semicompeting risks data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67 (1):145–163, 2018. → page 89

F. Lindgren, H. Rue, and J Lindstö. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial

differential equation approach. *Roy Statist Soc, Ser B*, page To appear, 2011a. → pages 57, 178

Finn Lindgren, Håvard Rue, and Johan Lindström. 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, 2011b. → pages 9, 10, 37, 45, 124, 166, 201, 213, 243

Finn Lindgren, Havard Rue, et al. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19):1–25, 2015. → pages 124, 166, 213, 243

Nicola Loperfido and Peter Guttorp. Network bias in air quality monitoring design. *Environmetrics*, 19(7):661–671, 2008. → pages 2, 42, 97

B Matern. Doubly stochastic poisson processes in the plane. *Statistical ecology*, 1:195–213, 1971. → page 107

Robert McMillan and Joshua Murphy. Measuring the effects of severe air pollution:evidence from the uk clean air act. June 2017. → pages 71, 82

Dana Michalcová, Samuel Lvončik, Milan Chytrý, and Ondřej Hájek. Bias in vegetation databases? a comparison of stratified-random and preferential sampling. *Journal of Vegetation Science*, 22(2):281–291, 2011. → page 46

David AW Miller, Krishna Pacifici, Jamie S Sanderlin, and Brian J Reich. The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10(1): 22–37, 2019. → pages 133, 134, 171

Rua S Mordecai, Brady J Mattsson, Caleb J Tzilkowski, and Robert J Cooper. Addressing challenges when studying mobile or episodic species: hierarchical bayes estimation of occupancy and use. *Journal of Applied Ecology*, 48(1):56–66, 2011. → page 132

Tomáš Mrkvička, Mari Myllymäki, and Ute Hahn. Multiple monte carlo testing, with applications in spatial point processes. *Statistics and Computing*, 27(5):1239–1255, 2017. → page 112

Mari Myllymäki, Tomáš Mrkvička, Pavel Grabarnik, Henri Seijo, and Ute Hahn. Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):381–404, 2017. → page 112

NOAA. Endangered species act status of puget sound killer whales. https://www.westcoast.fisheries.noaa.gov/protected_species/marine_mammals/killer_whale/esa_status.html. Accessed: 2019-09-27. → page 136

Canadian Council of Ministers of the Environment. Ambient air monitoring protocol for pm2.5 and ozone. canada-wide standards for particulate matter and ozone. Technical report, Canadian Council of Ministers of the Environment, 2011. → page 177

Cabinet Office. Modernising government, 1999. → page 178

Ricardo A Olea. Declustering of clustered preferential sampling for histogram and semivariogram inference. *Mathematical Geology*, 39(5): 453–467, 2007. → page 46

Jennifer K Olson, Jason Wood, Richard W Osborne, Lance Barrett-Lennard, and Shawn Larson. Sightings of southern resident killer whales in the salish sea 1976–2014: the importance of a long-term opportunistic dataset. *Endangered Species Research*, 37:105–118, 2018. → pages 138, 139

United States. Commission on Evidence-Based Policymaking. *The promise of evidence-based policymaking: Report of the Commission on Evidence-Based Policymaking*. Commission on Evidence-Based Policymaking, 2017. → page 178

Lucia Paci, Alan E Gelfand, Beamonte, María Asunción, Pilar Gargallo, and Manuel Salvador. Spatial hedonic modelling adjusted for preferential sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1):169–192, 2020. → page 2

Krishna Pacifici, Brian J Reich, David AW Miller, Beth Gardner, Glenn Stauffer, Susheela Singh, Alexa McKerrow, and Jaime A Collazo. Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, 98(3):840–850, 2017. → pages 133, 135, 145

Debdeep Pati, Brian J Reich, and David B Dunson. Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, 98(1):35–48, 2011. → pages 47, 51, 64

Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018. doi:10.32614/RJ-2018-009. URL https://doi.org/10.32614/RJ-2018-009. → page 128

Edzer J. Pebesma and Roger S. Bivand. Classes and methods for spatial data in R. *R News*, 5(2):9–13, November 2005. URL https://CRAN.R-project.org/doc/Rnews/. → pages 128, 172, 250

Maria Grazia Pennino, Iosu Paradinas, Janine B Illian, Facundo Muñoz, José María Bellido, Antonio López-Quílez, and David Conesa. Accounting for preferential sampling in species distribution models. *Ecology and evolution*, 9(1):653–663, 2019. → pages 2, 97, 128, 134, 153, 176, 178, 240

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/. → pages 136, 166, 172, 250

T.;Steinle S.;Carnell E.;Leaver-D.;Roberts E.;Vieno M.;Beck R.;Dragosits U. Reis, S.;Liska. Uk gridded population 2011 based on census 2011 and land cover map 2015. nerc environmental information data centre., 2017. URL https://doi.org/10.5285/0995e94d-6d42-40c1-8ed4-5090d82471e1. → pages 82, 124

Ian W Renner and David I Warton. Equivalence of maxent and poisson point process models for species distribution modeling in ecology. *Biometrics*, 69(1):274–281, 2013. → page 167

Ian W Renner, Jane Elith, Adrian Baddeley, William Fithian, Trevor Hastie, Steven J Phillips, Gordana Popovic, and David I Warton. Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379, 2015. → page 133

Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007. → page 29

J Andrew Royle, Marc Kery, and Jerome Guelat. Spatial capture-recapture models for search-encounter data. *Methods in Ecology and Evolution*, 2(6):602–611, 2011. → pages 132, 133

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009. → pages 30, 33, 35, 37, 38, 45, 57, 124, 166, 178, 201, 213, 243

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005. → page 9

Håvard Rue, Andrea Riebler, Sigrunn H Sørbye, Janine B Illian, Daniel P Simpson, and Finn K Lindgren. Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421, 2017. → pages 37, 45, 57, 201

Ramiro Ruiz-Cárdenas, Elias T Krainski, and Håvard Rue. Direct fitting of dynamic models using integrated nested laplace approximations—inla. *Computational Statistics & Data Analysis*, 56(6):1808–1828, 2012. → page 199

Mark J Schervish. *Theory of statistics*. Springer Science & Business Media, 2012. → page 37

Martin Schlather, Paulo J Ribeiro, and Peter J Diggle. Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):79–93, 2004. → pages 11, 97, 101, 128

P. Schumacher and J.V. Zidek. Using prior information in designing intervention detection experiments. *The Annals of Statistics*, pages 447–463, 1993. → pages 2, 42, 97

Elizabeth Seely, Richard W Osborne, Kari Koski, and Shawn Larson. Soundwatch: Eighteen years of monitoring whale watch vessel activities in the salish sea. *PloS one*, 12(12):e0189764, 2017. → pages 138, 163

T Sellke, MJ Bayarri, and JO Berger. Calibration of p-values for precise null hypotheses. *The American Statistician*, 2001. → page 114

Gavin Shaddick and James V Zidek. A case study in preferential sampling: Long term monitoring of air pollution in the uk. *Spatial Statistics*, 9: 51–65, 2014. → pages 42, 53, 54, 58, 60, 87, 123, 198

Gavin Shaddick, James V Zidek, and Yi Liu. Mitigating the effects of preferentially selected monitoring sites for environmental policy and health risk analysis. *Spatial and spatio-temporal epidemiology*, 18:44–52, 2016. → pages 2, 176

Gavin Shaddick, Matthew L Thomas, Amelia Green, Michael Brauer, Aaron van Donkelaar, Rick Burnett, Howard H Chang, Aaron Cohen, Rita Van Dingenen, Carlos Dora, et al. Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(1):231–253, 2018. → page 97

R.A Simons. Erddap. erddap. https://coastwatch.pfeg.noaa.gov/erddap, 2019. → page 166

Daniel Simpson, Janine B Illian, Finn Lindgren, Sigrunn H Sørbye, and Havard Rue. Going off grid: Computationally efficient inference for log-gaussian cox processes. *Biometrika*, 103(1):49–70, 2016. → pages 20, 21, 23, 103, 105, 143, 145, 242

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, Sigrunn H Sørbye, et al. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28, 2017. → pages 29, 69, 197

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4): 583–639, 2002. → page 167

Sophie Sutcliffe and Julius Court. *A toolkit for progressive policymakers in developing countries*. Research and Policy in Development Programme, 2006. → page 178

P Switzer. Estimation of spatial distributions from point sources with application to air pollution measurement. technical report no. 9. Technical report, Stanford Univ., CA (USA). Dept. of Statistics, 1977. → page 46

Benjamin M Taylor and Peter J Diggle. 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, 2014. → pages 36, 38

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58 (1):267–288, 1996. → page 29

Maria Nicolette Margaretha van Lieshout. *Theory of Spatial Statistics: A Concise Introduction*. CRC Press, 2019. → pages 6, 7, 10, 11

Vancouver Aquarium. BC Cetacean Sightings Network. http://wildwhales.org/. Accessed: 2019-03-29. → page 138

DJ von Lehmden and C Nelson. Quality assurance handbook for air pollution measurement systems. volume ii. ambient air specific methods. Technical report, Environmental Protection Agency, Research Triangle Park, NC (USA ..., 1977. → page 177

David I Warton, Leah C Shepherd, et al. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, 4(3):1383–1402, 2010. → pages 30, 41, 44, 49, 66, 133, 200

David I Warton, Ian W Renner, and Daniel Ramp. Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS one*, 8(11):e79168, 2013. → page 33

Ronald L Wasserstein and Nicole A Lazar. The asa statement on p-values: context, process, and purpose, 2016. → page 114

Kim Whoriskey, Eduardo G Martins, Marie Auger-Méthé, Lee FG Gutowsky, Robert J Lennox, Steven J Cooke, Michael Power, and Joanna Mills Flemming. Current and emerging statistical techniques for aquatic telemetry data: A guide to analysing spatially discrete animal detections. *Methods in Ecology and Evolution*, 10(7):935–948, 2019. → page 133

Christopher K Wikle. Modern perspectives on statistics for spatio-temporal data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):86–98, 2015. → page 6

Simon N Wood, Zheyuan Li, Gavin Shaddick, and Nicole H Augustin. Generalized additive models for gigadata: modeling the uk black smoke network daily data. *Journal of the American Statistical Association*, 112 (519):1199–1210, 2017. → page 6

Brian J Worton. Kernel methods for estimating the utilization distribution in home-range studies. *Ecology*, 70(1):164–168, 1989. → page 132

Lang Wu. *Mixed effects models for complex data*. Chapman and Hall/CRC, 2009. → pages 13, 89, 180

Yuan Yuan, Fabian E Bachl, Finn Lindgren, David L Borchers, Janine B Illian, Stephen T Buckland, Haavard Rue, Tim Gerrodette, et al. Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *The Annals of Applied Statistics*, 11 (4):2270–2297, 2017. → pages 21, 133, 134, 135, 145, 147, 154, 166, 244

H. Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004. ISSN 0162-1459. → pages 29, 199

Sha Zhe. Mesh creation including coastlines, Aug 2017. URL http://rstudio-pubs-static.s3.amazonaws.com/302639_3ccd091c277d4d6c9ad9c2cf524250b6.html. → page 198

J.V. Zidek, G. Shaddick, and C.G. Taylor. Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. *Ann Applied Statistics*, page To appear, 2014. → pages 2, 82, 87, 99

Botta-Duká Zoltán, Edit Kovács-Láng, Tamás Rédei, Miklós Kertész, and János Garadnai. Statistical and biological consequences of preferential sampling in phytosociology: theoretical considerations and a case study. *Folia Geobotanica*, 42(2):141–152, 2007. → page 46

Appendix A

Supporting Materials

A.1 Chapter 3 Supporting Materials

A.1.1 Chosen priors for the case study

For the Y process, we used weakly informative Gaussian priors for the γ_k's. We used a Gamma(a, b) prior for the precision parameter $1/\sigma_\epsilon^2$, where a denotes the shape parameter and b denotes the inverse-scale parameter. We chose $a = 1$ and $b = 5 \times 10^-5$. Under this parameterisation, the mean and variance of this distribution are a/b and a/b^2 respectively. Thus this prior assumption allows for very large and very small variances of the response to exist. Next, a 2D Wishart distribution is assumed for Σ_b^{-1} with four degrees of freedom. The prior matrix is given 0 off–diagonal elements and diagonal values of 1. This results in a prior mean for the two variance terms of the random effects $(\sigma_{b,1}^2, \sigma_{b,2}^2)$ of 4 with a prior variance for these terms equal to 8. The prior mean for the correlation term is 0 with variance for the logit transform of the correlation equal to 4. This allows for random effects with a large range of magnitudes and correlation structures to exist. We place the PC joint priors [Fuglstad et al., 2019, Simpson et al., 2017] on the two hyperparameters for the 3 independent Matern realisations, with prior belief that the 5th percentile for the range is 3.4km (a fifth of the smallest range found in previous analyses) and the 99th percentile for the standard

deviation of each field is 1 (noting that the data have been transformed). We fix the Matern roughness parameter to equal 1 since this is the largest smoothness value currently implemented in R-INLA, and we assume a-priori that the medium–range pollution process will be reasonably smooth. The lower prior bound on the range parameter, combined with the probabilistic upper bound on the variance, should help prevent the model from collapsing into a state that over–fits the data.

For the site–selection process R, our choice of priors follows the same objectives as for the observation process. Weakly informative Gaussian priors were placed on all the α terms. The same PC prior chosen for the observation process was placed on the $\beta_0^\star(\mathbf{s})$ field. For the first order autoregressive term $\beta_1^\star(t)$, we placed a Gamma$(1, 5 \times 10^-5)$ on the marginal precision and a $N(0, 0.15)$ prior was placed on the logit of the lag 1 correlation (i.e. on $\log((1+\rho_a)/(1-\rho_a))$) to allow for a large degree of flexibility. Finally, we consider two different sets of priors for the PS parameters d_b, d_β. For Implementation 1 we constrain these to equal 0 and thus we can view this as setting a point mass prior at 0. For implementations 2 and 3, we assign a $N(0, 10)$ prior to allow PS to be detected.

A.1.2 Details on the R-INLA implementation

We used the estimated ranges from Shaddick and Zidek [2014] to construct the Delauney triangulation mesh required for use in R–INLA. Following the advice of Bakka [2017], Zhe [2017], and trading it off with the need for maintaining a reasonable computation time, we set the edge lengths of the triangles throughout the domain to be around 5km, less than the minimum estimated range of 17km found in Shaddick and Zidek [2014]. This is important since it has been shown that the length of the triangle edges must be less than the range of any Matern field and should ideally be less than a quarter of this. Failure to do so leads to large errors in the approximation of the Gaussian random field. We are confident that with our choice of mesh, any changes to the inference in the unsampled regions will be a direct result of our joint model framework and not due to any

undesirable artifacts caused by a poor choice of triangulation mesh for the SPDE approximation.

It is well known that an empirical Bayes or maximum likelihood approach does not fully account for the uncertainties in the hyperparameters when performing predictions and inference, and these may be high in spatially correlated Gaussian random fields [Zhang, 2004]. Interestingly, for this dataset we compared the fully Bayesian approach with the empirical Bayes method using R-INLA and found little difference. The posterior credible intervals for the latent effects and parameters were slightly wider under the fully Bayesian approach, however the posterior credible intervals for the predictions were almost identical. Additionally we used the empirical Bayes approach in a small simulation study with good results. Thus for computational savings we opted to consider only empirical Bayes methods.

In R-INLA, copying across a linear combination of latent processes (potentially from a different time point) requires the use of dummy variables. In particular, the idea of Ruiz-Cárdenas et al. [2012] is required. This simply involves creating infinite precision Gaussian variables with observed values of zero and with linear predictor set equal to the (negative) linear combination of latent processes desired, plus an infinite variance random intercept process. It is not hard to see that the values of these random intercepts equal precisely the values of the linear combination of the desired processes. This approach proved vital for fitting implementations 2 and 3.

Note that in essence, for Implementation 3 we are modelling the initial site–placement process as a LGCP, but using a Bernoulli likelihood as a pseudo-likelihood instead of the usual Poisson likelihood to form the computational approximation. We use the conditional logistic regression approach, commonly used to fit Poisson point processes, placing the zeros in a regular (not a latticed) manner throughout Ω, independent from the observed site locations. In practice, we created a reasonably regular delauney triangulation mesh in R-INLA throughout Ω for our GMRF with mesh vertices placed independent from the observed site locations. Regularity was enforced through a combination of the choices of a minimum vertex length of 5km, an upper vertex length of 7km and a minimum angle of 25 degrees.

We then used the created mesh vertices as our pseudo–sites.

A somewhat undesirable property of using the logistic regression approach is that the likelihood value does not converge as the number of pseudo zeros tends towards infinity. Thus, unlike the result of using the Poisson approximation to a Point Process, convergence must instead be judged with the convergence of fixed parameter estimates, excluding the estimate of the intercept. However, if the Poisson approximation is chosen, then it cannot be used to simultaneously model the retention process alongside the site–placement process and hence a third Bernoulli likelihood modelling the retention–process would be required. Thus in either case, there is a trade-off. Given that the computational time required to fit the model in R-INLA using the SPDE approach is affected more by the resolution of the computational mesh than by the number of observations, we can increase the density of the pseudo–sites with a reasonably small effect on the total computation time.

Thus for fitting Implementation 3, we follow the advice given in the literature [Fithian and Hastie, 2013, Warton et al., 2010]. We repeatedly re–fit the joint model on an ever-increasing density of pseudo–sites until the parameters and predictions converge. We found that all estimates, except of course the site–selection intercept, stabilized once the average distance between pseudo–sites was decreased to 5km. This supports the claim that our estimates from our model are close to those of the joint triple model with a LGCP for the site–selection process, a Bernoulli likelihood for the site–retention process, and a Gaussian process for the observation process.

The correct placement of the zeros in the site–selection process is vital for the asymptotic convergence of the pseudo-likelihood to the LGCP. In particular, the asymptotics of the conditional logistic regression approximation used in our example with the logit link are only established when the zeros are either a realisation of a homogeneous Poisson point process, independent of the monitoring site locations [Baddeley et al., 2015], or when they are placed uniformly throughout the domain Ω [Warton et al., 2010]. In either case, the density of the zeros must be uniform (at least in probability) throughout Ω for each year $j \in \{1, ..., 31\}$ and be placed independent

from the observation locations.

A direct consequence of this is that for our site–selection process, we should not consider for selection at time j the subset of observed sites (i.e. the subset of Population 1) that are offline at year j (i.e. $S^C_{t_j}$). Put differently, we should not include the $R_{i,j}$'s in Population 1 in the likelihood such that $r_{i,j} = 0$. Erroneously doing so would lead to an increased density of zeros in the heavily sampled regions and thus a 'preferential sample' of zeros. Similarly, for the site–retention process at time j, we should only consider the sites online at the previous time $j-1$.

Putting these two processes together, the only zeros that should contribute to the joint Bernoulli likelihood at time j are the pseudo–sites and the sites that were online at the previous time $j-1$ and were removed from the network at time j. In fact, we tested the sensitivity of the results to the above, re–fitting the model once by following the advice given above, and again but ignoring the advice and considering all the observed sites (operational and offline) for selection at each time step j, along with the pseudo–sites. Despite the former being more appropriate, we found no differences in estimates, but we required a higher density of pseudo–sites, and hence an increased computational cost to reduce this bias in the parameter estimates. This advice is therefore of most importance for the modelling of very large datasets where the number of unique observed site locations through time could be much higher than seen here.

To form all of our predictions and maps, we simulated 1000 MCMC samples of all the parameters and latent effects from the fitted models. This feature is available in the R-INLA package [Lindgren et al., 2011b, Rue et al., 2009, 2017], by simply saving all the configuration settings generated by the software required to fit the model. We then formed all the site-specific trajectories by appropriately combining all latent effects and parameters in the linear predictor. We computed the mean, the empirical upper 97.5% and empirical lower 2.5% values of the 1000 linear predictor estimates to form our point estimates and credible intervals. Finally, to obtain the map of the pointwise expectations of the predictive distribution across GB, we used the MCMC samples of the latent effects and parameters (minus the IID site-

specific effects) and linearly interpolated the estimated field throughout Ω on a regular lattice grid covering the the map of GB, before taking the empirical mean and standard deviation across the 1000 maps. To compute the average BS across the Whole GB, we computed the mean (averaging across the pixels) of each the 1000 sampled/realized maps. Then, we computed the mean, the empirical 2.5% and the empirical 97.5% values of these 1000 (mean) values.

A.1.3 Posterior pointwise mean and pointwise standard deviation plots

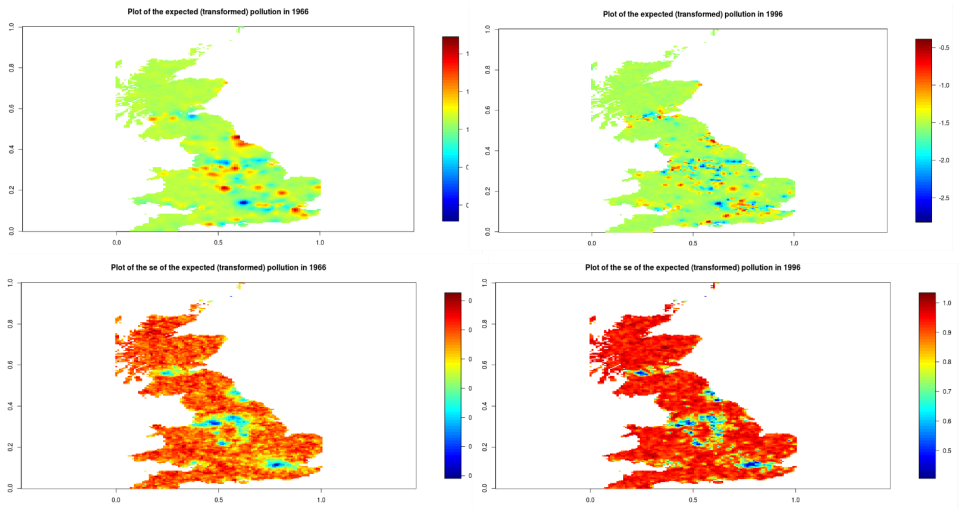

Figure A.1: A plot of the posterior mean black smoke in 1966 and 1996 under Implementation 1 with corresponding standard errors plotted below. Note that for visualisation purposes, the two plots have had their values scaled to put them on the same colour scale.

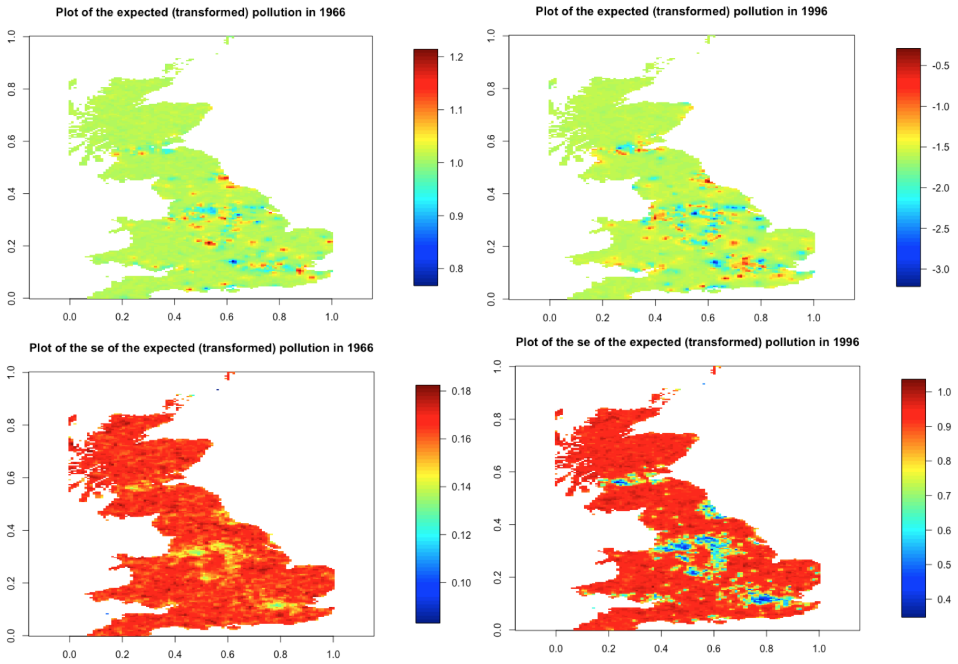

Figure A.2: A plot of the posterior mean black smoke in 1966 and 1996 with corresponding standard errors plotted below. Estimates are taken from Implementation 2. Note that for visualisation purposes, the two plots have had their values scaled to put them on the same colour scale.

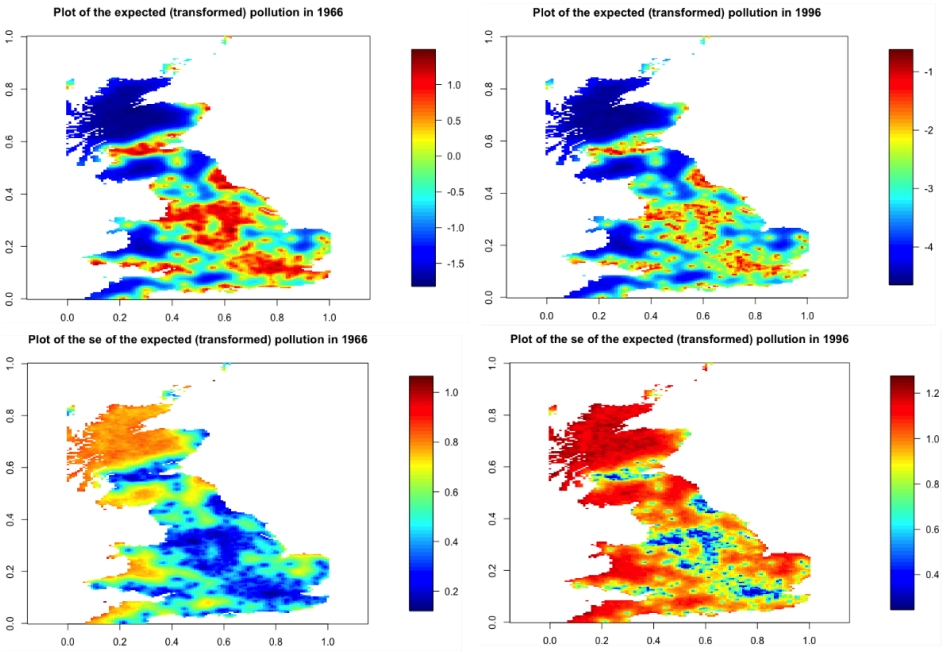

Figure A.3: A plot of the posterior mean black smoke in 1966 and 1996 with corresponding standard errors plotted below. Estimates are taken from Implementation 3. Note that for visualisation purposes, the two plots have had their values scaled to put them on the same colour scale. Notice the large drop in estimated BS in throughout much of the region.

A.1.4 Additional plot of the exceedance of the annual black smoke EU guide value

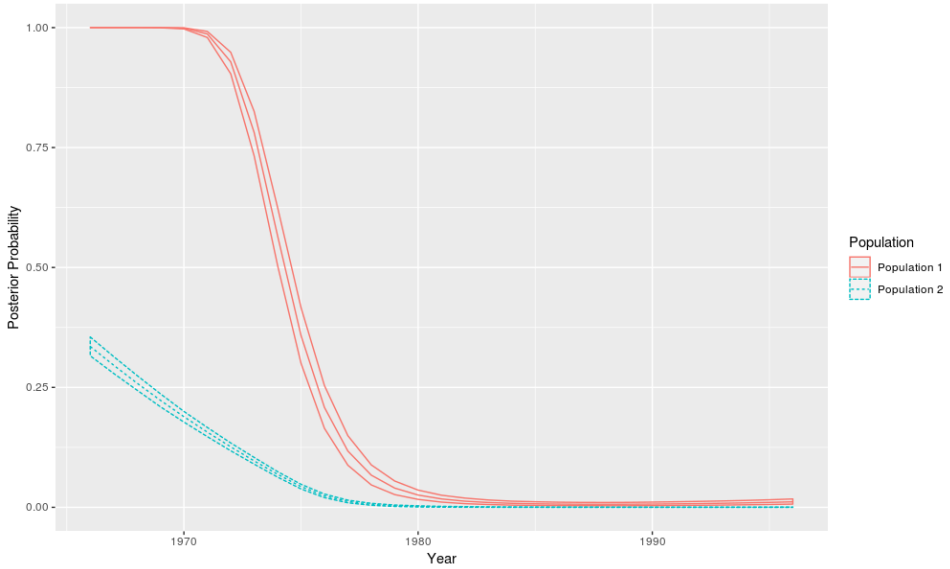

Figure A.4: A plot showing the posterior proportion of the total surface area of Great Britain with annual average black smoke level exceeding the EU guide value of $34\mu\text{gm}^{-3}$. Shown are the results from Implementation 2 (the red solid line) and Implementation 3 (the blue dashed line). Note that the line for Implementation 1 is almost identical to that from Implementation 1 and omitted.

A.1.5 Additional plot of annual average black smoke levels

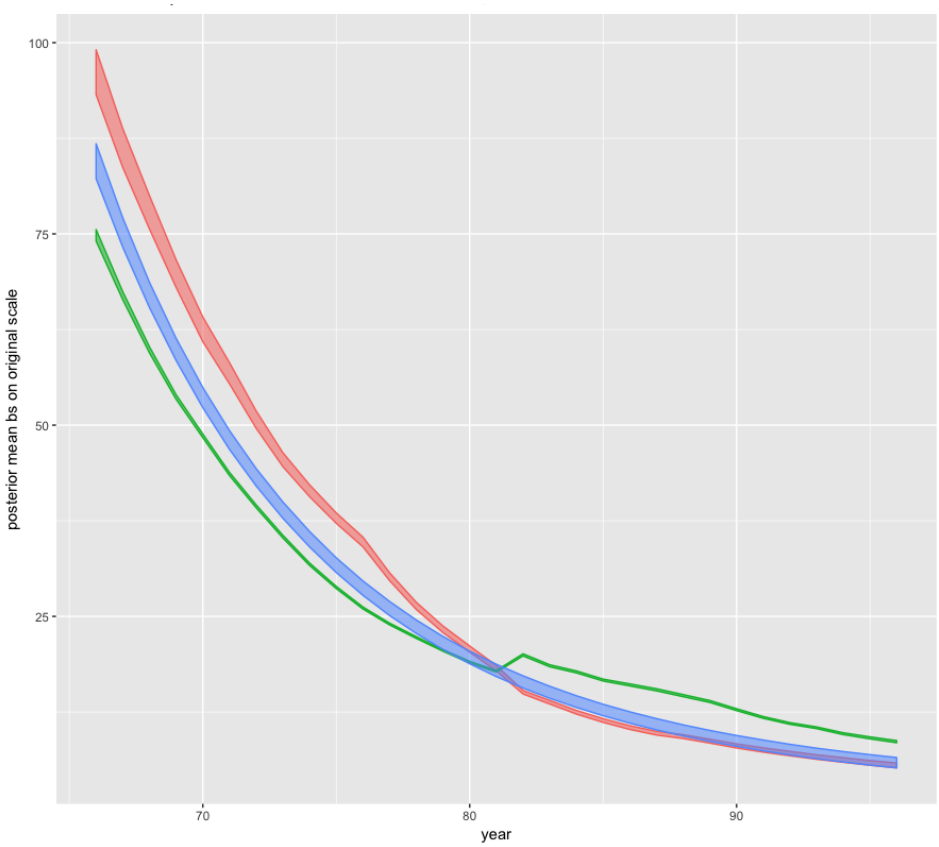

Figure A.5: Implementation 2. In green are the model–averaged BS levels averaged over sites that were selected in $\mathcal{P}_1$ (i.e. operational) at time t. In contrast, those in red are the model–averaged BS levels averaged over sites that were not selected in $\mathcal{P}_1$ (i.e. offline) at time t. Finally, in blue are the model–averaged BS levels averaged across Great Britain. Also included with the posterior mean values are their 95% posterior credible intervals. If printed in black-and-white, the green band is initially the lower line, the red band is the upper line and the blue band is initially the middle line.

A.1.6 Model diagnostic plots

We include, for each of the three implementations considered in this Chapter, residual plots to help diagnose poor model fit. Included are residuals vs. year plots, with a fitted lowess smoother to help show that the choice of a quadratic model adequately captured the temporal trend in the data. Also shown are normal QQ-plots of the residuals with fitted 99% confidence bands around the overlain QQ-line. Residuals are computed with respect to the posterior mean values. It is clear from this plot that a heavier tailed distribution on the response would have been more suitable. Finally, we include histograms and normal QQ-plots of the random effects. Here we see slightly left-skewed and right-skewed empirical marginal distributions for the random intercepts and slopes respectively. We have no strong cause for concern with these final plots.

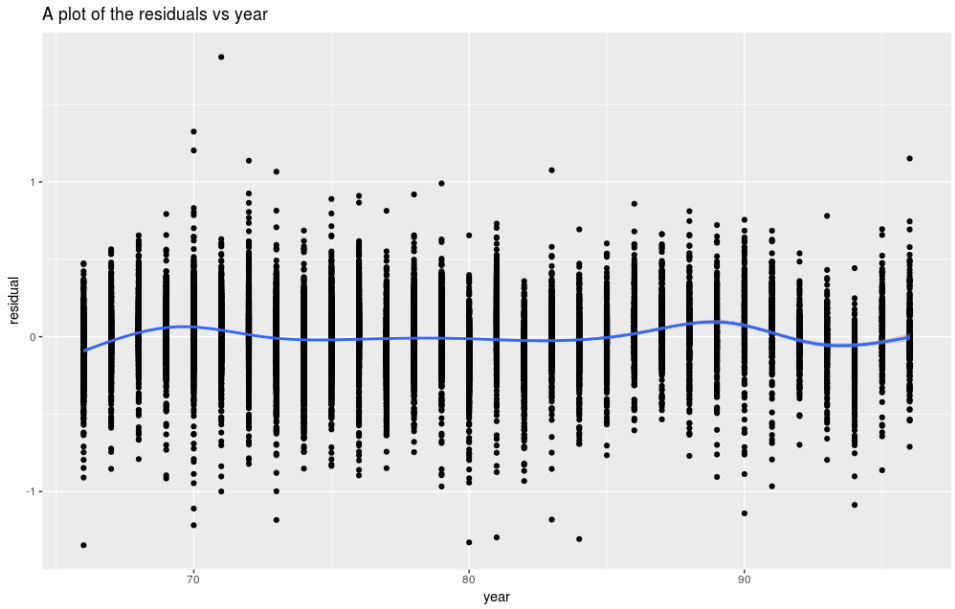

Figure A.6: A plot of the residuals vs. year from Implementation 1 with a fitted smoother.

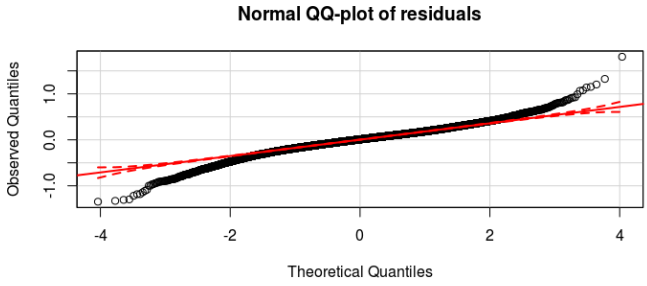

Figure A.7: A Normal Q–Q plot of the residuals from Implementation 1.

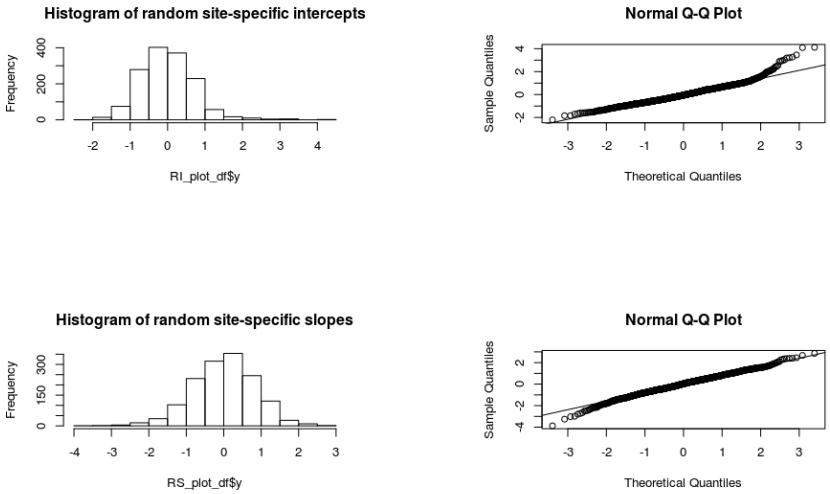

Figure A.8: Histograms of the spatially–uncorrelated random intercepts (top left) and slopes(bottom left), with corresponding Normal Q–Q plots shown on the right from Implementation 1.

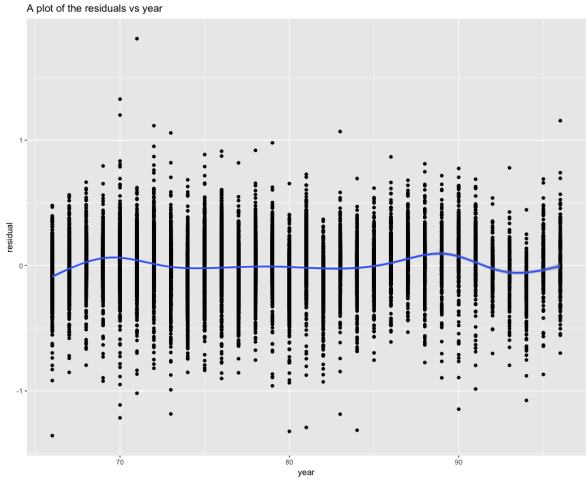

Figure A.9: A plot of the residuals vs. year for Implementation 2, with a fitted smoother.

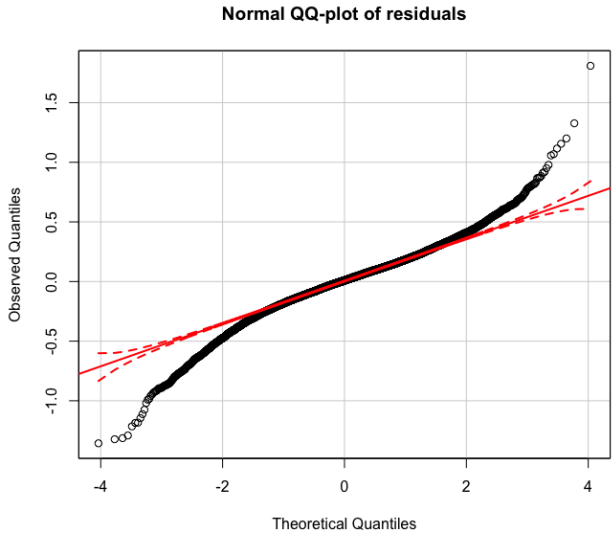

Figure A.10: A Normal Q–Q plot of the residuals from Implementation 2, with 95% confidence intervals shown in red.

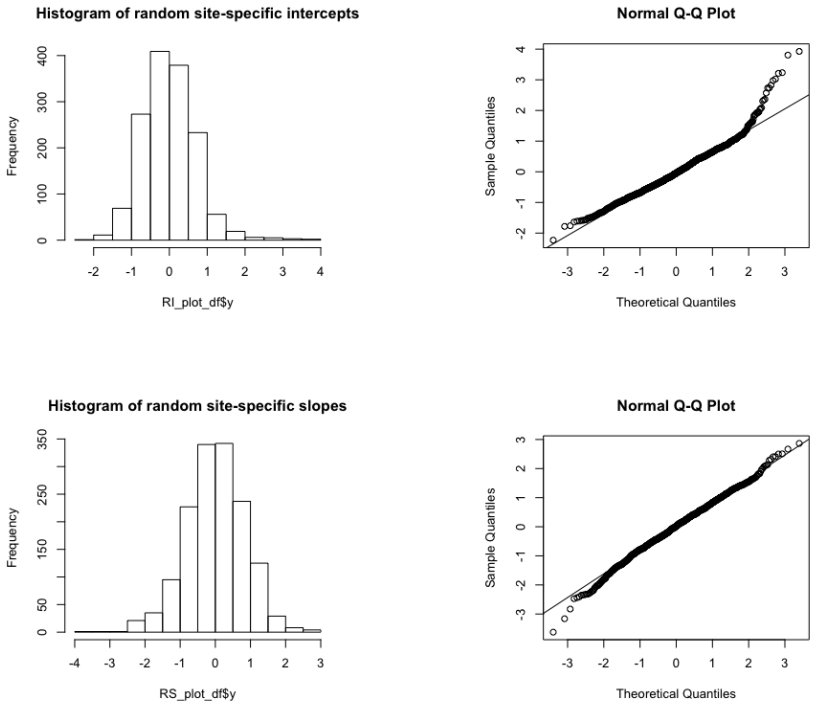

Figure A.11: Histograms of the spatially–uncorrelated random intercepts (top left) and slopes(bottom left), with corresponding Normal Q–Q plots shown on the right from Implementation 2.

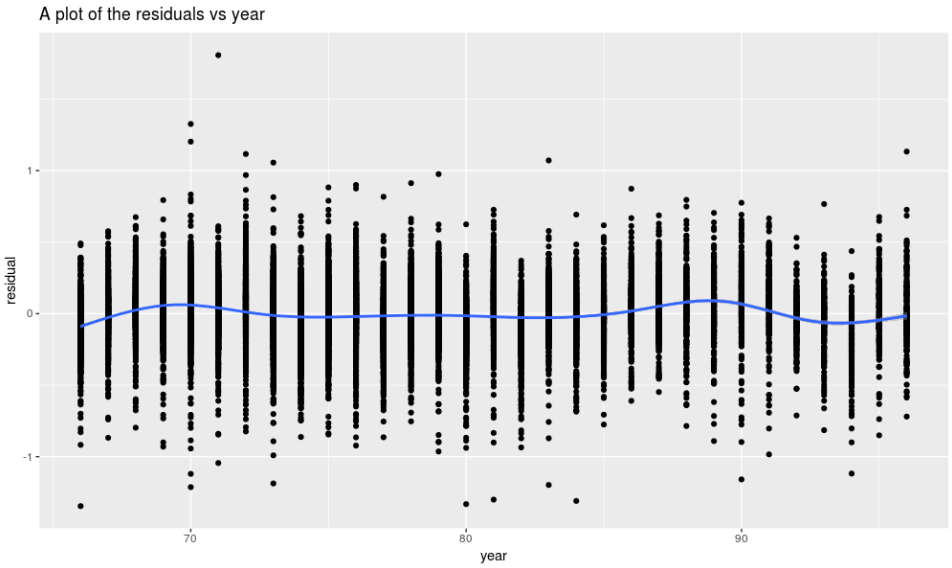

Figure A.12: A plot of the residuals vs. year for Implementation 3 with a fitted smoother.

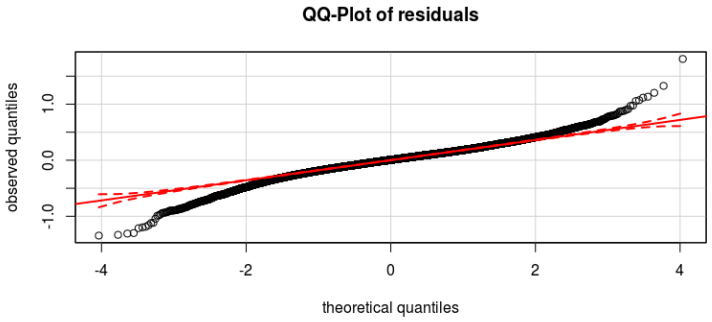

Figure A.13: A Normal Q-Q plot of the residuals from Implementation 3 with 95% confidence intervals shown in red.

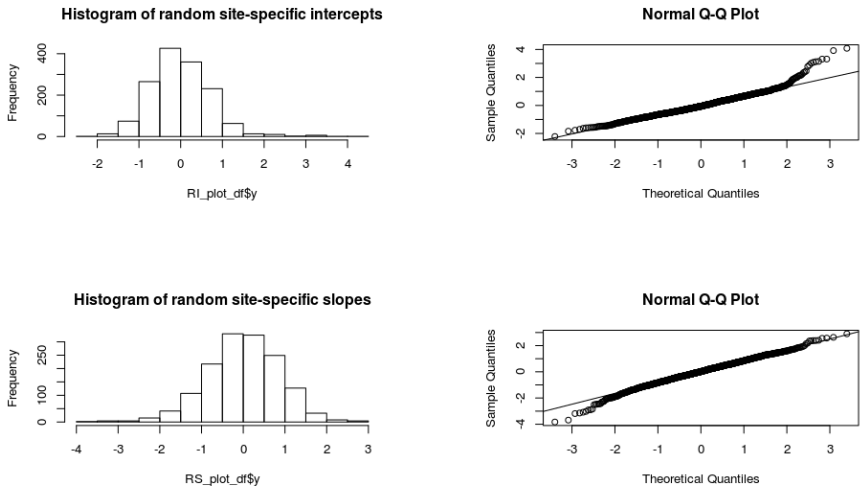

Figure A.14: Histograms of the spatially-uncorrelated random intercepts (top left) and slopes(bottom right), with corresponding Normal Q-Q plots shown on the right from Implementation 3.

A.2 Chapter 4 Supporting Materials

A.2.1 More details on the simulation study

For computational speed-ups, the R-INLA package is used for both the simulation and estimation of $Z(\mathbf{s})$ [Lindgren et al., 2011b, 2015, Rue et al., 2009]. A high resolution triangulation mesh (triangle lengths of 0.01) is defined for the SPDE approximation over the unit square Ω, and linear interpolation is used to impute the values at any point $\mathbf{s}$ within Ω. For the priors on the Gaussian process, PC priors [Fuglstad et al., 2018] are specified with prior probability of 0.1 that the spatial range is less than 0.1 and prior probability of 0.1 that the standard deviation is greater than 3. Gaussian errors on the responses $Y(\mathbf{s}, t)$ are added, with a weakly informative Gamma(1, 5e^{-}5) distribution placed on the precision of the error distribution. This is done

to reduce the risk of computational singularities. The NN test is performed at the 5% significance level using 19 Monte Carlo samples (i.e. $M = 19$). Each experimental setting is repeated 200 times.

Along with the NN test outlined in algorithm 1, a Monte Carlo test using the rank correlations between estimates of $Z(\mathbf{s})$ and estimates of the raw residuals of the assumed IPP under the null hypothesis is also compared. This may provide a more suitable test, when the assumed sampling mechanism is indeed a LGCP. Furthermore, such a test does not require a choice of K. To compute these residual values, an edge-corrected Gaussian kernel is used to smooth the raw residuals. The bandwidth is selected using leave-one-out cross-validation. This is performed using the *spatstat* package [Baddeley et al., 2015]. To compute the test, the NN_k values are simply replaced with the smoothed residual values, evaluated at the point locations. We refer to this test as the residual test hereafter. It is interesting to assess the relative performance of the NN test, given its generality across all point processes and to the discrete spatial setting.

First, the Type 1 error of the PS tests are assessed in the simplest setting without any covariate effects (i.e. $\alpha_1 = 0$) or PS effects (i.e. $\gamma = 0$). Results from four tests are compared. The first two are residual tests. The first computes the p-value directly using a standard permutation approach under the (false) assumption that the pairs of residuals and estimates $\hat{Z}(\mathbf{s})$ are an IID sample from some bivariate distribution. The positive spatial correlations due to the process $Z(\mathbf{s})$ violate this assumption, with the magnitude of violation increasing with the spatial range ρ_Z. The second attempts to correct for this spatial correlation. By forming realisations from the estimated sampling process, the spatial correlation in $Z(\mathbf{s})$ is accounted for. The third and fourth are NN tests. Once again, comparisons are made between the permutation-based and the Monte Carlo-based approaches.

Fig. A.15 shows the results for $n = 50$ across three increasing values of the spatial range ρ_Z and across different numbers of nearest neighbours K. It is apparent that both Monte Carlo tests attain Type 1 error at or below the 5% level. The two standard permutation tests attain a Type 1 error above the 5% level, and this increases dramatically with ρ_Z. At the highest

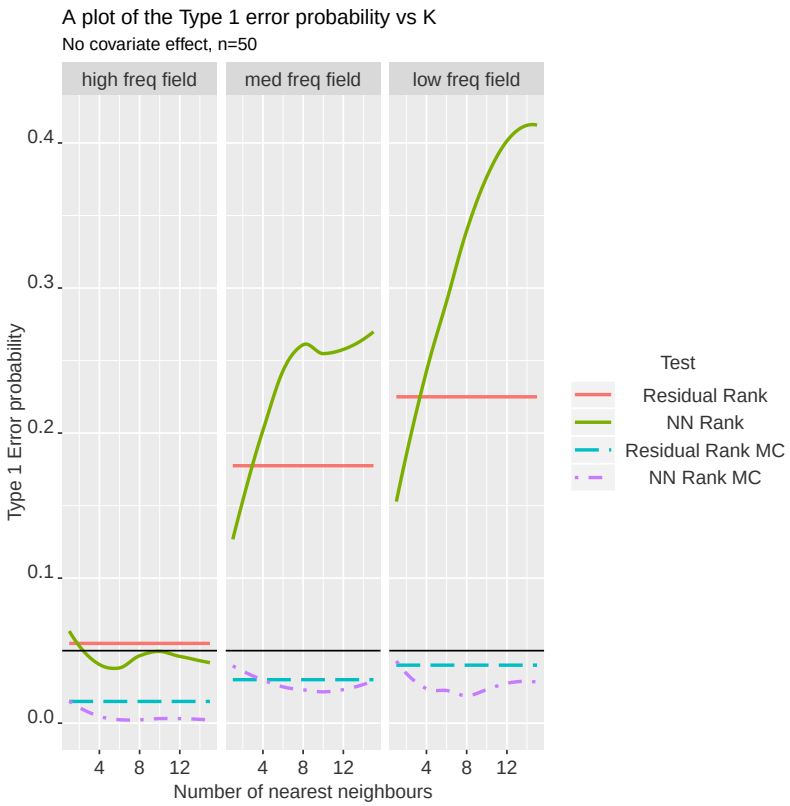

Figure A.15: A plot of the Type 1 error for four tests. The three boxes show the results for $\rho_Z \in \{0.02, 0.2, 1\}$, from left to right respectively for a sample size of 50. The two 'Residual' tests are computed using the kernel-density smoothed values of the residuals from the fitted homogeneous Poisson processes. Leave-one-out cross-validation was used to select the bandwidth. The 'NN' tests are those based on the K nearest neighbour values. The suffix 'MC' denotes the test has been computed from Monte Carlo realisations of the fitted point process.

value of $\rho_Z = 1$, equal to the length of the domain Ω, the Type 1 error can be higher than 40%. The results for the very low value of $\rho_Z = 0.02$, demonstrate that the type 1 error approaches the nominal 5% level when the spatial correlation approaches zero. This is due to the IID assumption

becoming more reasonable as the $Z(\mathbf{s})$ tends towards Gaussian white noise. When $\rho_Z = 0.02$, the prior distribution on the range parameter would reflect the case where a researcher incorrectly assumed spatially smooth data prior to the model-fitting. Fig. A.19 shows the results for $n = 100$. It is apparent that the Type 1 error increases with sample size for the permutation tests, while the Monte Carlo tests remain bounded above by 0.05.

Next, the power of the two Monte Carlo tests to detect a PS effect when the alternative hypothesis is true is assessed. The behaviours of the tests are first investigated in the setting where no covariate effects exist (i.e. $\alpha_1 = 0$), but where moderate positive preferential sampling occurrs (i.e. $\gamma = 1$). All tests are performed with the two-sided alternative hypothesis, namely that h is a monotonic function of Z in either direction.

Fig. A.16 shows the results for $n = 50$, this time with spatial ranges of $\rho_Z \in \{0.2, 1\}$, again across $K \in \{1, ..., 15\}$. The power results for $\rho_Z = 0.02$ are omitted in Figure A.17, since the power is consistently small (<0.1) for both. This demonstrates the need for the $Z(\mathbf{s}, t)$ term to be spatially-smooth for the test to detect PS. It is clear that the power of the NN test is sensitive to the choice of K value, especially for smaller sample sizes. Interestingly the optimum power achieved by the NN test with respect to K depends upon both the spatial range of Z and the sample size. The higher the spatial range of Z, and hence the smoother it is, the greater the value of K that is required to optimize the power. The optimum choice of K also increases with the sample size, since the number of realized points per cluster increases. For example, when $\rho_Z = 0.2$ and the sample size is 50, the test is optimized when $K = 1$. This increases to $K = 5$ when $\rho_Z = 1$ and the sample size is 100. Finally, for $n = 50$ it appears that the NN tests have a slightly lower power than the residual measure-based test.

Next, the spatial range is fixed to be very small ($\rho_Z = 0.02$), and the magnitude of PS is fixed to be very high ($\gamma = 2$). This set-up leads to very small clusters to form when $\gamma \neq 0$. The joint effects of sample size and K on the power of the NN test to detect PS is then demonstrated. Additionally, the power of the NN test is compared to the residual test. Three plots are shown to present the power vs. K in Fig. A.17. From left to right, these

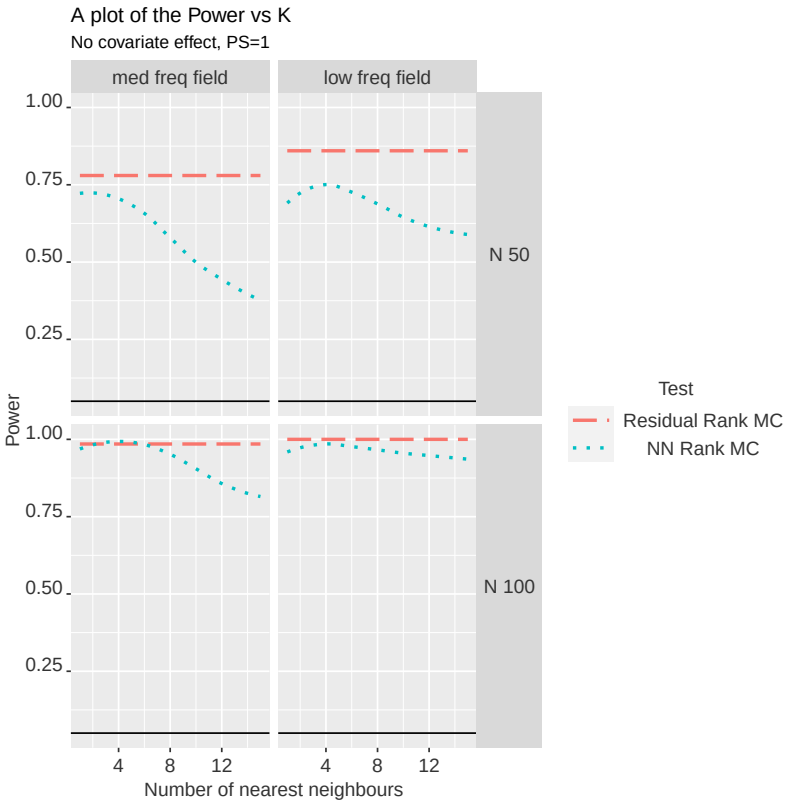

Figure A.16: A plot of the Power for two tests when the PS parameter γ equals 1. The two columns show the results for $\rho_Z \in 0.2, 1$, from left to right respectively. The two rows show results for a sample size of 50 and 100 respectively. The 'Residual' tests are computed using the kernel-density smoothed values of the residuals from the fitted homogeneous Poisson processes. Leave-one-out cross-validation was used to select the bandwidth. The 'NN' test is based on the K nearest neighbour values. The suffix 'MC' denotes the test has been computed from Monte Carlo realisations of the fitted point process.

show sample sizes of $50, 100$ and 250. For small sample sizes ($n \in \{50, 100\}$), both tests have low power to detect PS as expected. Interestingly however, the NN test outperforms the smoothed residual test for all three sample

sizes, achieving maximum powers of 0.21, 0.65 and 1 at $K = 1$ compared with 0.14, 0.40 and 0.97 for the residual test. Furthermore, the power of the NN test attains it maximum at $K = 1$, before dramatically diminishing to 0 as K increases. Fig. A.20 shows the equivalent plots for $\rho_Z = 0.2$. Here, the NN test is no longer more powerful, with the residual test performing better at $n = 50$. Note that the performance of the residual test may improve with a different choice of bandwidth-selection method.

Results are now presented for the case when a unique covariate effect exists for the sampling process. The magnitudes of the covariate effect and the PS effect are both set to 1 (thus $\alpha_1 = \gamma = 1$). The spatial range of the covariate effect is varied ($\rho_w \in \{1, 0.02\}$). The results on the power of three different tests to detect PS are shown. As before, the first two are the kernel-smoothed residual and NN rank tests. The third test is a rank test using kernel-smoothed estimated residuals, but this time using residuals computed from an incorrectly specified point process fit to the points. This is chosen to be a homogeneous Poisson process (HPP hereafter). Note that the Monte Carlo realized points S_t^m still come from the null IPP, fitted to the original observations S_t. Thus the Monte Carlo sampled realisations still come from the correct data-generating mechanism (correct up to parameter estimation error). Unlike the residuals from the first test, these residuals do not adjust for the covariate effect. The purpose of this comparison is to see if any improvements in the power of the test can be attained by considering computed quantities that directly adjust for any covariate effects.

The spatial range of the covariate field is changed for the following reason. When the spatial ranges of both the covariate field $\mathbf{w}(\mathbf{s})$ and the underlying spatial field $Z(\mathbf{s})$ are large and similar, then the magnitude of the empirical correlation of a single realisation of the two fields may be high. This is despite their realisations arising from independent distributions [Hanks et al., 2015]. A possible consequence of this is that the tests may be unable to distinguish between clustering due to an unknown process Z, and clustering due to the measured covariate $\mathbf{w}(\mathbf{s})$. This may affect the ability of tests to detect preferential sampling, when their computed quantities are not adjusted for the effects of covariates. The rank test of the residuals from

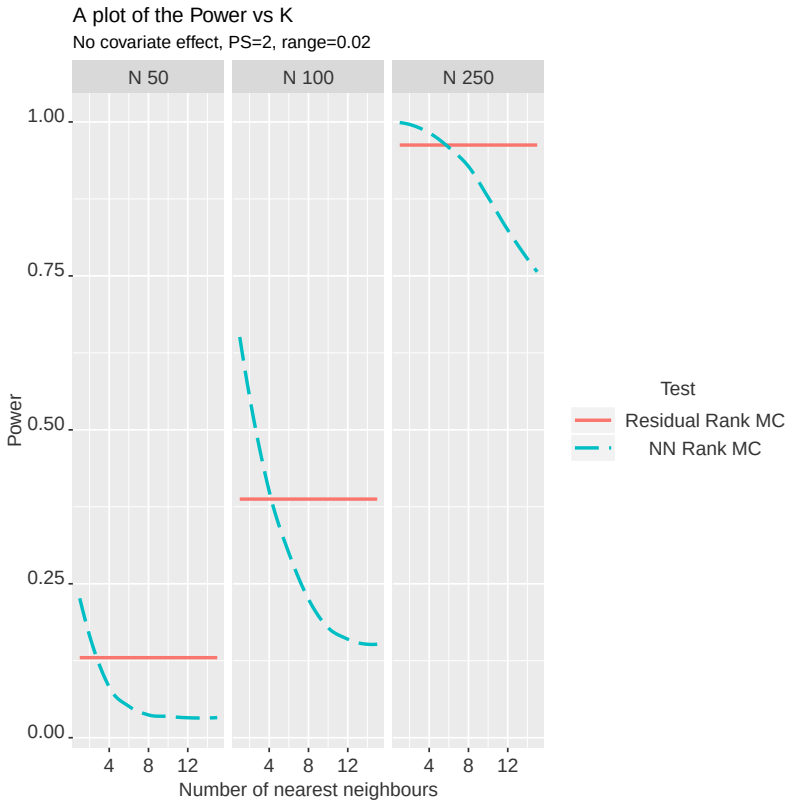

Figure A.17: A plot of the Power for two tests when the PS parameter γ equals 2 and $\rho_Z = 0.02$. The three columns show the results for the sample sizes 50, 100 and 250 from left to right respectively. The 'Residual' tests are computed from the kernel-density smoothed values of the residuals from the fitted homogeneous Poisson processes. Leave-one-out cross-validation was used to select the bandwidth. The 'NN' test is based on the K nearest neighbour values. The suffix 'MC' denotes the test has been computed from Monte Carlo realisations of the fitted point process.

the correctly specified IPP model is the only test that directly adjusts for the covariate effects.

Fig A.18 presents a complex interaction of different factors. When the covariate field has very low spatial range (i.e. $\rho_w = 0.02$), and hence has high

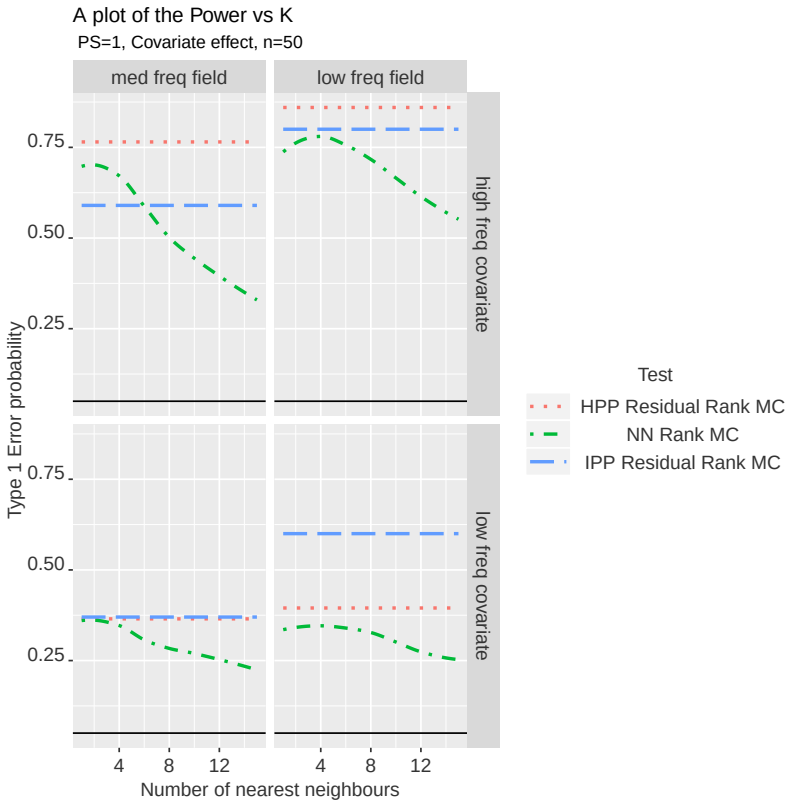

Figure A.18: A plot of the Power for two tests when the PS parameter γ equals 1, the covariate effect α_1 equals 1 and when the sample size is 50. The two columns show the results for the spatial range $\rho_Z \in \{0.2, 1\}$ from left to right respectively. The two rows show the results for the spatial range of the covariate $\rho_w \in \{0.02, 1\}$ from top to bottom respectively. The two 'Residual' tests are computed from the kernel-density smoothed values of the raw residuals from the fitted Homogeneous (HPP) and Inhomogeneous Poisson processes (IPP). Leave-one-out cross-validation was used to select the bandwidth. The 'NN' test is based on the K nearest neighbour values. The suffix 'MC' denotes the test has been computed from Monte Carlo realisations of the fitted point process.

frequency, negligible correlation can exist between $Z(\mathbf{s})$ and $w(\mathbf{s})$. Consequently, no gains in power are seen when tests use covariate-adjusted mea-

sures of clustering relative to when they use unadjusted measures. However, when both the covariate w and underlying field Z are very smooth (i.e. $\rho_w = 1, \rho_Z = 1$), large increases in power are seen with the covariate-adjusted measure of clustering. The power increases to 0.57 compared with 0.40 and 0.33 for the HPP residual and the NN methods respectively. The results are similar for $n = 250$ (see Fig. A.22 in the supplementary material). In conclusion then, in cases where the spatial ranges of informative covariates are large and similar in size to the underlying $Z(\mathbf{s})$, computed quantities other than NN_k should be considered to improve the power to detect PS.

Finally, the performance of the tests is assessed in settings where the response is non-Gaussian, and when the true sampling process is not an IPP. The $Y(\mathbf{s})$ values now take the form of counts and (13) is replaced with a Poisson distribution. The log-transformed mean at location $\mathbf{s} \in \Omega$ is set equal to the random field $Z(\mathbf{s})$ plus a constant intercept of 2. The intercept is chosen to ensure non-zero counts occur often. The true sampling process is set equal to a Hardcore process with two different radii of interactions, denoted R, compared (0.025 and 0.05). Under these two processes, points within S_t cannot be sampled closer than a distance apart of 0.025 or 0.05 respectively. With $n = 100$, these two constraints enforce moderate and strict levels of regularization of the points respectively, violating the IPP assumption of no inter-point interaction.

The Hardcore process is chosen to highlight the fact that the choice of nearest-neighbour distances to capture additional clustering will be poor in some settings. Here, due to the nearest neighbour distances being lower bounded, the contrast in their observed values will decrease as R is increased. Estimates of the smoothed residuals are not directly affected by an increase in R and hence the residual test is expected to far outperform the NN test as R increases.

After sampling the data, the tests under two scenarios are compared. The first considers the case where the researcher assumes the correct Hardcore process sampling mechanism for the Monte Carlo realisations of S_t. The second considers the case when the researcher misspecifies it as an IPP (i.e. assumes no inter-point interaction).

The results from this simulation study, repeated 100 times, are shown in Fig. A.21. As expected, the residual tests far outperform the NN test when the radius of interaction is 0.05. This is due to the lack of contrast in the nearest neighbour distances that leads to a reduction in the power of the NN test. When the radius of interaction is 0.025, the contrast in the nearest neighbour distances is restored and both methods perform very well again. In this case, the power exceeds 0.95 when the correct Hard Core process is fitted. Interestingly, the residual test that use the raw residuals from the correctly specified Hard Core processes perform no better than the residual test that use raw residuals of the incorrectly specified HPP. On the other hand, the performance of the NN test improves when the class of point process is correctly specified.

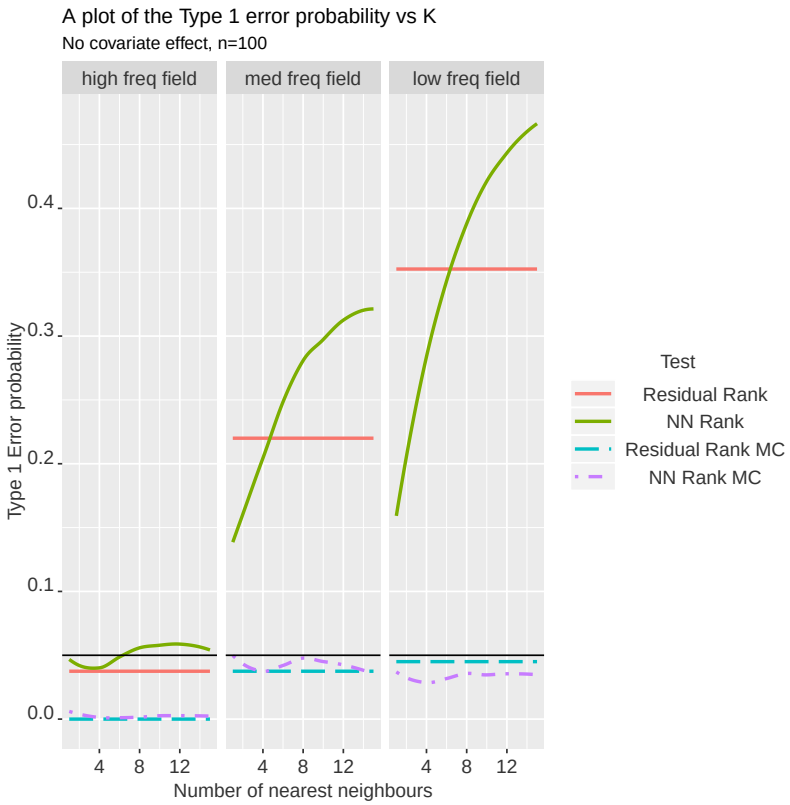

Figure A.19: A plot of the Type 1 error for four tests. The three boxes show the results for $\rho_Z \in 0.02, 0.2, 1$, from left to right respectively for a sample size of 100. The 'Residual' tests are computed from the kernel-density smoothed values of the residuals from the fitted homogeneous Poisson processes. Leave-one-out cross-validation was used to select the bandwidth. The 'NN' tests are those based on the K nearest neighbour values. The suffix 'MC' denotes the test has been computed from Monte Carlo realisations of the fitted point process.

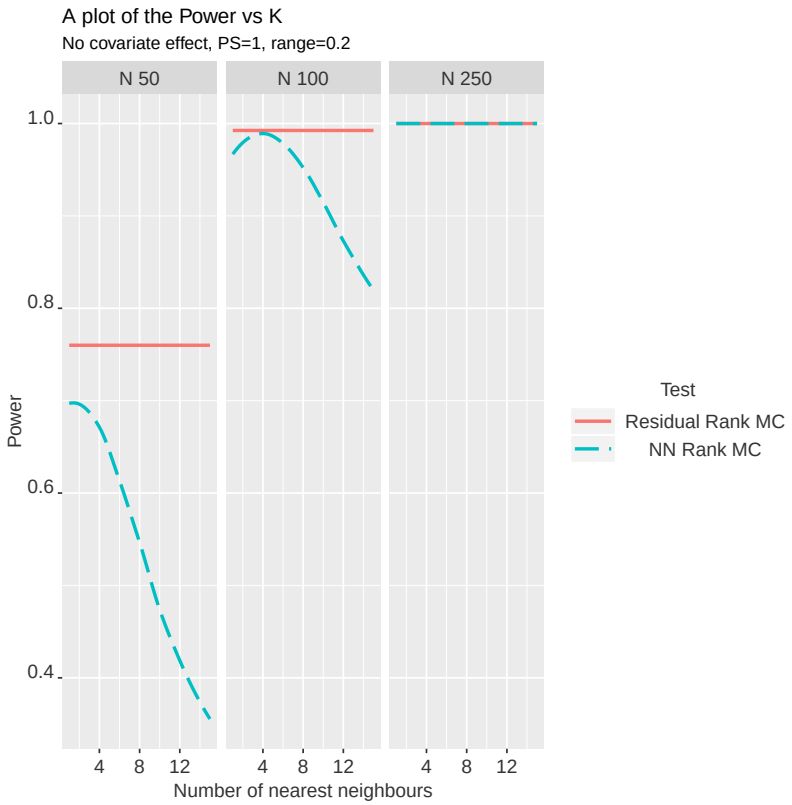

Figure A.20: A plot of the Power for two tests when the PS parameter γ equals 1 and $\rho_Z = 0.2$. The three columns show the results for the sample sizes 50, 100 and 250 from left to right respectively. The two rows show results for a sample size of 50 and 100 respectively. The 'Residual' test denotes the kernel-density smoothed values of the raw residuals from the homogeneous Poisson process. Leave-one-out cross-validation was used to select the bandwidth. The 'NN' tests are those based on the K nearest neighbour values. The suffix 'MC' denotes the test has been computed from Monte Carlo realisations of the fitted point process.

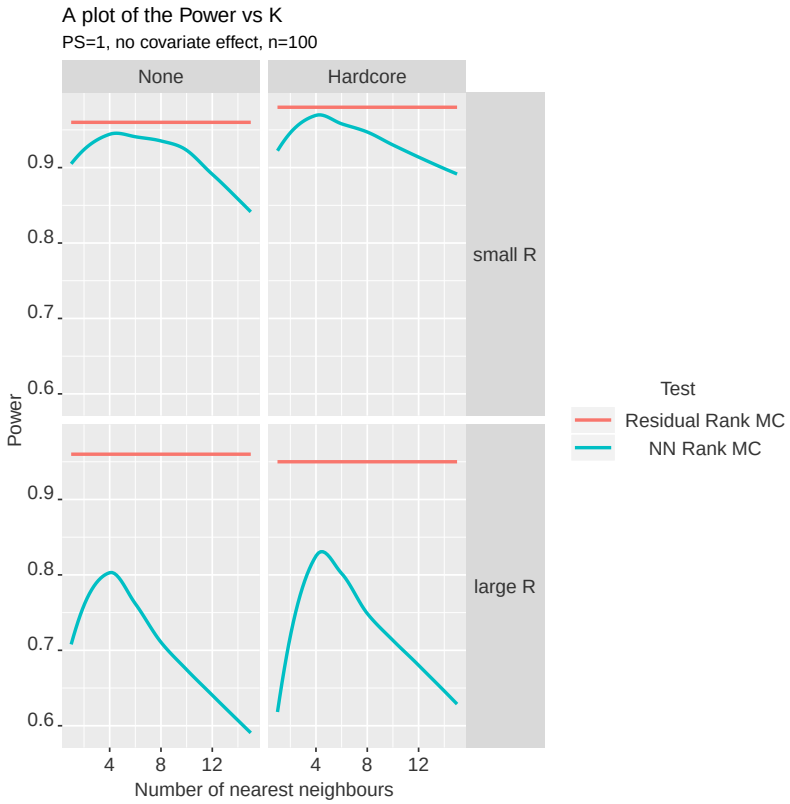

Figure A.21: A plot of the Power for two tests when the true sampling process is a Hard Core point process, with PS parameter γ equals 1 and $\rho_Z = 1$. The two columns show the results when a Poisson process model ('None'), and when a Hard Core process are fitted and then used for Monte Carlo sampling. From top to bottom, the rows denote the case where the true radius of interaction for the Hard Core process equals 0.025 and 0.05. The 'Residual' test is computed using kernel-density smoothed values of the raw residuals from the fitted point process. Leave-one-out cross-validation was used to select the bandwidth. The 'NN' tests are those based on the K nearest neighbour values. The suffix 'MC' denotes the test has been computed from Monte Carlo realisations of the fitted point processes.

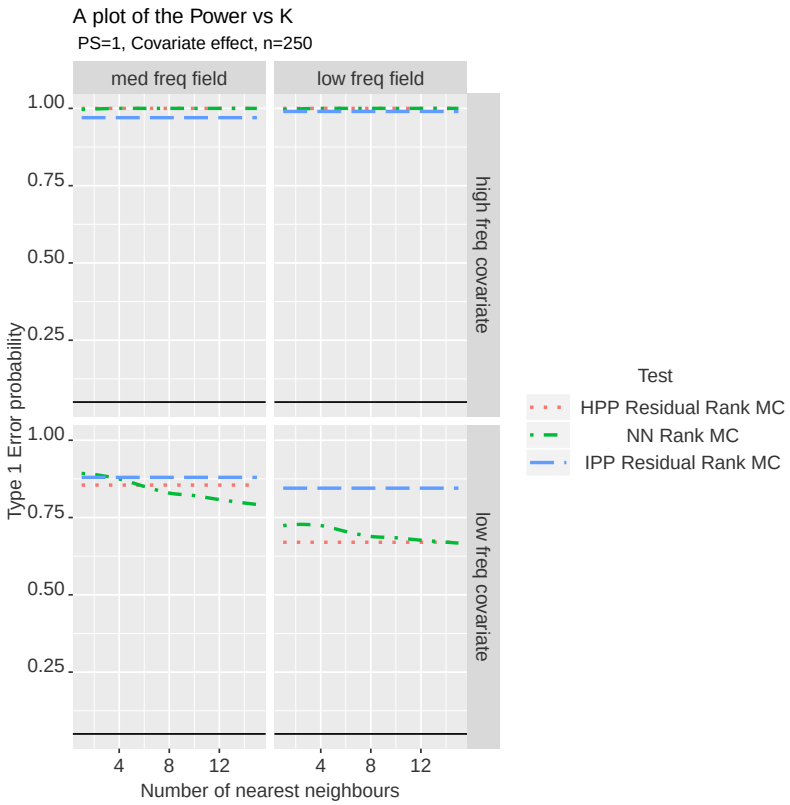

Figure A.22: A plot of the Power for two tests when the PS parameter γ equals 1, the covariate effect α_1 equals 1 and when the sample size is 250. The two columns show the results for the spatial range $\rho_Z \in \{0.2, 1\}$ from left to right respectively. The two rows show the results for the spatial range of the covariate $\rho_w \in \{0.02, 1\}$ from top to bottom respectively. The 'Residual' test is computed using kernel-density smoothed values of the raw residuals from the fitted Homogeneous and Inhomogeneous Poisson processes. Leave-one-out cross-validation was used to select the bandwidth. The 'NN' test is based on the K nearest neighbour values. The suffix 'MC' denotes the test has been computed from Monte Carlo realisations of the fitted point process.

A.3 Chapter 5 Supporting Materials

A.3.1 Additional theory on marked point processes

Start with a Poisson process Y on Ω with intensity $\lambda(\mathbf{s})$. Next, take a probability distribution $p(\mathbf{s},.)$ on M depending on $\mathbf{s} \in \Omega$ such that, for $B \subset M$, $p(.,B)$ is a measurable function on Ω. A marking of Y is a random subset of $\Omega \times M$ such that the projection onto Ω is Y and such that the conditional distribution of $Y^\star$, given Y makes the marks $m_y : \mathbf{y} \in Y$ independent with respective distributions $p(\mathbf{y},.)$. We now have the following theorems (adapted from Kingman [1994]):

Theorem A.1 (Marking Theorem) *The random subset $C \subset Y^\star$ is a Poisson process on $\Omega \times M$ with mean measure $\Lambda^\star$ defined:*

$$\Lambda^\star(C) = \int\int_{(\boldsymbol{s},m)\in C} \lambda(\boldsymbol{s}) p(\boldsymbol{s}, dm) d\boldsymbol{s} \tag{A.1}$$

Theorem A.2 (Mapping Theorem) *If the points (Y, m_Y) form a Poisson process on $\Omega \times M$, then the marks form a Poisson process on M and the mean measure is obtained by setting $C = \Omega \times B$ in (A.1):*

$$\mu_m(B) = \int_\Omega \int_B \lambda(\boldsymbol{s}) p(\boldsymbol{s}, dm) d\boldsymbol{s} \tag{A.2}$$

if the marks take on only K different values, then the theorem specializes for the i^{th} mark to:

$$\Lambda_i(A) = \int_A \lambda(\boldsymbol{s}) p(\boldsymbol{s}, \{m_i\}) d\boldsymbol{s} \qquad A \subset \Omega \tag{A.3}$$

A.3.2 Extra results of the main simulation study

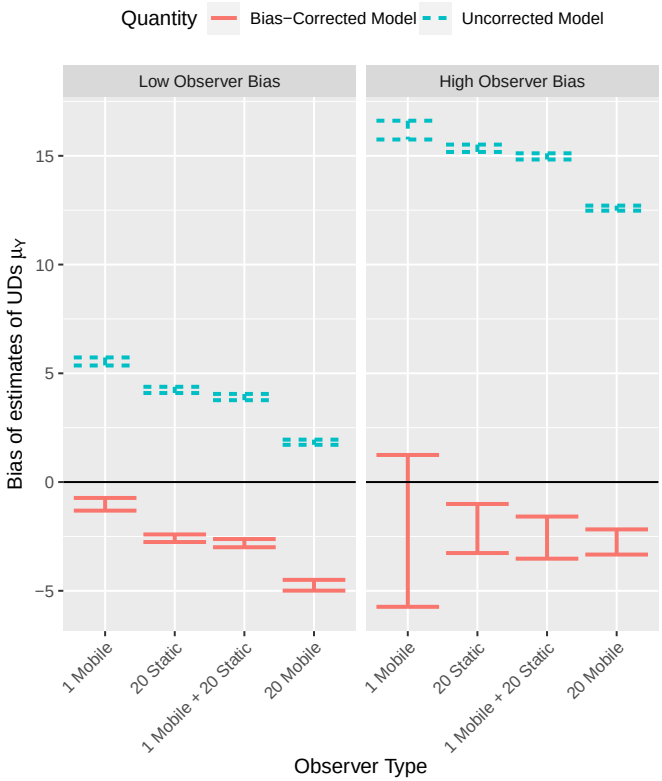

Figure A.23: A plot showing the bias of the estimated y-coordinate of the animal's UD center μ_y under the bias-corrected and bias-uncorrected models vs the types of observers. From left to right are the results from one mobile observer, twenty static observers, twenty static with one mobile observers, and twenty mobile observers. The degree of observer bias is changed from low to high in the columns. The red solid lines and the blue dashed lines show the median bias along with robust intervals computed as $\pm 2c$MAD from the bias-corrected and uncorrected models across the 100 simulation replicates respectively. The MAD has been scaled by $c = 1.48$. This ensures that the intervals are asymptotically equivalent to the 95% confidence intervals that would be computed if the biases were normally distributed. Note that here all the analyst's assumptions correctly match the true data-generating mechanism, albeit with any overlap in the observers' efforts ignored. Note the large reduction in Bias offered by the bias-corrected method in the 'High Observer Bias' setting.

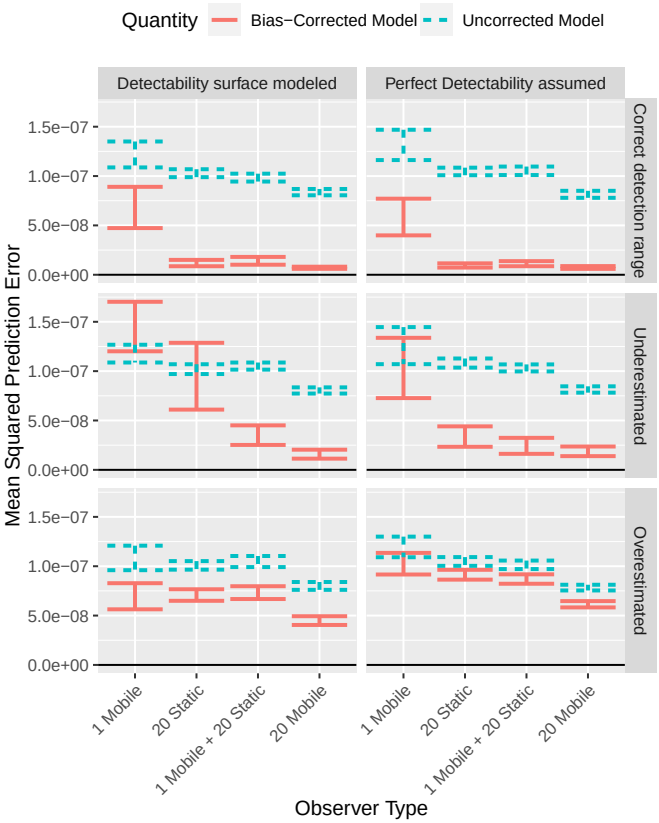

Figure A.24: A plot showing the mean squared prediction error (MSPE) of the estimated animal's UD under the bias-corrected and bias-uncorrected models vs the types of observers. From left to right are the results from one mobile observer, twenty static observers, twenty static with one mobile observers, and twenty mobile observers. The distance sampling function has either been modeled or ignored in the two columns from left to right and the observers' detection range has been assumed to be 10, 2 and 50 across the rows. The red solid lines and the blue dashed lines show the median MSPE along with robust intervals computed as $\pm 2c\text{MAD}$ from the bias-corrected and uncorrected models across the 100 simulation replicates respectively. The MAD has been scaled by $c = 1.48$. This ensures that the intervals are asymptotically equivalent to the 95% confidence intervals that would be computed if the MSPE values were normally distributed. The results are shown for 150 trips with high observer bias.

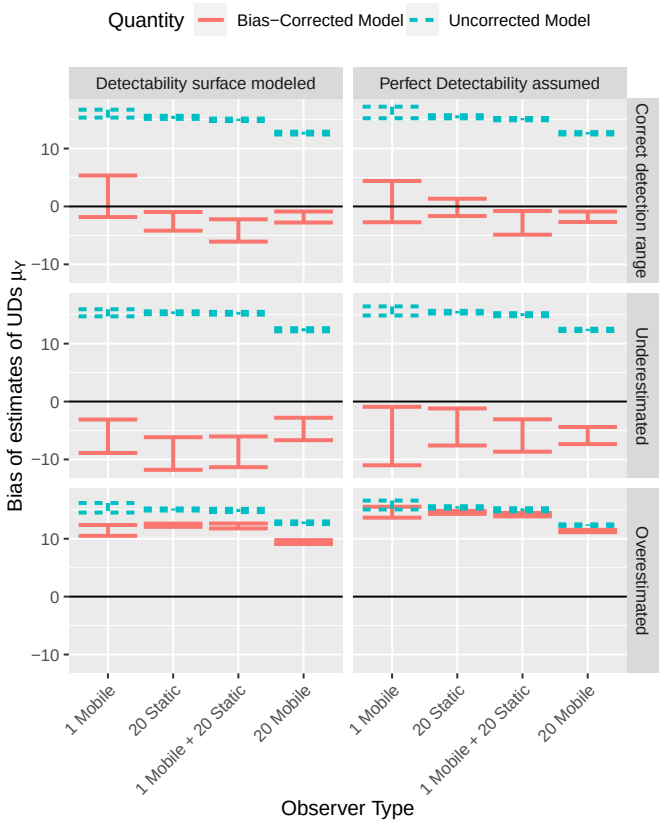

Figure A.25: A plot showing the bias of the estimated animal's UD center μ_y under the bias-corrected and bias-uncorrected models vs the types of observers. From left to right are the results from one mobile observer, twenty static observers, twenty static with one mobile observers, and twenty mobile observers. The distance sampling function has either been modeled or ignored in the two columns from left to right and the observers' detection range has been assumed to be 10, 2 and 50 across the rows. The red solid lines and the blue dashed lines show the median bias along with robust intervals computed as $\pm 2c$MAD from the bias-corrected and uncorrected models across the 100 simulation replicates respectively. The MAD has been scaled by $c = 1.48$. This ensures that the intervals are asymptotically equivalent to the 95% confidence intervals that would be computed if the biases were normally distributed. The results are shown for 150 trips with high observer bias.

A.3.3 Details of the additional simulation study

In the previous simulation study, we argued that two major sources of prediction bias were the autocorrelations between the encounter/non-encounter events, and the overlap between the observers' fields-of-view. We demonstrate these claims in a second simulation study. Unlike the previous simulation study, this one is designed to ensure that the encounter/non-encounter events at each time step are approximately independent of each other. This is achieved by increasing the average distance travelled by the animal at each discrete time step. This could also be interpreted as the setting where observers attempt encounters at discrete sampling times and wait for a sufficiently long amount of time between sampling times to reduce the autocorrelation between the encounter/non-encounter events.

We simulate the movements of observers from the same stochastic differential equation model. For the animal, we change the variance of the potential function to 1 and increase the variance of the Brownian motion terms to 400. This leads to the animal moving an average distance of 23 units, compared with 1.75 and 3.5 units for the high bias and low bias mobile observers respectively. Given that each simulation trip ends when the first encounter is made, these simulation settings ensure that the autocorrelation between the encounter/non-encounter events is greatly reduced.

Fig SA.26 clearly demonstrates that a reduction in the autocorrelation between the encounter/nonencounter events leads to a reduction in the bias of the estimated UD center. Furthermore, a large increase is witnessed in the relative MSPE of the effort-corrected approach compared with the uncorrected approach. In fact, the effort-corrected approach outperforms the uncorrected approach in all settings. However, there is some remaining bias in the estimates of the UD center from the effort-corrected approach and the magnitude of this bias increases with the numbe of observers.

To demonstrate that this bias is in fact caused by overlap in the observers' fields-of-view, we implement a method for adjusting for overlap when estimating observer effort. In particular, let $p_{det}(o, \mathbf{s}, t)$ denote the detection probability function for observer o, evaluated at space-time coordinate $(\mathbf{s}, t)$.

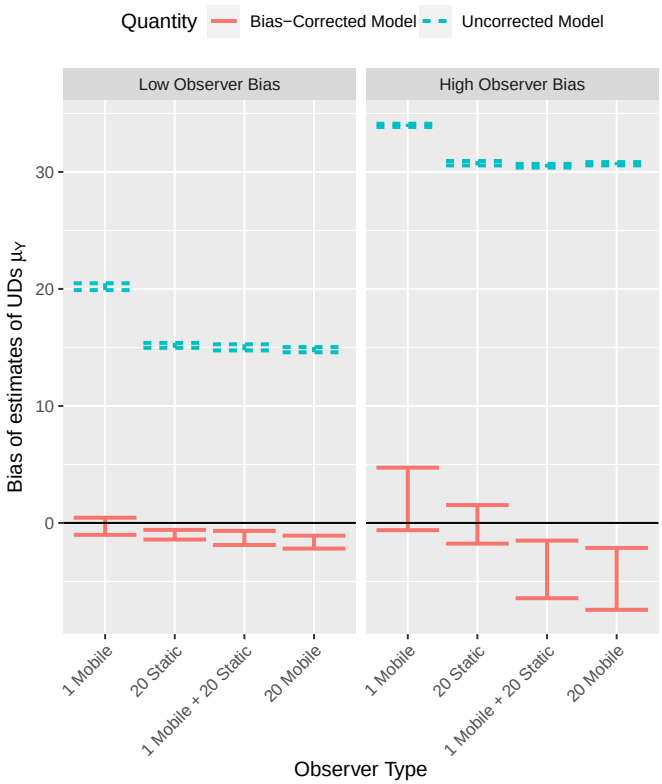

Figure A.26: A plot showing the bias of the estimated y-coordinate of the animal's UD center μ_y under the bias-corrected and bias-uncorrected models vs the types of observers. The results shown here are for the second simulation study. From left to right are the results from one mobile observer, twenty static observers, twenty static with one mobile observers, and twenty mobile observers. The degree of observer bias is changed from low to high in the columns. The red solid lines and the blue dashed lines show the median bias along with robust intervals computed as $\pm 2c$MAD from the bias-corrected and uncorrected models across the 100 simulation replicates respectively. The MAD has been scaled by $c = 1.48$. This ensures that the intervals are asymptotically equivalent to the 95% confidence intervals that would be computed if the biases were normally distributed. Note that here all the analyst's assumptions correctly match the true data-generating mechanism, albeit with any overlap in the observers' efforts ignored.

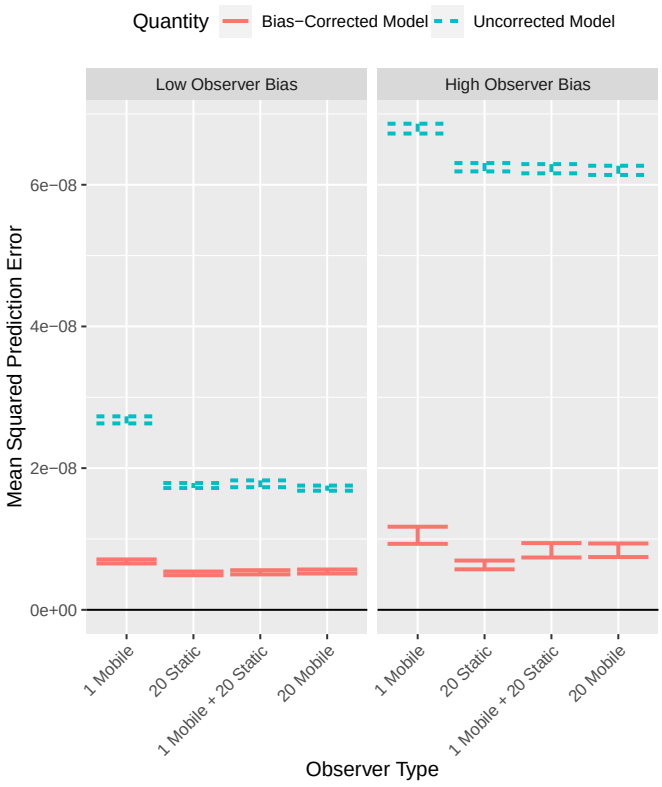

Figure A.27: A plot showing the mean squared prediction error (MSPE) of the animal's UD under the bias-corrected and bias-uncorrected models vs the types of observers. The results shown here are for the second simulation study. From left to right are the results from one mobile observer, twenty static observers, twenty static with one mobile observers, and twenty mobile observers. The degree of observer bias is changed from low to high in the columns. The red solid lines and the blue dashed lines show the median bias along with robust intervals computed as $\pm 2c\text{MAD}$ from the bias-corrected and uncorrected models across the 100 simulation replicates respectively. The MAD has been scaled by $c = 1.48$. This ensures that the intervals are asymptotically equivalent to the 95% confidence intervals that would be computed if the MSPE values were normally distributed. Note that here all the analyst's assumptions correctly match the true data-generating mechanism, albeit with any overlap in the observers' efforts ignored.

The standard bias correction approach simply estimates effort as

$$E(\mathbf{s}) = \sum_t \sum_{o \in O} p_{det}(o, \mathbf{s}, t).$$

However, the probabilities of detection from overlapping observers do not sum. We correct for this and compute:

$$E(\mathbf{s}) = \sum_t \left(1 - \prod_{o \in O} \left(1 - p_{det}(o, \mathbf{s}, t)\right)\right).$$

We implement this approach for 50 simulation iterations in the settings with twenty mobile observers. This setting is chosen since it suffers from the largest degree of overlap. Fig SA.28 demonstrates that the overlap corrected method indeed eliminates the bias in estimates of the animals UD center in both the low observer bias and high observer bias settings. However, no improvement in the MSPE is seen in Fig SA.29.

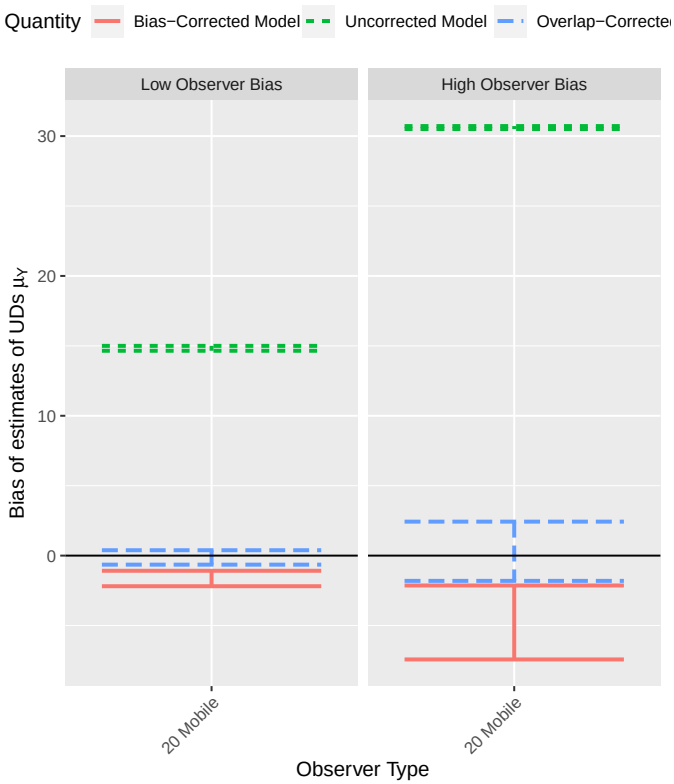

Figure A.28: A plot showing the bias of the estimated animal's UD center μ_y under the bias-corrected, bias-uncorrected, and the overlap-corrected models for the twenty mobile obervers. From left to right are the results when the degree of observer bias was either low or high. The red solid lines, the blue dashed lines, and the green dotted lines show the median bias along with robust intervals computed as $\pm 2c\text{MAD}$ from the bias-corrected, uncorrected, and overlap-corrected models across the 50 simulation replicates respectively. The MAD has been scaled by $c = 1.48$. This ensures that the intervals are asymptotically equivalent to the 95% confidence intervals that would be computed if the biases were normally distributed. Note that here the correct data-generating mechanism was assumed by the analyst. Notice that the bias is completely eliminated by the overlap-corrected method.

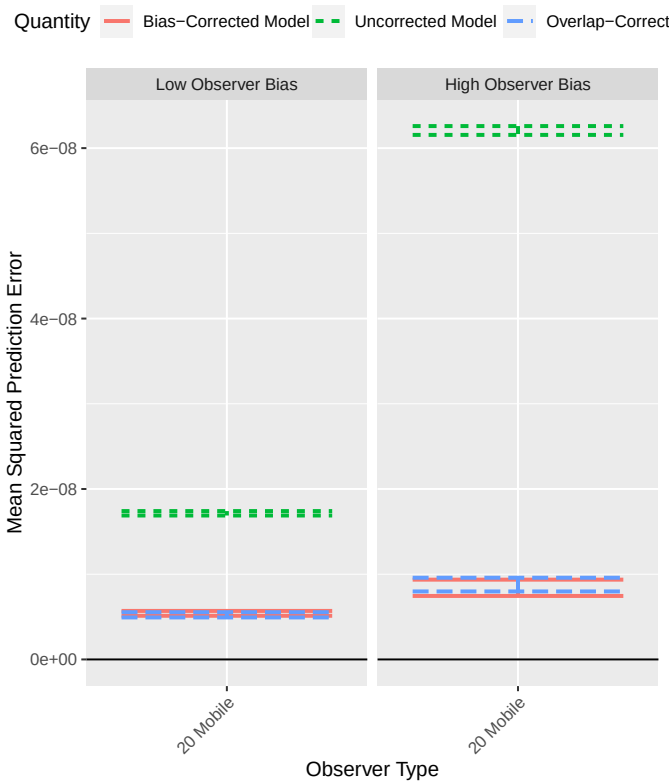

Figure A.29: A plot showing the mean squared prediction error (MSPE) of the estimated animal's UD under the bias-corrected, bias-uncorrected, and the overlap-corrected models for the twenty mobile obervers. From left to right are the results when the degree of observer bias was either low or high. The red solid lines, the blue dashed lines, and the green dotted lines show the median MSPE along with robust intervals computed as $\pm 2c\text{MAD}$ from the bias-corrected, uncorrected, and overlap-corrected models across the 50 simulation replicates respectively. The MAD has been scaled by $c = 1.48$. This ensures that the intervals are asymptotically equivalent to the 95% confidence intervals that would be computed if the MSPE values were normally distributed. Note that here the correct data-generating mechanism was assumed by the analyst. Notice that there is little-to-no improvement in the MSPE between the two corrected methods.

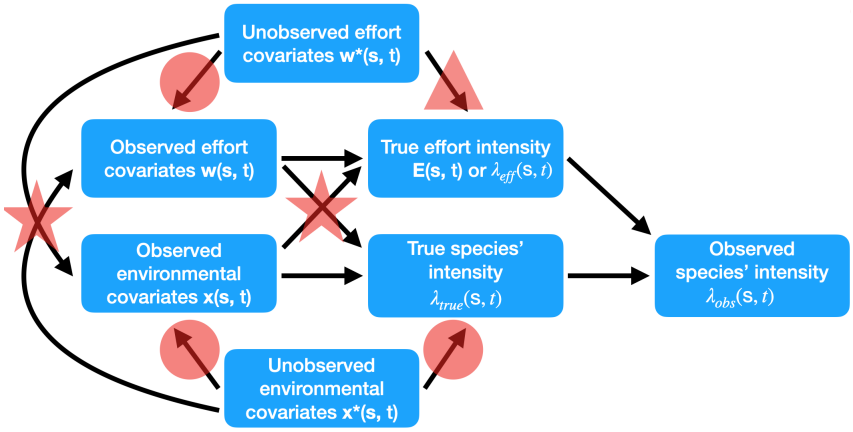

Figure A.30: A plot showing the assumed causal DAG for the proposed framework with the detection probability assumed constant. An arrow between a variable set A and a variable set B indicates that at least one variable exists in both sets with a direct causal effect between them. The causal Markov assumption is made such that a variable is independent of its non-descendants, when conditioned on its parents [Hernan and Robins, 2020].

A.3.4 Additional comments on the causal DAG

Model (5.10) is fit to a set of observed environmental covariates $\mathbf{x}(\mathbf{s},t)$ and observed effort covariates $\mathbf{w}(\mathbf{s},t)$, but in general, there may exist unobserved covariates $\mathbf{x}^\star(\mathbf{s},t)$ and $\mathbf{w}^\star(\mathbf{s},t)$. These unobserved covariates, in conjunction with the causal paths contained in the stars, the circles, and the triangle of Fig A.30 may cause problems. For example, the lower causal path denoted by the arrow within the red star on the left combined with the causal path within the red circle on the bottom right opens a back-door pathway between the effort intensity surface and the true species' intensity surface. This pathway passes through the unobserved environmental covariates $\mathbf{x}^\star(\mathbf{s},t)$ causing estimates of $\boldsymbol{\gamma}_1^T$ to be confounded by $\mathbf{x}^\star(\mathbf{s},t)$.

A similar conclusion may be drawn by considering the upper of the two arrows within the left had star, combined with the arrow seen in the triangle. Here, estimates of $\boldsymbol{\beta}^T$ will be confounded by $\mathbf{w}^\star(\mathbf{s},t)$. Further problems

would occur due to the two causal paths within the right star. These would lead to $\lambda_{true}(\mathbf{s},t)$ and $E(\mathbf{s},t)$ becoming non-identifiable. The existence of a subset of covariates $\tilde{\mathbf{w}}(\mathbf{s},t)$ within $\mathbf{w}(\mathbf{s},t)$ driving both $\lambda_{true}(\mathbf{s},t)$ and $E(\mathbf{s},t)$ causes neither intensity surface to be estimable. This is because only the sum of the effects of $\tilde{\mathbf{w}}(\mathbf{s},t)$ are estimable within the loglinear model (5.10). Thus for the true species' intensity surface to not be confounded by the effort intensity, none of the causal paths within the red stars can exist.

Yet more problems can occur if the four causal paths within the red circles and the red triangle exist. The upper two paths lead to estimates of $\boldsymbol{\gamma}_2^T$ being confounded by unmeasured effort covariates $\mathbf{w}^\star(\mathbf{s},t)$ and the bottom two paths lead to estimates of $\boldsymbol{\beta}^T$ being confounded by unmeasured environmental covariates $\mathbf{x}^\star(\mathbf{s},t)$. Furthermore, the existence of the causal path in the red triangle alone may lead to estimates of $\lambda_{true}(\mathbf{s},t)$ within (10) to be confounded by $\mathbf{w}^\star(\mathbf{s},t)$. This is because if a Gaussian process $Z(\mathbf{s},t)$ is included within the linear predictor for $\lambda_{true}(\mathbf{s},t)$ then any residual spatio-temporal correlations in the sightings data due to $\mathbf{w}^\star(\mathbf{s},t)$ may be erroneously captured by $Z(\mathbf{s},t)$.

Note that similar issues occur if the detection probability is not constant and a detection probability surface $p_{det}(\mathbf{s},t)$ is estimated with its own set of covariates. An extension to this causal DAG would allow for similar conclusions to be drawn. For the later case study, we assume that none of the causal paths within the stars or the triangle are present.

A.3.5 Deriving site occurrence and site count likelihoods

Likelihoods for site occurrence and site count data can all be derived from the modelling framework if the true locations of the target species follow a log-Gaussian Cox process. We ignore time for notational simplicity. With Ω our study region, with a known sampled region (e.g. a transect) $A_i \subset \Omega$, and with a known or estimable observer effort captured by $\lambda_{eff}(\mathbf{s},\mathbf{m})$, define the following quantity:

$$\Lambda_{obs}(A_i, \mathbf{m}|Z) = \int_{A_i} \lambda_{true}(\mathbf{s}, \mathbf{m}) p_{det}(\mathbf{s}, \mathbf{m}) \lambda_{eff}(\mathbf{s}, \mathbf{m}).$$

This is referred to as the integrated observed intensity function, conditional upon knowing the Gaussian process $\mathrm{Z}(\mathbf{s})$. Importantly, this represents the expected number of observed sightings within A_i. Following Hefley and Hooten [2016], we can then derive the target likelihoods. Firstly, suppose that an observer records the number of sightings made within A_i, denoted $N(A_i)$. Then, the distribution of the number of counts, conditioned upon knowing Z, is:

$$[N(A_i, \mathbf{m})|Z] \backsim \text{Poisson}(\Lambda_{obs}(A_i, \mathbf{m}|Z)).$$

In practice, the Gaussian process is not known and thus needs to be estimated. Consequently, the above likelihood is an example of a spatio-temporal generalized linear mixed effects model (STGLMM). Multiple software packages exist to fit such models (e.g. *R-INLA*). Next, suppose that instead of recording the number of sightings made within A_i, a binary presence/absence indicator of presence (denoted $P(A_i)$) was recorded. The distribution of this indicator variable can also be derived from the conditional Poisson distribution on the counts. In particular, let $O(A_i) = I\left(N(A_i) > 0\right)$, with I denoting the indicator function. Then the probability statement $P(O(A_i|Z) = 1) = P(N(A_i|Z) > 0) = 1 - \exp[\Lambda_{obs}(A_i|Z)]$ implies the following conditional distribution on the indicator variables:

$$[O(A_i, \mathbf{m})|Z] \backsim \text{Bernoulli}(1 - \exp\left[\Lambda_{obs}(A_i, \mathbf{m}|Z)\right]).$$

Once again, the likelihood is of the STGLMM format which can be computed using standard software packages. Note also that computing the integrated observed intensity function is critical across the likelihoods.

A.3.6 Comments on preferential sampling

In the setting of this Chapter, preferential sampling would be defined as a stochastic dependence between the observer effort and the underlying species intensity. An example would be a setting where observers focused their observer effort in areas with high species density, perhaps due to some prior knowledge on their likely locations. The biasing effects of preferential sampling on spatial prediction [Diggle et al., 2010] and on the estimation of the mean intensity in ecological applications [Pennino et al., 2019] have been shown.

In many situations this modelling framework will suitably adjust inference for any heterogeneous observer effort across Ω, removing the biasing effects of preferential sampling. In cases where nonzero observer effort exists throughout the study region (i.e. where $E(\mathbf{s}, \mathbf{m}) > 0 \, \forall \, (\mathbf{s}, \mathbf{m}) \in (\Omega \times M)$), the estimation of $\lambda_{true}(\mathbf{s}, \mathbf{m})$ will be unaffected by preferential sampling. However, when a subregion $B \subset \Omega$ is never visited, (i.e. when $E(\mathbf{s}, \mathbf{m}) = 0 \, \forall \, (\mathbf{s}, \mathbf{m}) \in (B \times M)$), the estimation of $\lambda_{true}(\mathbf{s}, \mathbf{m}))$ within B may be biased. To highlight this fact, suppose our study region Ω is split into a northern region A and a southern region B. Suppose that the true intensity $\lambda_{true}(\mathbf{s})$ takes value 2 within A and value 1 within B. If only A is visited, then without the availability of strong covariates explaining the differences across A and B, then any model will wrongly overestimate the true intensity in B, namely the model will predict that $\lambda_{true}(\mathbf{s}) = 1 \, \forall \, \mathbf{s} \in B$.

To minimize the impacts of preferential sampling on any conclusions made using this modelling framework, extrapolating predictions into unsampled regions should be done with care, especially if it is believed that the intensity of observer effort may depend upon the underlying species' intensity. This is standard advice in any statistical analysis and is not a limitation unique to this framework.

A.3.7 More notes on estimating the whale-watch observer effort

Two strong assumptions are required to allow us to multiply the total observer effort field by the fraction of total observer effort observed in a given month/year. We first assume that the expected spatial positions of the boats are constant throughout the time period of interest 9am - 6pm. We know that at the starts and ends of the days the boats will likely be closer to port. We assume however, that the whale-watch boats are travelling independently in equilibrium (represented by our estimated observer effort field $E_{WW}(\mathbf{s}, T_l, y)$).

Second, we assume that the boats are spread out throughout Ω sufficiently, such that the total observer effort from all the vessels (assumed equal) is additive. In other words, we assume the whale-watch boats are sufficiently spread out, such that their observation ranges do not overlap. In reality the whale-watch vessels often visit similar nature 'hotspots' and hence traverse similar routes. As a consequence, they may travel close together at certain times. At these times, their combined observer effort may not scale linearly with the number of the boats.

Given that we have chosen months to be our discretization of time, we must estimate the monthly observer effort across space, adding up the contributions of effort across the years of interest (2009 - 2016).

$$E^{obs}_{WW}(\mathbf{s}, T_l, m)) = \sum_{y=2009}^{2016} E^{obs}_{WW}(\mathbf{s}, y, T_l, m))$$

We define boat hours to be our unit of observer effort, kilometers to be our unit of distance and month to be our unit of time. Thus, $E^{obs}_{WW}(\mathbf{s}, T_l, m)$ denotes the number of WW boat hours of observer effort, per unit area that occurred for pod m, at location $\mathbf{s}$ and month T_l, summed over all the years of the study. In a similar flavour to the intensity surface, the effort surface is not really defined pointwise, but defined over regions of non-zero area as an integral. In particular, for a region $A \subset \Omega$ and month T_l, we define the total observer effort that occured inside A in boat hours to be:

$$\tilde{E}^{obs}_{WW}(A, T_l, m) = \int_A E^{obs}_{WW}(\mathbf{s}, T_l, m) d\mathbf{s} \tag{A.4}$$

For later computation of the LGCP, we approximate the stochastic integral required for the likelihood over a finite set of integration points. Thus, we are required to compute the integrals of the effort field over the integration points.

A.3.8 Computational steps for approximating the likelihood

The LGCP likelihood above (1) is analytically intractable, as it requires the integral of the intensity surface, which typically cannot be calculated explicitly. However, various methods exist for approximating this integral. We consider the approximation method from Simpson et al. [2016]. We present the spatial-only setting (i.e. $L = 1$) and ignore marks for notational convenience. The results generalize easily to the spatio-temporal case with marks. First, p suitable integration points are chosen in Ω with known corresponding areas $\{\tilde{\alpha}_j\}_{j=1}^{p}$. Then, the first p indices are defined to be the chosen integration points with the last n indices chosen as the observed locations of the sightings $\mathbf{s}_i \in \Omega$. Then, define $\boldsymbol{\alpha} = (\tilde{\boldsymbol{\alpha}}^T_{p\times 1}, \mathbf{0}^T_{n\times 1})^T$ and $\mathbf{y} = (\mathbf{0}^T_{p\times 1}, \mathbf{1}^T_{n\times 1})^T$. We define $\log(\eta_i) = \log(\lambda_{true}(\mathbf{s}_i) p_{det}(\mathbf{s}_i) \lambda_{eff}(\mathbf{s}_i))$. We obtain:

$$\pi(\mathbf{y}|\mathbf{z}) \approx K \prod_{i=1}^{n+p} \eta_i^{y_i} \exp(-\alpha_i \eta_i). \tag{A.5}$$

We can see that the stochastic integral is only approximated across the first p integration points, hence the name. The expected count around an integration point scales linearly with the area $\{\tilde{\alpha}_j\}$ associated with it. This is under the assumption that for fixed intensity, doubling the area of a region, doubles the expected number of encounters occurring within the region. The problem of evaluating (5.1) is reduced to a problem similar to evaluating $n + p$ independent Poisson random variables, conditional on

$\mathbf{Z} = \mathbf{z}$, with means $\alpha_i \eta_i$ and 'observed' values y_i. This is a Riemann sum approximation to the integral. In standard software, the natural logarithm of the weights α_i is added as an offset in the model and equation (A.5) can be fit if one defines the minor modification that $\log(\alpha_i)$ is defined to be zero if $\alpha_i = 0$. This is implemented as standard in the R-INLA package [Lindgren et al., 2011b, 2015, Rue et al., 2009].

Including known or estimated effort from the O observers in the model simply requires evaluating the areal-averaged effort that occurred at each encounter location and around each of the p chosen integration points $\mathbf{s}_i : i \in \{1, ..., p\}$. We denote the regions around the integration points, $A_j \subset \Omega$. These may correspond to regular lattice cells or as in our example, irregular Voronoi polygons (Fig A.31). Often there will be uncertainty surrounding the effort. We use the Monte Carlo sampling procedure seen in Chapter 5 to account for this uncertainty in our application.

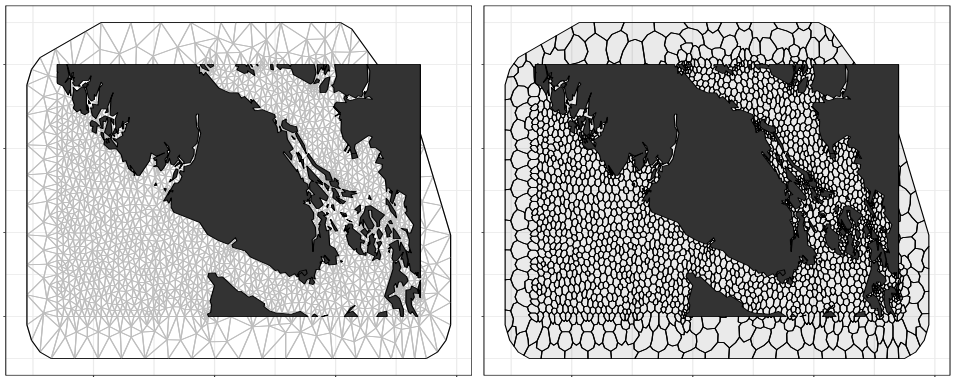

Figure A.31: The computational mesh on the left and the corresponding dual mesh on the right, formed by constructing Voronoi polygons around the mesh vertices. The Voronoi polygons form our integration points A_i.

A.3.9 Additional details on the results and additional tables

Environmental covariates were mapped to the integration points A_i and to the sighting locations $\mathbf{y}$ for modelling. In cases where we had noisy covariates with missing values, we chose the median covariate value (out of those that

spatially-intersect the Voronoi polygon) as the polygon's 'representative'. For missing covariates at observation locations, we mapped the non-missing value which was closest in distance to the observation location. Sea-surface temperature (SST) and (log) chlorophyll-A (chl-A) levels were obtained. Monthly chl-A and SST were obtained for each year and averaged over the years. Log transformed covariates were centered to have mean 0 and scaled to have unit variance. Sea surface temperature was not scaled for interpretation reasons.

Next we performed hierarchical centering of our SST and chl-A covariates. This is following the advice of Yuan et al. [2017], where it was shown that three unique biological insights can be obtained per covariate. In particular, we performed two types of centering: spatial and spacetime centering. Centering covariates like this can also improve the predictive performance of models. The 2 hierachical centering schemes applied to both SST and chl-A were compared. We refer to these as covariate sets 1 and 2.

Models that included a wide range of different latent effects within (5.13) were compared. A unique (sum-to-zero constrained) random walk of second order for each pod was tested, alongside a shared spatial and/or spatio-temporal Gaussian (markov) field across the pods and a unique spatial field for pod L. For the random walk term, we shared the precision parameter across the pods. We put INLA's default logGamma(1, 5e-05) prior distribution on this shared (log) precision. Finally a unique intercept was allowed for each pod. The unique intercepts per pod allow for a different global intensity for each pod to exist across the months, whilst the unique random walk terms per pod allow for a changing relative intensity of each pod across the months. This is chosen based on previous work that found pod J to be the most likely to be present in the Salish Sea year-round [Ford et al., 2017].

We also fitted the models without covariates included in the linear predictor (5.13) and hence only with spatial and spatio-temporal terms included in the model. We also fitted models with covariates present in (5.13), but with the spatial random fields removed. These are inhomogeneous Poisson processes. We did this to attempt to show how the variability seen in the data is captured by covariates and random effects. We also did this to

investigate whether or not the spatial distribution of the SRKW intensity (conditioned on the observer effort), changes with month, or whether or not it is spatially static across the months.

For all spatial fields, we placed PC priors [Fuglstad et al., 2018] on the GMRF, with a prior probability of 0.01 that the ranges of the fields are less than 15km. We also placed a prior probability of 0.1 that the standard deviations of the fields exceed 3. Thus, our prior beliefs were that the fields are smooth (i.e. the ranges are not too small) and are not too large in amplitude (i.e. the standard deviations are not too large). We did this to reduce the risk of over-fitting the data. The PC priors penalize departures from our prior beliefs under the Occam's razor principle; penalizing models with greater complexity than that specified in our prior.

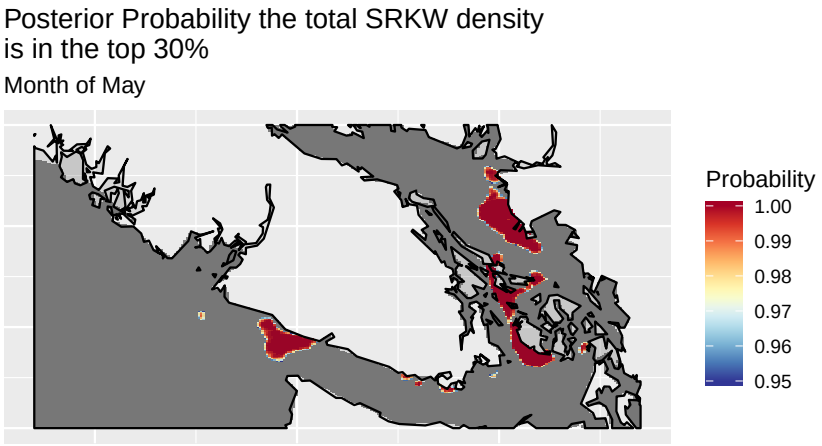

Figure A.32: A plot showing the posterior probability that the sum of the three pod's intensities across the region takes value in the upper 30% for the month of May. The 30% exceedance value is computed across all the months. Shown are the probabilities of exceedance, with only the probabilities greater than 0.95 displayed. Results shown are for the 'best' model, adjusted for Monte Carlo observer effort error.

Now we display the table of coefficients from the 'best' model, and the table of DIC values of all tested candidate models. Finally, we display our model-estimated number of sightings per pod and per month, with 95% credible intervals. We also display the observed number of sightings to check the model's calibration.

Table A.1: A table of posterior estimates of the fixed effects $\boldsymbol{\beta}$, with their 95% posterior credble intervals for the final model (Model 8 in Table A.2) repeatedly fit with 1000 Monte Carlo estimated observer effort fields. Note that the symbol * denotes 'significance' such that the 95% credible intervals do not cover 0 (no effect). For the pods, * indicates that a difference was found between the relative pod intensities with respect to their 95% credible intervals. The 'change' column displays the change in 'significance' of the effect size compared with the results from model 8 without the additional MC error from the observer effort. The '-' symbol denotes no change in significance. None of the directions and hence qualitative conclusions of the effect estimates change.

	Mean	SD	0.025 Q	0.5 Q	0.975 Q	Δ
Pod J	-3.84	0.71	-5.23	-3.83	-2.44	-
Pod K	-4.57	0.71	-5.95	-4.56	-3.20	-
Pod L	-3.95	0.98	-5.81	-3.96	-1.99	-
SST month avg	0.03	0.26	-0.49	0.04	0.54	-
SST spatial avg*	-0.37	0.17	-0.70	-0.37	-0.05	-
chl-A month avg	0.31	1.11	-1.89	0.26	2.58	-
chl-A spatial avg*	-1.03	0.32	-1.67	-1.02	-0.38	-
SST ST residual*	-0.67	0.05	-0.77	-0.67	-0.57	-
chl-A ST residual*	-0.23	0.07	-0.38	-0.23	-0.09	-

Table A.2: A table showing the DIC values of all the models tested, with the model formulations summarized in the columns. A value of NA implies that model convergence issues occurred.

Model	*DIC*	ΔDIC	*Covariate Set*	*Shared Field*	*Field for L*
0	3614	5554	×	×	×
1	2843	4783	1	×	×
2	2707	4642	2	×	×
3	-1633	307	×	Spatial	×
4	-1730	210	×	Spatio-temporal	×
5	-1842	98	1	Spatial	×
6	-1851	89	2	Spatial	×
7	-1931	9	1	Spatial	Spatial
8	**-1940**	0	2	Spatial	Spatial
9	NA	NA	1	Spatio-temporal	×
10	NA	NA	2	Spatio-temporal	×
11	NA	NA	1	Spatio-temporal	Spatial
12	NA	NA	2	Spatio-temporal	Spatial

A.3.10 Pseudo-code for computing the modelling framework in inlabru

Fitting log-Gaussian Cox processes within a Bayesian framework models is greatly simplified with the use of the R package inlabru [Bachl et al., 2019]. Furthermore, inlabru can fit joint models containing many (possibly different) likelihoods, and is able to share parameters and latent effects between them with ease. Numerous other features exist and helper functions are provided to help produce publication-quality plots. For full information and for free tutorials, visit the inlabru website at inlabru.org. The following pseudo-code is largely based on the available tutorials.

In this section we will demonstrate the simplicity of fitting a joint model to a dataset comprised of a distance sampling survey, and a presence-only

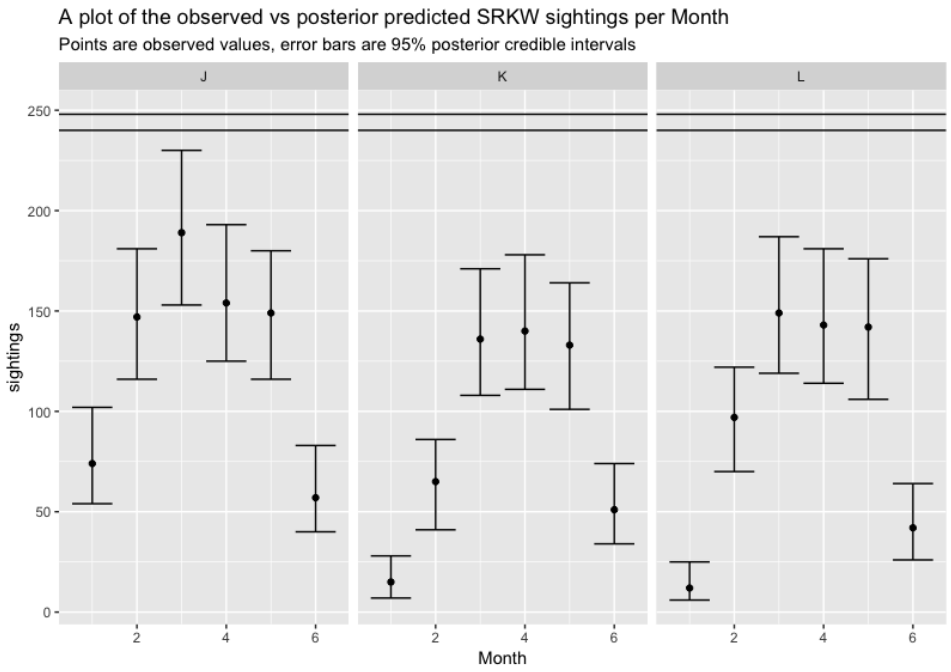

Figure A.33: A plot showing the total observed number of sightings made per month with the posterior 95% credible intervals shown. Results shown are for Model 8 with MC observer effort error. The posterior predictions are made, given the identical observer effort to that estimated for the observed data. Also shown are the horizontal lines showing the maximum possible number of sightings that could be made in months with 30 and 31 days respectively. Posterior credible intervals extending above this upper bound imply the Poisson model is severely misspecified.

dataset with a corresponding observer effort field using inlabru's syntax. For simplicity, suppose we have 1 continuous environmental covariate, called covar1 (e.g. SST), and that it is in the 'SpatialGridDataFrame' or 'SpatialPixelsDataFrame' class. Next, suppose we have an estimate of the natural logarithm of the observer effort for the presence-only data called logeffort_po, also of class 'SpatialGridDataFrame' or 'SpatialPixelsDataFrame'. We assume that the effort took values strictly greater than 0 everywhere before

taking the logarithm (or that we have added a small constant to enforce this).

Suppose we have the observed sighting locations of the individual of interest as two separate objects of class 'SpatialPointsDataFrame', one for the survey sightings and one for the presence-only sightings. Call these surv_points and po_points and suppose we have thinned the data to ensure any autocorrelation has been removed. Suppose also that we have our transect lines from the survey as an object called surveylines in the class 'SpatialLinesDataFrame', and that we know the transect strip half-width (denoted W). Finally, suppose that our spatial domain of interest is described by an object called boundary of 'SpatialPolygonsDataFrame' class. All 'Spatial' objects in the sp package [Bivand et al., 2013, Pebesma and Bivand, 2005] must be in the same coordinate reference system.

Suppose we wished to estimate a half-normal detection probability function, as a function of distance. To program this in inlabru, we must first define the half-norm detection probability function in R [R Core Team, 2019]. Let 'logsigma' denote the natural logarithm of the standard deviation and 'distance' denote the perpendicular distance from the transect to the observed point. Then our function is:

```
halfnorm = function(distance, logsigma){
  exp(-0.5*(distance/exp(logsigma))^2)
}.
```

Next, given a well constructed Delauney triangulation mesh called 'mesh', we construct the spatial random field for the LGCP. Helper functions exist for creating appropriate meshes in inlabru. The code for creating the spatial field, with Matern covariance structure is:

```
matern <- inla.spde2.pcmatern(mesh,
 prior.sigma = c(upper_sigma, prior_probs),
 prior.range = c(lower_range, prior_probr)).
```

Here, upper_sigma, prior_probs, lower_range and prior_probr all define the parameters of the PC prior on the random field [Fuglstad et al., 2018]. Once again, the tutorials help assist with the choice of prior. Now, we define all the parameters and terms in the model that must be estimated:

```
mod_components <- ~ mySpatialField(map = coordinates, model = matern) +
  beta.covar1(map = covar1, model = 'linear') +
  po_search_effort(map = logeffort_po, model = 'linear',
                   mean.linear=1, prec.linear=1e20) +
  logsigma + Intercept_Survey + Intercept_PO.
```

Note here that we choose the prior mean and precision of the 'po_search_effort' field to enforce it to enter the model as an offset. Now we can create the likelihood objects for both data types, each with their own formulae, but sharing components.

```
lik_surv <- like('cp',
              formula = coordinates ~ Intercept_Survey + mySpatialField +
                        beta.covar1 + log(halfnorm(distance, logsigma)) +
                        log(1/W),
              data = surv_points,
              components = mod_components,
              samplers = surveylines,
              domain = list(coordinates = mesh))
lik_po <- like('cp',
              formula = coordinates ~ Intercept_PO + mySpatialField +
                        beta.covar1 + po_search_effort,
              data = po_points,
              components = mod_components,
              samplers = boundary,
              domain = list(coordinates = mesh)
```

And then we can fit the joint model and simulate M samples of all of the parameters and latent effects from the posterior distribution.

```
fit_joint <- bru(mod_components, lik_surv, lik_po)
posterior_samples <- generate(fit_joint, n.samples = M).
```

Note that once the model object (fit_joint) is created, the estimated field can be easily plotted, and predictions can easily be made on new datasets and at new locations. Stochastic integration of the field to estimate abundance (for suitable datasets) is also possible using inlabru helper functions. Details can be found in the inlabru tutorials. The above code can scale up to include multiple environmental covariates (including categorical predictors), spatio-temporal fields, and/or temporal effects. Likelihoods of different type (e.g. Bernoulli, Poisson, Gaussian etc.,) can all be included, with this feature becoming especially useful for when the joint estimation of data of differing type is desired.

A.3.11 Additional figures

Covariate plots

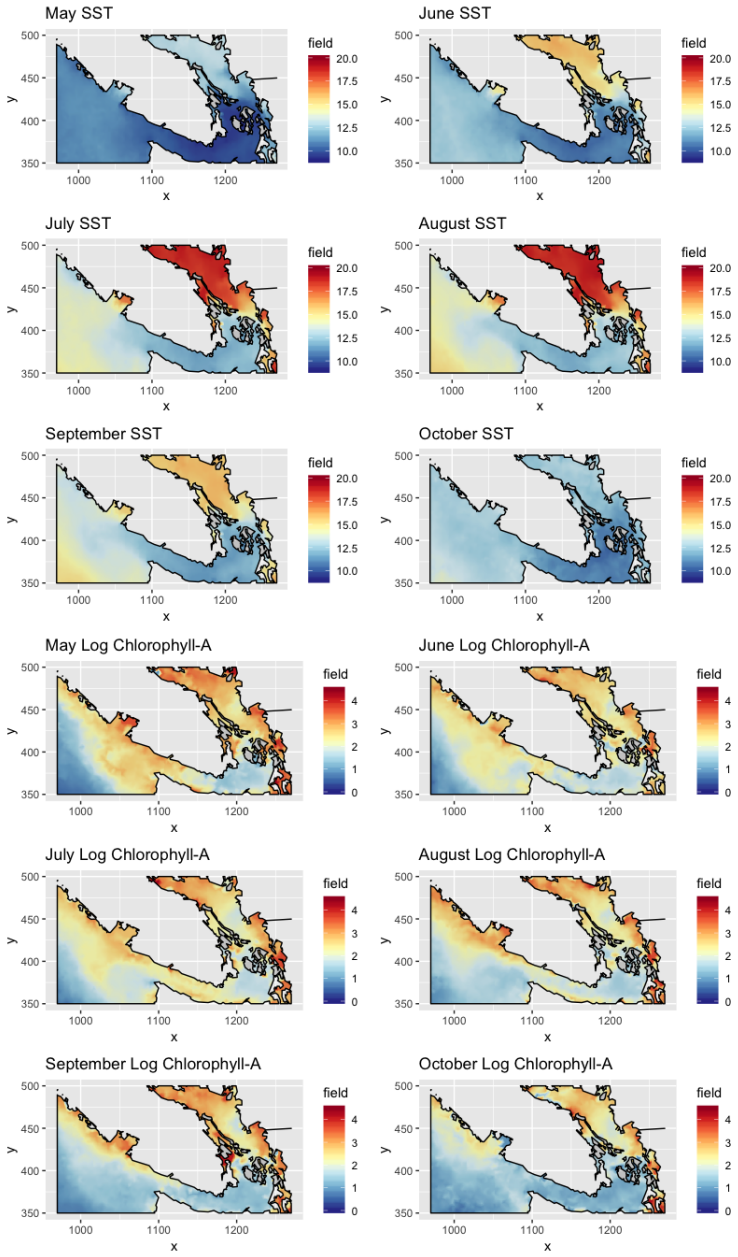

Figure A.34: Plots showing the average monthly sea-surface temperatures in degrees Celsius (top 6) and the natural logarithm of chlorophyll-A concentrations in mgm^{-3} (bottom 6). The averages have been taken over the years 2009-2016.

Plot of the pod-specific random walk effects

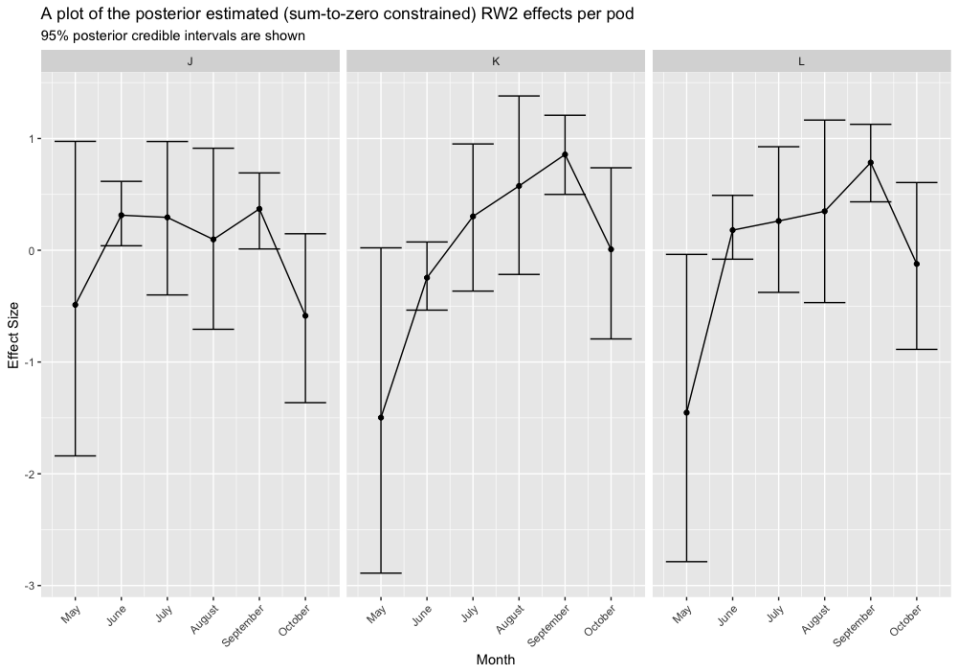

Figure A.35: A plot showing the posterior mean and posterior 95% credible intervals of the pod-specific (sum-to-zero constrained) random walk monthly effect from the 'best' model with Monte Carlo observer effort error included.

Plot of model standard deviation

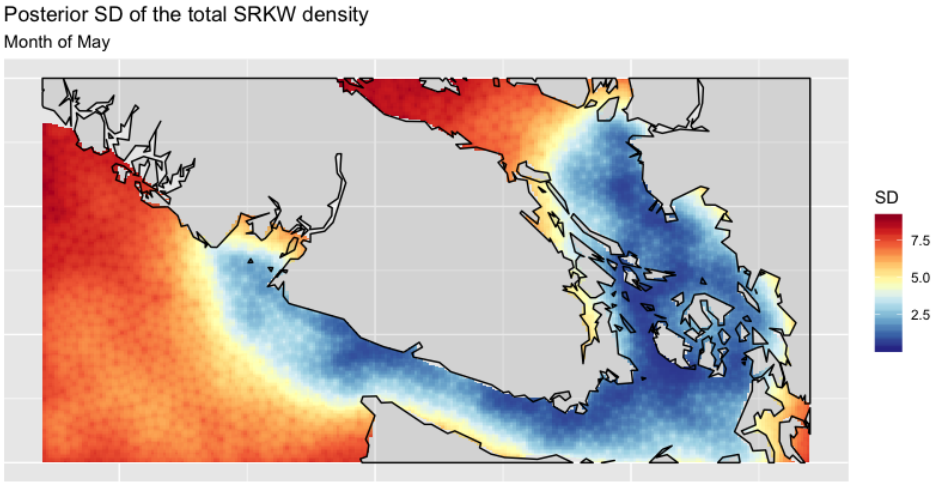

Figure A.36: A plot showing the posterior standard deviation of the sum of the SRKW intensities for the three pods, for the month of May. The qualitative behaviour is almost identical across all pods and across all months, so we omit them. Results shown are for Model 8 with MC observer effort error. Note that the computational mesh is visible in the plot as we linearly interpolated the standard deviations from the computational mesh vertices to the pixel locations, instead of approximating the full posterior distributions at each pixel location. This was done to reduce computation time.