A Markov Random Fields Approach to Modelling Habitat

by

Andrés E. Sánchez-Ordoñez

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

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL
STUDIES
(Statistics)

The University of British Columbia
(Vancouver)

August 2015

© Andrés E. Sánchez-Ordoñez, 2015

Abstract

Habitat modelling presents a challenge due to the variety of data available and their corresponding accuracy. One option is to use Markov random fields as a way to incorporate these distinct types of data for habitat modelling. In this work, I provide a brief overview of the intuition, mathematical theory, and application considerations behind modelling habitat under this framework. In particular, an auto-logistic model is built and applied to modelling sea lion habitat using synthetic data. First, we explore modelling one sample of data. Afterwards, the framework is extended to the multi-sample scenario. Finally, the theory for the methodology is presented, the results of the applied implementation are presented.

Preface

This dissertation is original, unpublished, independent work by the author Andrés E. Sánchez-Ordoñez under the supervision of Prof. Matías Salibián-Barrera.

Table of Contents

List of Tables

List of Figures

Glossary

MCMC Markov chain Monte Carlo

MLE Maximum Likelihood Estimator

MPLE Maximum Pseudo-Likelihood Estimator

Acknowledgments

Special gratitude is owed to my parents and my brother without whose continued support this work would not be possible. I offer great gratitude to Prof. Matías Salibián-Barrera for his immense and continued help and advice from conception to completion of the work presented here. Finally, I want to thank the faculty, staff, and my fellow colleagues for all of their encouragement.

Chapter 1

Motivation

The determination of animal habitat is an open problem in the ecological sciences. For ecologists habitat can have a variety of interpretations. One common interpretation is that of critical habitat, which is of particular importance for governmental regulation and sustainability of animal life [14]. Another interpretation is that of habitat as the place where animals reside the majority of their time [2]. This latter interpretation has scientific significance as it allows ecologists to understand the behaviour and habits of the animals under study. We focus on exploring this latter definition with regards to marine mammals and, in particular, sea lions.

In order to comprehend animal habitat, a variety of distinct data has been gathered by researchers. Broadly, the data can be categorised into two types: systematic surveys and opportunistic sightings.

The first of these types of data, namely, systematic surveys consist of a meticulous combing of a target area believed to be of significance for animal habitat. Either through aerial or marine surveys, ecologists discretise and travel the area of potential habitat in a pre-specified order. Afterwards, the number of animals, as well as characteristics of these (e.g., gender or approximate age), are recorded. Included within systematic surveys are counts of animal populations on land. Since systematic surveys are usually carried out by experts on the field, the data recorded can often be assumed to have high accuracy. Furthermore, the discretised nature of the data facilitates the modelling of spatial auto-correlation in a region [19] and the incorporation of environmental covariates [13].

As opposed to systematic surveys, opportunistic sightings data, also known as platform of opportunity data, are gathered in a less rigorous manner. Opportunistic sightings data consists of approximate locations at which an animal was observed by an individual passing by a specified area [7]. The data are usually gathered by non-experts including members of fishing vessels or members of the general public. For example, members of commercial fishing vessels may voluntarily record any sea lion sighted during their daily routine. Given that this type of data is gathered by non-experts voluntarily, the data is usually not believed to be highly accurate as it is difficult to verify that the individual who observed the animal is sufficiently knowledgeable to distinguish similar looking animal species [7]. Furthermore, the location at which the animal is observed is generally not recorded with high precision by non-experts as a consequence of the voluntary aspect of the data gathering mechanism. More problematic still, opportunistic sightings do not record the *effort* exerted into observing the animal. That is, the records do not communicate whether the animal was observed after spending a day looking for an animal at the location or whether the animal was observed in the few minutes taken for a vessel to cross the location. This also leads to an over-abundance of zeroes (i.e., absence observations) in platform-of-opportunity data, generally denoted as pseudo-absences to account for the limitations in the data gathering mechanism [23]. Some studies have incorporated ad-hoc measures of effort when predicting habitat from opportunistic sightings [7]. Other studies, however, have opted to use opportunistic sightings as a tool for model validation rather than for model building [14] [20]. A third viable alternative, which we later consider, is to aggregate opportunistic sightings over time and use these as covariates in a larger framework as a way to dampen the noise in the data.

A major drawback of both systematic surveys and opportunistic sightings is that the data gathered only accounts for the presence of an animal at the specified location at the specified time. More specifically, systematic surveys may record an animal at one location on the first day and may also record the same animal at another location a week later. Since without some sort of tagging, it may be difficult to identify whether the animal recorded during a survey has been recorded before, systematic surveys can lead to biased data on animal habitat due to over-estimation of the animals in the area and a misrepresentation of their habitat preferences.

Similarly, since opportunistic sightings may be recorded by distinct groups of non-synchronised individuals, the records on a certain day may show a multitude of animals observed at a location while in reality it may just be the same animal observed by multiple individuals at the location. Finally, it is important to note that both systematic surveys and opportunistic sightings cannot account with certainty the absence of an animal. That is, if an animal is not observed at a location, an observer cannot state with certainty that the location is unfavourable habitat for the animal [23] [22]. All that can be stated is that at the specified time an animal was not observed at the location. This issue is more problematic for the opportunistic sightings data since there is no indication of where the animals may or may not have travelled.

An additional drawback of both systematic surveys and opportunistic sightings is that neither of these types of data are driven by the behaviour of the animal itself. Systematic surveys sample locations believed to be pertinent to animal habitat, and opportunistic sightings record by chance observations of animals. However, ecological theory suggests that habitat is driven by the availability of resources [3]. In this setting resources refers to characteristics of the environment that benefit or facilitate the survival and growth of an animal population [4] [8]. Generally, these may refer to the availability of food or lack of predators. Quantifying these, however, can be rather difficult. Therefore, resources are usually inferred by physical environmental variables such as topography, bathymetry, temperature, chlorophyll abundance, etc. [20] [15] [21]. The correlation between physical environmental variables and resources is general justified by the observed or assumed correlation, which is based on past literature and scientific theory [14].

Lately, a new type of data has gained popularity that can alleviate some of the drawbacks faced by systematic surveys and opportunistic sightings. Using tags attached to animals, ecologists are now recording telemetry or tracking data. These consist of satellite-based locations of an animal, usually sampled multiple times a day, through an extended period of time. In addition to recording the animal's location, several of the tags currently in use also record environmental variables at the location such as temperature, bathymetry, light intensity, etc. Since telemetry data originates from the location travelled by the animal itself, the data is believed to provide a less biased view of the animal's habitat preferences. However, there are

a few limitations to telemetry data. First, the tags are usually attached to a specific subset of the animal population giving a potentially biased data sample. For example, in the case of Steller sea lions, tags are usually attached to females or pups since attaching the tags to the male sea lions is a dangerous task [21]. Second, due to the cost of the tags, only a few animals in a colony are tagged providing a sample that may not be representative of the population. Nevertheless, the abundance of information sampled by the telemetry data tags has allowed certain studies to use off-the-shelf methods to model habitat [21]. While these models incorporate the animals' trajectory and environmental variables, the spatial auto-correlation of the data is absent.

In order to get a more complete understanding of animal habitat, it would be ideal to be able to incorporate distinct types of data into a single model. We propose the use of an auto-logistic model with covariates as a reasonable habitat model. The response variable in the model is an indicator for whether the animal occupied a discrete spatial location based on the telemetry data. The joint probability of the telemetry data locations are modelled by an exponential function of the covariates. The covariates, in this case, could consist of environmental data or come from opportunistic sightings or systematic surveys. Moreover, spatial auto-correlation is also incorporated into the model by way of a Markov neighbourhood structure of the telemetry data. Furthermore, by using an auto-logistic model, important covariates could be reasonably determined within a hypothesis testing framework.

Chapter 2

Background

This section constitutes a review of the literature necessary for deriving and understanding the auto-logistic model that is implemented in our work. The background presented here draws mainly on the results found in [5] and [9]. First, the theoretical foundation of Markov random fields is introduced including the definitions of the spatial neighbourhood structure and potential function utilised for building an auto-logistic model. Afterwards, the framework of conditional dependence of the model is briefly discussed. Finally, using the foundations discussed, an auto-logistic model is formally defined.

2.1 Markov Random Fields

In order to define a Markov random field, it is necessary to first define the concept of a neighbour with respect to a distribution as opposed to a topology. Let $\boldsymbol{Y}$ be a discrete random variable defined over a squared lattice $\mathscr{D}$ on $\mathbb{R}^2$. The sites $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N\} \subset \mathscr{D}$ denote the N spatial locations at which observations are available. Then, following the notation used in [9], write a conditional probability process $P(Y(\boldsymbol{s}_1) = y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_I) = y(\boldsymbol{s}_I) | Y(\boldsymbol{s}_{I+1}) = y(\boldsymbol{s}_{I+1}), \ldots, Y(\boldsymbol{s}_{I+J}) = y(\boldsymbol{s}_{I+J}))$ as $P(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_I) | y(\boldsymbol{s}_{I+1}), \ldots, y(\boldsymbol{s}_{I+J}))$.

Definition 1 (Neighbour) *A site k is a neighbour of site i if the conditional distribution of $Y(\boldsymbol{s}_i)$, given all other site values, depends functionally on $Y(\boldsymbol{s}_k)$ for $i \neq k$.*

Definition 2 (Clique) *A clique denotes a set of sites consisting of either a single site or sites that are neighbours to each other.*

Having now defined a neighbourhood structure for a site, we can now formally define a Markov random field as follows:

Definition 3 *Any probability measure whose conditional distributions define a neighbourhood structure* $\{\mathcal{N}_i : i = 1,\ldots,N\}$ *is defined to be a Markov random field.*

In other words, a Markov random field is a set of random variables each of whose conditional distribution depends on each other based on a Markov neighbourhood structure. In other words, a Markov random field is an undirected graph with the Markov property.

Given the conditional dependence of a Markov random field, we make use of the factorisation theorem [5] below to specify a joint probability mass or density function as the product of conditional probabilities. This enables us to compute difficult joint probabilities as the product of simpler conditional probabilities.

Theorem 2.1.1 (Factorisation Theorem) *Let* $P(\cdot)$ *denote the joint probability mass function of a random variable* $\{Y(\boldsymbol{s}_i) : i = 1,\ldots,N\}$ *whose support* ζ *satisfies the positivity condition (see below). Then,*

$$\frac{P(\boldsymbol{y})}{P(\boldsymbol{z})} = \prod_{i=1}^{N} \frac{P(y(\boldsymbol{s}_i)|y(\boldsymbol{s}_1),\ldots,y(\boldsymbol{s}_{i-1}),z(\boldsymbol{s}_{i+1}),\ldots,z(\boldsymbol{s}_N))}{P(z(\boldsymbol{s}_i)|z(\boldsymbol{s}_1),\ldots,z(\boldsymbol{s}_{i-1}),y(\boldsymbol{s}_{i+1}),\ldots,y(\boldsymbol{s}_N))}, \tag{2.1}$$

for all $\boldsymbol{y},\boldsymbol{z} \in \zeta$ *where* $\boldsymbol{y} = (y(\boldsymbol{s}_1),\ldots,y(\boldsymbol{s}_N))^T$ *and* $\boldsymbol{z} = (z(\boldsymbol{s}_1),\ldots,z(\boldsymbol{s}_N))^T$ *are all possible realisations of* $\boldsymbol{Y}$.

In order for the factorisation theorem to be true, however, the positivity condition below must be satisfied.

Definition 4 (Positivity Condition) *Let* $\{y(\boldsymbol{s}_i) \ : i = 1,\ldots,N\}$ *be the values of a discrete random variable Y observed at the locations on a lattice. Define* $\zeta \equiv \{\boldsymbol{y} : P(\boldsymbol{y}) > 0\}$ *and* $\zeta_i \equiv \{y(\boldsymbol{s}_i) : P(y(\boldsymbol{s}_i)) > 0\}$ *for* $i = 1,2,\ldots,N$. *The positivity condition is satisfied if* $\zeta = \zeta_1 \times \zeta_2, \times \cdots \times \zeta_N$.

In essence, the positivity condition ensures that when the joint probability function is factorised as in Theorem 2.1.1, none of the individual conditional probabilities are zero and, thus, the product of the conditionals will be non-zero. Although the necessity of the positivity condition makes intuitive sense, it is nevertheless worth illustrating the condition through an example adapted from [9] in order to see possible complications in the specification of a model based on a Markov random field.

Let $y(\boldsymbol{s}_i) = \mathbb{I}(s_i \text{ is occupied})$, that is, an indicator for a presence being observed at site i, and assume that site $y(\boldsymbol{s}_j)$ can only be occupied if one of its immediate neighbours in the four cardinal directions is occupied. Under these conditions, the following arrangement would need to have a probability of zero.

$$\begin{matrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{matrix} \tag{2.2}$$

This is because the only way the centre site could be occupied is through contact with an occupied neighbour in the four cardinal directions. Hence, unless this is the starting state for the occupancy process and assuming that occupied sites cannot revert back to being unoccupied, the probability of the arrangement above occurring should be zero by the positivity condition.

Now, assume that the initial occupied site is somewhere on the lattice other than the observed occupied site and that we arrive at the following configuration:

$$\begin{matrix} y(\boldsymbol{s}_1) & 0 & y(\boldsymbol{s}_2) & y(\boldsymbol{s}_3) \\ 0 & 1 & y(\boldsymbol{s}_4) & y(\boldsymbol{s}_5) \\ y(\boldsymbol{s}_6) & 0 & y(\boldsymbol{s}_7) & y(\boldsymbol{s}_8) \end{matrix}$$

Note that under this configuration, we should observe

$$P\left(y(\boldsymbol{s}_4) = 0 | \{y(\boldsymbol{s}_j) : j \neq 4\}\right) = 0,$$

as otherwise we would get the same situation as in the Eq. 2.2 above where the positivity condition is violated. However, note that given the neighbours of $\boldsymbol{s}_4$ in the four cardinal directions, namely, $N(\boldsymbol{s}_4) : \{y(\boldsymbol{s}_2), 1, y(\boldsymbol{s}_5), y(\boldsymbol{s}_7)\}$, we have

$P(y(\boldsymbol{s}_4) = 0|N(\boldsymbol{s}_4)) > 0$ since we only observe one occupied neighbour for $\boldsymbol{s}_4$. Since we cannot ensure that $P(y(\boldsymbol{s}_4) = 0|\{y(\boldsymbol{s}_j) : j \neq 4\}) = 0$, the positivity condition is violated in this case. Therefore, when specifying conditional probability models based on Markov random fields, it is important to check that the framework of the stochastic process allows the positivity condition to be satisfied.

2.2 The Negative Potential Function

Having now covered the necessary foundation for Markov random fields, we now introduce the (negative) potential function which will play an important role in the specification of the auto-logistic model.

Assumption 1 *Without loss of generality, assume zero can occur at any site [9]; that is,* $\mathbf{0} \in \zeta$.

Definition 5 *Let Q be the negative potential function where*

$$Q(\boldsymbol{y}) \equiv \log\left\{\frac{P(\boldsymbol{y})}{P(\mathbf{0})}\right\} \textit{ for } \boldsymbol{y} \in \zeta.$$

It follows from Def. 5 that

$$\exp(Q(\boldsymbol{y})) = \frac{P(\boldsymbol{y})}{P(\mathbf{0})} \text{ for } \boldsymbol{y} \in \zeta,$$

such that,

$$P(\boldsymbol{y}) = \frac{\exp(Q(\boldsymbol{y}))}{\sum_{\boldsymbol{z}\in\zeta}\exp(Q(\boldsymbol{z}))},$$

where the normalising constant $\sum_{\boldsymbol{z}\in\zeta}\exp(Q(\boldsymbol{z}))$ is commonly called the partition function.

Now, we make use of the factorisation theorem to derive some important properties of the negative potential function Q that allow a conditional probability model to be later derived.

Proposition 1 *The negative potential function Q satisfies the following two equa-*

tions:

$$\frac{P(y(\boldsymbol{s}_k)|\{y(\boldsymbol{s}_j) : j \neq k\})}{P(0(\boldsymbol{s}_k)|\{y(\boldsymbol{s}_j) : j \neq k\})} = \frac{P(\boldsymbol{y})}{P(\boldsymbol{y}_{0(\boldsymbol{s}_k)})} = \exp\{Q(\boldsymbol{y}) - Q(\boldsymbol{y}_{0(\boldsymbol{s}_k)})\}, \tag{2.3}$$

where $0(\boldsymbol{s}_k) \equiv \{y(\boldsymbol{s}_k) = 0\}$ *and* $\boldsymbol{y}_{0(\boldsymbol{s}_k)} \equiv (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_{k-1}), 0, y(\boldsymbol{s}_{k+1}), \ldots, y(\boldsymbol{s}_N))^T$. *Furthermore, Q can be expanded uniquely on* ζ *as*

$$\begin{aligned} Q(\boldsymbol{y}) = &\sum_{1 \leq i \leq N} y(\boldsymbol{s}_i) G_i(y(\boldsymbol{s}_i)) + \sum_{1 \leq i < j \leq N} y(\boldsymbol{s}_i) y(\boldsymbol{s}_j) G_{ij}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)) \\ &+ \sum_{1 \leq i < j < l \leq N} y(\boldsymbol{s}_i) y(\boldsymbol{s}_j) y(\boldsymbol{s}_l) G_{ijl}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j), y(\boldsymbol{s}_l)) \\ &+ \cdots + y(\boldsymbol{s}_1) \cdots y(\boldsymbol{s}_N) G_{1 \ldots N}(y(\boldsymbol{s}_1), \ldots y(\boldsymbol{s}_N)), \end{aligned} \tag{2.4}$$

where $G_i(\cdot)$ *is defined as* $y(\boldsymbol{s}_i) G_i(y(\boldsymbol{s}_i)) \equiv Q(0, \ldots, 0, y(\boldsymbol{s}_i), 0, \ldots, 0)$ *and*

$$\begin{aligned} y(\boldsymbol{s}_i) y(\boldsymbol{s}_j) G_{ij}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)) \equiv &\, Q(0, \ldots, 0, y(\boldsymbol{s}_i), 0, \ldots, 0, y(\boldsymbol{s}_j), \ldots) \\ &- Q(0, \ldots, 0, y(\boldsymbol{s}_i), 0, \ldots, 0) \\ &- Q(0, \ldots, 0, y(\boldsymbol{s}_j), 0, \ldots, 0), \end{aligned}$$

where Q is a unique function on ζ *if* $G_{ij}(\cdot) = 0$ *whenever* $y(\boldsymbol{s}_i) = 0$ *or* $y(\boldsymbol{s}_j) = 0$.

Therefore, from Eq. 2.3 and Eq. 2.4, we can specify $Q(\boldsymbol{y})$ in terms of the conditional probabilities:

$$y(\boldsymbol{s}_k) G_k(y(\boldsymbol{s}_k)) = \log \left\{ \frac{P(y(\boldsymbol{s}_k)|\{0(\boldsymbol{s}_j) : j \neq k\})}{P(0(\boldsymbol{s}_k)|\{0(\boldsymbol{s}_j) : j \neq k\})} \right\}. \tag{2.5}$$

The equations described in Proposition 1 are important in that they allow one to precisely define forms of the negative potential function for which the partition function can be defined as a closed form of the parameters of Q. This result is proved in an unpublished paper by Hammersley and Clifford [17] as follows:

Theorem 2.2.1 (Hammersley-Clifford Theorem) *Suppose* $\boldsymbol{Y}$ *is distributed according to a Markov random field* ζ *that satisfies the positivity condition. Then, the negative potential function* $Q(\cdot)$ *must satisfy the following: if sites* $(s_i, s_j, \ldots, s_l)$ *do not form a clique, then* $G_{i,j,\ldots,l}(\cdot) \equiv 0$, *where cliques are defined by the neighbourhood*

structure $\{\mathcal{N}_i : i = 1, \ldots, N\}$.

In other words, the Hammersley-Clifford theorem states that conditional spatial models can be specified in terms of a few non-zero functions $G_{i,j,\ldots,l}(\cdot)$. The results of the theorem are perhaps clearer when we can specify the negative potential function as follows:

Let κ be a clique and define $\boldsymbol{y}_\kappa \equiv (y(\boldsymbol{s}_i) : i \in \kappa)^T$ and $V_\kappa(\boldsymbol{y}_\kappa) = \{\prod_{i\in\kappa} y(\boldsymbol{s}_i)\}\, G_\kappa(\boldsymbol{y}_\kappa)$. Then,

$$Q(\boldsymbol{y}) = \sum_{\kappa\in\mathscr{C}} V_\kappa(\boldsymbol{y}_\kappa).$$

where $\mathscr{C}$ is the set of all cliques. Thus, from Eq. 2.3 we find that

$$P\left(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}\right) \propto \exp\left(\sum_{\kappa:i\in\kappa} V_\kappa(\boldsymbol{y}_\kappa)\right) \text{ for } i = 1, 2, \ldots, N.$$

A corollary of the above result is obtained if ζ is countable in $\mathbb{R}^d$ or ζ is a Lebesgue measurable subset of $\mathbb{R}^d$, and a set of well-defined functions G can be specified from the conditional probabilities and neighbourhoods $\mathcal{N}_i : i = 1, \ldots, N$. Then, the resulting negative potential function yields a unique well-defined joint probability function provided that $\sum_{\boldsymbol{y}\in\zeta} \exp(Q(\boldsymbol{y})) < \infty$ holds [9]. In general, however, the partition function cannot be written in closed form as a function of the parameters of the negative potential function which makes maximising the exact likelihood computationally prohibitive [9].

2.3 Auto-logistic Model

We now have the appropriate foundation to specify the auto-logistic model that forms the basis of our proposed model. Since the probability mass function of Y has the form $P(\boldsymbol{y}) = \exp(Q(\boldsymbol{y}))/\sum_{\boldsymbol{z}\in\zeta}\exp(Q(\boldsymbol{z}))$, exponential family models can be trivially expressed in terms of the negative potential function. Furthermore, this allows us to specify pairwise-only dependence models for conditional exponential distributions [5]. First, note that, as shown in [9], the conditional distribution of

exponential family models can written as:

$$\begin{aligned} &P(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}) = \\ &\quad \exp\left\{A_i(\{y(\boldsymbol{s}_j) : j \in \mathscr{N}_i\})B_i(y(\boldsymbol{s}_i)) + D_i(\{y(\boldsymbol{s}_j) : j \in \mathscr{N}_i\}) + C_i(y(\boldsymbol{s}_i))\right\}, \end{aligned} \tag{2.6}$$

where for $i = 1, 2, \ldots, N$ we have A_i and D_i which depend on the neighbours of the i^{th} site and B_i and C_i which do not.

Now, assuming Eq. 2.6 and assuming pair-wise only dependence, we have

$$A_i(\{y(\boldsymbol{s}_j) : j \in \mathscr{N}_i\}) = \alpha_i + \sum_{j=1}^{N} \theta_{ij} B_j(y(\boldsymbol{s}_j)), \tag{2.7}$$

where $\theta_{ij} = \theta_{ji}$, $\theta_{ii} = 0$, and $\theta_{ik} = 0$ for $k \notin \mathscr{N}_i$ and $i = 1, 2, \ldots, N$.

Using the above, an auto-logistic model is defined as follows [5] [9]. Let $G_i(1) \equiv \alpha_i$ and $G_{ij}(1,1) \equiv \theta_{ij}$. Thus, the negative potential function in terms of α_i and θ_{ij} is:

$$Q(\boldsymbol{y}) = \sum_{i=1}^{N} \alpha_i y(\boldsymbol{s}_i) + \sum_{1 \leq i < j \leq N} \theta_{ij} y(\boldsymbol{s}_i) y(\boldsymbol{s}_j),$$

where $\theta_{ij} = 0$ if $j \notin \mathscr{N}_i$ and $\theta_{ii} = 0$ to maintain identifiability. Consequently, from Eq. 2.3 it follows that

$$Q(\boldsymbol{y}) - Q(\boldsymbol{y}_{0(\boldsymbol{s}_k)}) = \alpha_k y(\boldsymbol{s}_k) + \sum_{j=1}^{N} \theta_{kj} y(\boldsymbol{s}_k) y(\boldsymbol{s}_j).$$

Under the same assumptions, we can apply the above general result to the binary case where $y(\boldsymbol{s}_i) = 0$ or 1. First, note that we can express:

$$\frac{P(y(\boldsymbol{s}_k)|\{y(\boldsymbol{s}_j) : j \neq k\})}{P(0(\boldsymbol{s}_k)|\{y(\boldsymbol{s}_j) : j \neq k\})} = \exp\left\{\alpha_k y(\boldsymbol{s}_k) + \sum_{j=1}^{N} \theta_{kj} y(\boldsymbol{s}_k) y(\boldsymbol{s}_j)\right\}.$$

And similarly,

$$\frac{1 - P(0(\boldsymbol{s}_k)|\{y(\boldsymbol{s}_j) : j \neq k\})}{P(0(\boldsymbol{s}_k)|\{y(\boldsymbol{s}_j) : j \neq k\})} = \exp\left\{\alpha_k + \sum_{j=1}^{N} \theta_{kj} y(\boldsymbol{s}_j)\right\}.$$

After a bit of algebra, we reach the following form of the conditional probability:

$$P\left(y(\boldsymbol{s}_k)|\{y(\boldsymbol{s}_j) : j \neq k\}\right) = \frac{\exp\left\{\alpha_k y(\boldsymbol{s}_k) + \sum_{j=1}^{N} \theta_{kj} y(\boldsymbol{s}_k) y(\boldsymbol{s}_j)\right\}}{1 + \exp\left\{\alpha_k + \sum_{j=1}^{N} \theta_{kj} y(\boldsymbol{s}_j)\right\}}, \tag{2.8}$$

which is defined to be an auto-logistic model due to its similarity to the classical form of the logistic model.

Chapter 3

Current Work

In this section, we derive our proposed model. First, we adapt the model specified by Huffer and Wu [18] to more reasonably model telemetry data. Next, we deal with issues of identifiability in the proposed likelihood and adjust the model to alleviate these. We then show that the proposed likelihood is an auto-logistic model as defined in Section 2.3. After this, we derive the usual elements of the model needed for maximum likelihood estimation, namely, the log-likelihood, gradient, and information matrix. Following this, we extend the model to the multi-sample case where more than one lattice of spatial location data is available. Finally, we touch upon numerical issues and methods to resolve these in the multi-sample case.

3.1 Auto-logistic Model for Tracking Data

Our proposed method has a foundation in the unnormalised auto-logistic likelihood function put forth in [18] shown below (Eq. 3.1). In order to simplify the notation, we let $\boldsymbol{y} = (y_1, \ldots, y_N)^T$, where $y_i = y(\boldsymbol{s}_i) = \mathbb{I}(\text{presence was observed at site } i)$ for cells on a lattice or grid indexed by $i = 1, 2, \ldots, N$. Furthermore, $\boldsymbol{x}_i$ denotes a vector of the value of the explanatory variables observed at the i^{th} site. Finally, $j \in \mathcal{N}_i$ denotes sites neighbouring site i based on a pre-defined neighbourhood structure. In our model, the first-order neighbours in any direction are considered neighbours.

$$f(\boldsymbol{y}) \propto \exp\left\{ \sum_{i=1}^{N} [y_i \beta_0 + y_i \boldsymbol{x}_i \boldsymbol{\beta}_1] + (\gamma/2) \sum_{j \in \mathscr{N}_i} [y_i y_j + (1-y_i)(1-y_j)] \right\}. \quad (3.1)$$

Although the model above is a viable method for modelling binary data in a lattice, we want a more fine-tuned model which differentiates between occupied sites, where an animal has traversed, from unoccupied sites (i.e., unvisited by an animal) assigning a separate parameter to each as this may fit telemetry data more closely. Initial experimentation showed that the parameter specifying the spatial structure of the data

$$(\gamma/2) \sum_{i=1}^{N} \sum_{j \in \mathscr{N}_i} [y_i y_j + (1-y_i)(1-y_j)] = (\gamma/2) \sum_{i=1}^{N} y_i \sum_{j \in \mathscr{N}_i} y_j + (1-y_i) \sum_{j \in \mathscr{N}_i} (1-y_j),$$

did not differentiate between zones of spatially auto-correlated occupied sites and those of empty sites. For example, using Eq. 3.1, the two data grids shown in Fig. 3.1 would have the same estimated coefficient since the term

$$\sum_{i=1}^{N} \sum_{j \in \mathscr{N}_i} [y_i y_j + (1-y_i)(1-y_j)],$$

equals 14 in both cases. Table 3.1 breaks down the process with $\boldsymbol{y}_1$ calculations in the left column and $\boldsymbol{y}_2$ in the right column.

Given that our aim is to use telemetry data as the source of our presence/absence response variable, the difference between a track of occupied sites and a track of empty sites is critical. In order to make this distinction clear in the estimation procedure, we choose to split the spatial structure parameter into two distinct parameters. The first, θ_1, tracks series of occupied sites surrounded by empty sites. The second, θ_2, tracks zones composed entirely of empty sites.

Note that given our specification of the spatial parameters, we still meet the positivity condition since a site can turn to one or zero so long as there is at least a one (presence) or a zero (absence) surrounding it.

With our new parametrisation of the spatial structure parameters, the likelihood

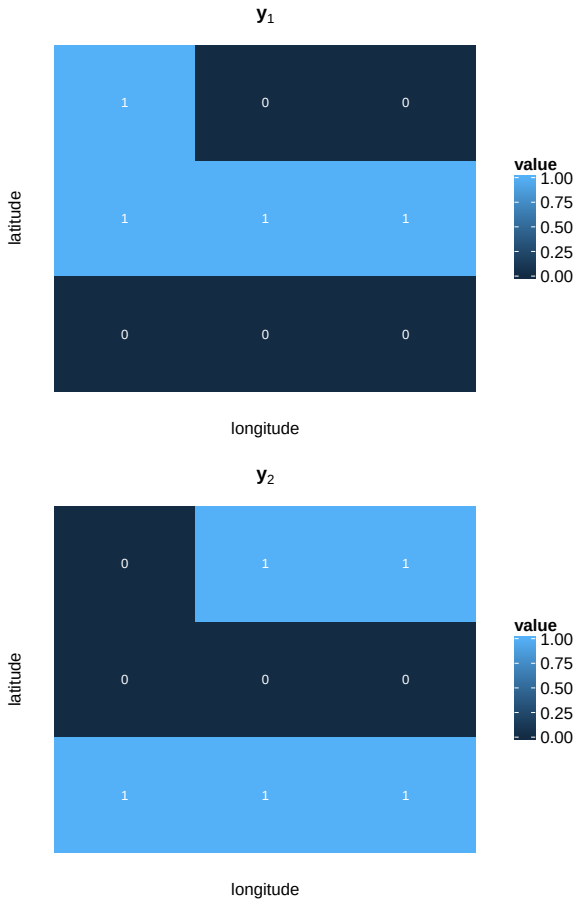

Figure 3.1: Two configurations of absence-presence locations on a lattice that yield equivalent spatial auto-correlation parameter value for the auto-logistic model in Eq. 3.1. The observed data $\boldsymbol{y}_1$ (top) and $\boldsymbol{y}_2$ (bottom) are used in the example shown in Table 3.1.

becomes:

$$f(\boldsymbol{y}) \propto \exp\left\{\sum_{i=1}^{N} y_i\beta_0 + y_i\boldsymbol{x}_i\boldsymbol{\beta}_1 + \theta_1 \sum_{j\in\mathcal{N}_i} y_i(1-y_j) + \theta_2 \sum_{j\in\mathcal{N}_i} (1-y_i)(1-y_j)\right\}. \tag{3.2}$$

The likelihood shown in Eq. 3.2 has an issue of identifiability where the in-

tercept term is not identifiable due to the form of the spatial structure parameters. Dropping the intercept term, we are left with the following likelihood whose identifiability we show next:

$$f(\boldsymbol{y}) = \exp\left\{\sum_{i=1}^{N} y_i \boldsymbol{x}_i \boldsymbol{\beta} + \theta_1 \sum_{j\in\mathcal{N}_i} y_i(1-y_j) + \theta_2 \sum_{j\in\mathcal{N}_i} (1-y_i)(1-y_j)\right\} \Big/ z(\boldsymbol{\eta}), \tag{3.3}$$

where $\boldsymbol{\eta} = (\boldsymbol{\beta}, \theta_1, \theta_2)^T$.

3.2 Identifiability of The Likelihood

We now verify the identifiability of the likelihood. It is clear that under some weak assumptions on the $\boldsymbol{x}_i$'s, such as all of the $\boldsymbol{x}_i$ not being equal to each other, terms depending on the parameter $\boldsymbol{\beta}$ are identifiable due to their dependence on the particular covariate being modelled. Hence, to show the identifiability of the model, we must show that the spatial auto-correlation parameters are identifiable.

Showing identifiability of the model is trivial and involves a simple expansion of the spatial auto-correlation parameters.

$$\begin{aligned}
&\theta_1 \sum_i \sum_{j\in\mathcal{N}_i} (y_i - y_i y_j) + \theta_2 \sum_i \sum_{j\in\mathcal{N}_i} (1-y_i)(1-y_j) \\
&= \theta_1 \sum_i y_i |\mathcal{N}_i| + (\theta_2 - \theta_1) \sum_i \sum_{j\in\mathcal{N}_i} y_i y_j + \theta_2 \left[\sum_i |\mathcal{N}_i| - \sum_i \sum_{j\in\mathcal{N}_i} y_j - \sum_i y_i |\mathcal{N}_i|\right] \\
&= (\theta_2 - \theta_1) \sum_i \sum_{j\in\mathcal{N}_i} y_i y_j + (\theta_1 - \theta_2) \sum_i y_i |\mathcal{N}_i| + \theta_2 \sum_i |\mathcal{N}_i| - \theta_2 \sum_i \sum_{j\in\mathcal{N}_i} y_j
\end{aligned}$$

where $|\mathcal{N}_i|$ denotes the cardinality of the neighbours of the i^{th} site, and $\sum_i$ is a shorthand notation for $\sum_{i=1}^{N}$. Note that the number of neighbours depends on the location of the site and the given neighbourhood structure. For example, for the first-order neighbourhood structure we use, we have eight neighbours for sites not on the edge of the lattice, five neighbours for sites on the edge but not on a corner, and three neighbours for sites on the corners. Finally, from the expansion above, we can see that there are four terms which contribute to the likelihood in addition

to the term containing the covariate. Since each of the four terms has a distinct sufficient statistic associated each with its own parameter (or a difference between parameters), it is clear that the model is identifiable for distinct parameter values.

3.3 Likelihood as an Auto-logistic Model

Having shown the identifiability of the model, we now show that the proposed likelihood is an auto-logistic model as previously specified and, thus, can be expressed as the product of conditional probabilities.

First, recall from Def. 5 that the probability mass function of a Markov random field can be written in terms of the negative potential function and the partition function. We denote the (exponentiated) negative potential function as:

$$\exp\{Q(\boldsymbol{y})\} = \exp\left\{\sum_{i=1}^{N} y_i \boldsymbol{x}_i\boldsymbol{\beta} + \theta_1 \sum_{j\in\mathcal{N}_i} y_i(1-y_j) + \theta_2 \sum_{j\in\mathcal{N}_i} (1-y_i)(1-y_j)\right\}. \tag{3.4}$$

In our case, the partition function is just a normalising constant defined as the sum over all possible configurations of the data. Consequently, we denote the normalising constant $z(\boldsymbol{\eta})$ as a function of the parameters of the model $\boldsymbol{\eta}$. Note that computation of $z(\boldsymbol{\eta})$ is usually intractable due to the dimension of the data.

Next, fix k and let $\boldsymbol{y}_{0_k} = (y_1, \ldots, y_{k-1}, 0, y_{k+1}, \ldots, y_N)^T$ and using Eq. 2.3, we take the difference of the negative potential functions at $\boldsymbol{y}$ and $\boldsymbol{y}_{0_k}$ and exponentiate these to find:

$$\frac{P(\boldsymbol{y})}{P(\boldsymbol{y}_{0_k})} = \exp\Big\{\sum_i y_i\boldsymbol{x}_i\boldsymbol{\beta} - \sum_{i\neq k} y_i\boldsymbol{x}_i\boldsymbol{\beta} + \theta_1\sum_i\sum_j y_i(1-y_j) - \theta_1\sum_{i\neq k}\sum_j y_i(1-y_j) + \theta_2\sum_i\sum_j (1-y_i)(1-y_j) - \theta_2\sum_{i\neq k}\sum_j (1-y_i)(1-y_j) - \theta_2\sum_{j\in\mathcal{N}_k}(1-y_j)\Big\}.$$

We now simplify the equation by parts. For the terms containing the parameter

β we have:

$$\sum_i y_i \boldsymbol{x}_i \boldsymbol{\beta} - \sum_{i \neq k} y_i \boldsymbol{x}_i \boldsymbol{\beta} = y_k \boldsymbol{x}_k \boldsymbol{\beta}. \tag{3.5}$$

Next, for the terms containing the parameter θ_1, we break the summation into those sites not neighbouring site k, those which have site k as a neighbour and the contribution from the neighbours of these, and the neighbours of site k.

$$\begin{aligned}
&\theta_1 \sum_i \sum_j y_i(1-y_j) - \theta_1 \sum_{i \neq k} \sum_j y_i(1-y_j) \\
&\quad = \theta_1 \sum_{i \notin \mathcal{N}_k} \sum_{j \in \mathcal{N}_i} y_i(1-y_j) + \theta_1 \sum_{i \in \mathcal{N}_k} \sum_{j \in \mathcal{N}_i} y_i(1-y_j) + \theta_1 \sum_{j \in \mathcal{N}_k} y_k(1-y_j) \\
&\quad - \theta_1 \sum_{i \notin \mathcal{N}_k} \sum_{j \in \mathcal{N}_i} y_i(1-y_j) - \theta_1 \sum_{i \in \mathcal{N}_k} \sum_{j \in \mathcal{N}_i, j \neq k} y_i(1-y_j) - \theta_1 \sum_{i \in \mathcal{N}_k} y_i \\
&\quad = \theta_1 \sum_{i \in \mathcal{N}_k} \sum_{j \in \mathcal{N}_i} y_i(1-y_j) + \theta_1 \sum_{j \in \mathcal{N}_k} y_k(1-y_j) - \theta_1 \sum_{i \in \mathcal{N}_k} \sum_{j \in \mathcal{N}_i, j \neq k} y_i(1-y_j) \\
&\quad - \theta_1 \sum_{i \in \mathcal{N}_k} y_i.
\end{aligned} \tag{3.6}$$

Note that for sites in the neighbourhood of the k^{th} site, we must take into account the value of the k^{th} site. Moreover, note that:

$$\theta_1 \sum_{i \in \mathcal{N}_k} \sum_{j \in \mathcal{N}_i} y_i(1-y_j) = \theta_1 \sum_{i \in \mathcal{N}_k} y_i \left(\sum_{j \in \mathcal{N}_i, j \neq k} (1-y_j) + (1-y_k) \right).$$

Hence, Eq. 3.6 simplifies to

$$\begin{aligned}
&\theta_1 \sum_{i \in \mathcal{N}_k} \sum_{j \in \mathcal{N}_i, j \neq k} y_i(1-y_j) + \theta_1 \sum_{i \in \mathcal{N}_k} y_i(1-y_k) \\
&\qquad + \theta_1 \sum_{j \in \mathcal{N}_k} y_k(1-y_j) - \theta_1 \sum_{i \in \mathcal{N}_k} \sum_{j \in \mathcal{N}_i, j \neq k} y_i(1-y_j) - \theta_1 \sum_{i \in \mathcal{N}_k} y_i \\
&\qquad = -\theta_1 \sum_{i \in \mathcal{N}_k} y_i y_k + \theta_1 \sum_{j \in \mathcal{N}_k} y_k(1-y_j) = \theta_1 \sum_{j \in \mathcal{N}_k} (y_k - 2 y_j y_k).
\end{aligned}$$

For the terms containing the parameter θ_2, we similarly break down the summation based on the neighbours of the k^{th} site.

$$
\begin{aligned}
&\theta_2 \sum_i \sum_j (1-y_i)(1-y_j) - \theta_2 \sum_{i\neq k} \sum_j (1-y_i)(1-y_j) \\
&\quad = \theta_2 \sum_{i\notin\mathcal{N}_k} \sum_{j\in\mathcal{N}_i} (1-y_i)(1-y_j) - \theta_2 \sum_{i\notin k} \sum_{j\in\mathcal{N}_i} (1-y_i)(1-y_j) \\
&\quad + \theta_2 \sum_{i\in\mathcal{N}_k} \sum_{j\in\mathcal{N}_i} (1-y_k)(1-y_j) + \theta_2 \sum_{j\in\mathcal{N}_k} (1-y_k)(1-y_j) \\
&\quad - \theta_2 \sum_{i\in\mathcal{N}_k} \sum_{j\in\mathcal{N}_i} (1-y_i)(1-y_j) - \theta_2 \sum_{j\in\mathcal{N}_k} (1-y_k)(1-y_j).
\end{aligned} \tag{3.7}
$$

As was the case with θ_1, we can simplify sites which have site k as a neighbour as:

$$
\theta_2 \sum_{i\in\mathcal{N}_k} \sum_{j\in\mathcal{N}_i} (1-y_i)(1-y_j) = \theta_2 \sum_{i\in\mathcal{N}_k} \sum_{j\in\mathcal{N}_i, j\neq k} (1-y_i)(1-y_j) + \theta_2 \sum_{i\in\mathcal{N}_k} (1-y_i)(1-y_k).
$$

Hence, Eq. 3.7 simplifies to

$$
\begin{aligned}
&\theta_2 \sum_{i\in\mathcal{N}_k} \sum_{j\in\mathcal{N}_i, j\neq k} (1-y_i)(1-y_j) + \theta_2 \sum_{i\in\mathcal{N}_k} (1-y_i)(1-y_k) + \theta_2 \sum_{j\in\mathcal{N}_k} (1-y_k)(1-y_j) \\
&\quad - \theta_2 \sum_{i\in\mathcal{N}_k} \sum_{j\in\mathcal{N}_i, j\neq k} (1-y_i)(1-y_j) - \theta_2 \sum_{i\in\mathcal{N}_k} (1-y_i) - \theta_2 \sum_{j\in\mathcal{N}_k} (1-y_j) \\
&\quad = 2\theta_2 \sum_{j\in\mathcal{N}_k} (1-y_k)(1-y_j) - 2\theta_2 \sum_{j\in\mathcal{N}_k} (1-y_j).
\end{aligned}
$$

Therefore, taking the simplified forms of Eq. 3.5, Eq. 3.6, and Eq. 3.7, we find the ratio of the conditional probabilities at the k^{th} site given all other sites to be:

$$
\begin{aligned}
\frac{P(\boldsymbol{y})}{P(\boldsymbol{y}_{0_k})} &= \exp\left\{ y_k \boldsymbol{x}_k \boldsymbol{\beta} + \theta_1 \left(\sum_{j\in\mathcal{N}_k} y_k(1-y_j) - y_j y_k \right) \right. \\
&\qquad \left. + 2\theta_2 \left(\sum_{j\in\mathcal{N}_k} (1-y_k)(1-y_j) - (1-y_j) \right) \right\} \\
&= \frac{P(y_k | \{y_j : j\neq k\}, \boldsymbol{x}_k)}{P(0_k | \{y_j : j\neq k\}, \boldsymbol{x}_k)},
\end{aligned}
$$

where $0_k \equiv 0(\boldsymbol{s}_k)$.

Furthermore, since our data are binary we note that

$$\frac{1-P(0_k|\{y_j : j \neq k\},\boldsymbol{x}_k)}{P(0_k|\{y_j : j \neq k\},\boldsymbol{x}_k)} = \exp\left\{\boldsymbol{x}_k\boldsymbol{\beta} + \theta_1 \sum_{j\in\mathcal{N}_k}(1-2y_j) - 2\theta_2 \sum_{j\in\mathcal{N}_k}(1-y_j)\right\},$$

so that we arrive at the conditional likelihood of the k^{th} site given all other sites to be:

$$P(y(\boldsymbol{s}_k)|\{y(\boldsymbol{s}_j) : j \neq k\},\boldsymbol{x}_k) =$$
$$\frac{\exp\left\{y_k\boldsymbol{x}_k\boldsymbol{\beta} + \theta_1\left(\sum_{j\in\mathcal{N}_k} y_k(1-y_j) - y_jy_k\right) + 2\theta_2\left(\sum_{j\in\mathcal{N}_k}(1-y_k)(1-y_j) - (1-y_j)\right)\right\}}{1+\exp\left\{\boldsymbol{x}_k\boldsymbol{\beta} + \theta_1\sum_{j\in\mathcal{N}_k}(1-2y_j) - 2\theta_2\sum_{j\in\mathcal{N}_k}(1-y_j)\right\}},$$

which is an auto-logistic model as previously defined. Furthermore, it is worth noting that the conditional probability of observing presence, that is, $y_k = 1$ at any site k, is simply an inverse-logit function of a linear function $g_k(\cdot)$ of the covariates $\boldsymbol{x}_k$ at a location y_k:

$$P(y_k = 1|\{y_j : j \neq k\},\boldsymbol{x}_k) = \frac{\exp\{g_k(\boldsymbol{x}_k)\}}{1+\exp\{g_k(\boldsymbol{x}_k)\}}, \tag{3.8}$$

where

$$g_k(\boldsymbol{x}_k) = \boldsymbol{x}_k\boldsymbol{\beta} + \theta_1 \sum_{j\in\mathcal{N}_k}(1-2y_j) - 2\theta_2 \sum_{j\in\mathcal{N}_k}(1-y_j) \tag{3.9}$$

By the factorisation theorem (Theorem 2.1.1), it follows that we can estimate the parameters of the joint likelihood as the product of the conditional probabilities at each site. Since each conditional probability is essentially a logistic function, estimation of the parameters in the factorised likelihood can be trivially achieved via logistic regression. This estimate, known as the Maximum Pseudo-Likelihood Estimate (MPLE), has the advantage of not requiring one to compute the normalising constant of the joint likelihood. Therefore, the MPLE serves as the starting point in our MCMC-MLE optimisation procedure described later in Section 4.1.

3.4 Estimation of the Joint Likelihood

We now examine the estimation of the joint likelihood via Monte Carlo methods. Since in order to estimate model parameters a tractable form of the likelihood function is needed, we follow the material presented in [18] and [12] in examining available methods to obtain, or at least estimate, a normalised joint likelihood function to optimise.

By definition, one can decompose the normalised likelihood function into the two components shown below: the unnormalised likelihood, $h_\eta(\boldsymbol{y})$, and the normalising constant, $z(\boldsymbol{\eta})$:

$$f(\boldsymbol{y},\boldsymbol{\eta}) = \frac{1}{z(\boldsymbol{\eta})} h_\eta(\boldsymbol{y}),$$

Here, the unnormalised likelihood is the exponentiated negative potential function from Eq. 3.4:

$$\begin{aligned} h_\eta(\boldsymbol{y}) &= \exp\{Q(\boldsymbol{y})\} \\ &= \exp\left\{ \sum_{i=1}^{N} y_i \boldsymbol{x_i}\boldsymbol{\beta} + \theta_1 \sum_{j\in\mathcal{N}_i} y_i(1-y_j) + \theta_2 \sum_{j\in\mathcal{N}_i} (1-y_i)(1-y_j) \right\}, \end{aligned} \tag{3.10}$$

and the normalising constant is

$$z(\boldsymbol{\eta}) = \int h_\eta(\boldsymbol{y}) d\mu,$$

where μ is the Lebesgue measure on $\mathbb{R}$.

As is common in the field of spatial statistics, direct computation of the normalising constant is intractable. In our case, computing the normalising constant is particularly difficult due to the constant being a function of our parameter to be estimated.

In estimating the normalising constant, we proceed as suggested in [18]. Rather than estimating the normalising constant on its own, we introduce an importance distribution and estimate the ratio of the normalising constant of the original likelihood over that of the importance distribution. Therefore, we estimate the ratio of the likelihood with the parameters to be estimated $\boldsymbol{\eta}$ over a likelihood with ar-

bitrarily chosen importance parameters $\boldsymbol{\eta}^* = (\boldsymbol{\beta}^*, \theta_1^*, \theta_2^*)^T$. Under this construct, maximising the likelihood $f(\boldsymbol{y}, \boldsymbol{\eta})$ is equivalent to maximising the ratio of the two likelihoods:

$$\arg\max_{\boldsymbol{\eta}} f(\boldsymbol{y}, \boldsymbol{\eta}) = \arg\max_{\boldsymbol{\eta}} \frac{f(\boldsymbol{y}, \boldsymbol{\eta})}{f(\boldsymbol{y}, \boldsymbol{\eta}^*)},$$

where

$$\frac{f(\boldsymbol{y}, \boldsymbol{\eta})}{f(\boldsymbol{y}, \boldsymbol{\eta}^*)} = \frac{h_\eta(\boldsymbol{y})}{h_{\eta^*}(\boldsymbol{y})} \cdot \frac{z(\boldsymbol{\eta}^*)}{z(\boldsymbol{\eta})} = \frac{h_\eta(\boldsymbol{y})}{\frac{z(\boldsymbol{\eta})}{z(\boldsymbol{\eta}^*)} h_{\eta^*}(\boldsymbol{y})}. \tag{3.11}$$

Now, instead of estimating the normalising constant directly, we estimate the ratio of the normalising constants for the two likelihoods. Estimating the ratio of the normalising constant is easier since

$$\frac{z(\boldsymbol{\eta})}{z(\boldsymbol{\eta}^*)} = \int \frac{h_\eta(\boldsymbol{y})}{z(\boldsymbol{\eta}^*)} f_{\eta^*}(\boldsymbol{y}) d\mu(\boldsymbol{y}) = \mathbb{E}_{\eta^*}\left[\frac{h_\eta(\boldsymbol{Y})}{h_{\eta^*}(\boldsymbol{Y})}\right] \approx \frac{1}{M} \sum_{m=1}^{M} \frac{h_\eta(\boldsymbol{y}_m^*)}{h_{\eta^*}(\boldsymbol{y}_m^*)}, \tag{3.12}$$

where $\boldsymbol{y}_m^*$ is a sample from $f(\cdot, \boldsymbol{\eta}^*)$. Substituting Eq. 3.12 into Eq. 3.11, we find:

$$\frac{f(\boldsymbol{y}, \boldsymbol{\eta})}{f(\boldsymbol{y}, \boldsymbol{\eta}^*)} \approx \frac{h_\eta(\boldsymbol{y}_{obs}) / h_{\eta^*}(\boldsymbol{y}_{obs})}{\frac{1}{M} \sum_{m=1}^{M} h_\eta(\boldsymbol{y}_m^*) / h_{\eta^*}(\boldsymbol{y}_m^*)}. \tag{3.13}$$

Note that by making use of the ergodic theorem [10], the expected value we wish to find can be computed from a non i.i.d.sample $\boldsymbol{y}_m^*$. Therefore, for our estimation we sample pseudo-data generated via a Gibbs-sampler (i.e., MCMC) from $\boldsymbol{Y}^* \sim f(\cdot, \boldsymbol{\eta}^*)$. The arbitrarily chosen $\boldsymbol{\eta}^*$ ought to be close to the true parameter $\boldsymbol{\eta}$ so that the maximisation procedure works adequately, and the obtained estimate $\hat{\boldsymbol{\eta}}_{MLE}$ is accurate. Hence, as previously mentioned, we set the MPLE as the initial value for $\boldsymbol{\eta}^*$.

For conciseness, we can summarise the observed data into a vector of sufficient statistics, $t(\boldsymbol{y})$, such that

$$t(\boldsymbol{y}) = \left(\sum_i y_i \boldsymbol{x}_i, \sum_i \sum_{j \in \mathcal{N}_i} y_i (1 - y_j), \sum_i \sum_{j \in \mathcal{N}_i} (1 - y_i)(1 - y_j) \right)^T.$$

Using $t(\boldsymbol{y})$, we can approximate our proposed likelihood in the form of Eq. 3.13 as a function of $\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\eta}$, and $\boldsymbol{\eta}^*$:

$$L(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{\eta}^*) \approx \frac{\exp\{(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T t(\boldsymbol{y}_{obs})\}}{\frac{1}{M}\sum_m \exp\{(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T t(\boldsymbol{y}^*_m)\}}, \tag{3.14}$$

where $\boldsymbol{y}_{obs}$ are the observed data and $\boldsymbol{y}^*_m$ is the m^{th} sample of pseudo-data generated via MCMC for $m = 1, \ldots, M$. Similarly, we can derive the log-likelihood as:

$$\ell(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{\eta}^*) \approx (\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T t(\boldsymbol{y}_{obs}) - \log\left(\frac{1}{M}\sum_m \exp\left\{(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T t(\boldsymbol{y}^*_m)\right\}\right).$$

Now that we have an estimate for the log-likelihood function in closed form, we can compute the maximum likelihood estimates as usual. First, we compute the gradient of the log-likelihood function. That is, we compute the vector of partial derivatives where the partial derivative of the log-likelihood with regards to the j^{th} parameter is:

$$\frac{\partial \ell(\boldsymbol{\eta})}{\partial \eta_j} = t_j(\boldsymbol{y}_{obs}) - \frac{\sum_{m=1}^{M} \exp\{(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T t(\boldsymbol{y}^*_m)\} t_j(\boldsymbol{y}^*_m)}{\sum_{m=1}^{M} \exp\{(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T t(\boldsymbol{y}^*_m)\}} = t_j(\boldsymbol{y}_{obs}) - \sum_{m=1}^{M} w_m t_j(\boldsymbol{y}^*_m). \tag{3.15}$$

Note that we use

$$w_m = \frac{\exp\{(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T t(\boldsymbol{y}^*_m)\}}{\sum_{m=1}^{M} \exp\{(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T t(\boldsymbol{y}^*_m)\}},$$

to denote the weight of the m^{th} MCMC sample.

Similarly, we compute the Hessian matrix of second-order partial derivatives:

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\eta})}{\partial \eta_i \eta_j} =& -\frac{\sum_{m=1}^{M} \exp\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T t(\mathbf{y}_m^*)\} t_i(\mathbf{y}_m^*) t_j(\mathbf{y}_m^*)^T}{\sum_{m=1}^{M} \exp\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T t(\mathbf{y}_m^*)\}} \\ &+ \frac{\left(\sum_{m=1}^{M} \exp\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T t(\mathbf{y}_m^*)\} t_i(\mathbf{y}_m^*)\right)\left(\sum_{m=1}^{M} \exp\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T t(\mathbf{y}_m^*)\} t_j(\mathbf{y}_m^*)^T\right)}{\left(\sum_{m=1}^{M} \exp\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T t(\mathbf{y}_m^*)\}\right)^2} \\ =& -w_m t_i(\mathbf{y}_m^*) t_j(\mathbf{y}_m^*)^T + w_m^2 t_i(\mathbf{y}_m^*) t_j(\mathbf{y}_m^*)^T.\end{aligned} \tag{3.16}$$

The MLE is found by maximising the log-likelihood, which is equivalent to finding the value of $\boldsymbol{\eta}$ where the gradient is zero and the Hessian is negative definite.

Finally, standard errors of the MLE estimates can then be computed via the Fisher information defined as the expectation of the negative Hessian of the log-likelihood:

$$I(\boldsymbol{\eta}_0) = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \boldsymbol{\eta}^T}\right].$$

However, we estimate the Fisher information by taking the expectation of the negative Hessian evaluated at the MLE. Thus, the observed Fisher information is:

$$\hat{I}(\boldsymbol{\eta}_0) = -\left.\frac{\partial^2 \ell(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \boldsymbol{\eta}^T}\right|_{\boldsymbol{\eta}=\hat{\boldsymbol{\theta}}} = \mathbb{E}_{\hat{\theta}}\left[w_m t_i(\mathbf{y}_m^*) t_j(\mathbf{y}_m^*)^T - w_m^2 t_i(\mathbf{y}_m^*) t_j(\mathbf{y}_m^*)^T\right]. \tag{3.17}$$

3.5 Multi-sample Joint Likelihood

Thus far, we have focused on the case described in [18], where we have only a single lattice or grid of data representing the sites $\boldsymbol{y}$ where there is either an observed presence or absence. Now, we move on to the multi-sample case where the data consists of a set of grids. Each of these grids can vary in terms of presence or absence at a site, the covariate values at sites, etc. However, we can consider all these grids as having been generated from a common true stochastic process.

In the multi-sample case, we have K grids of available data each of which we assume to have been discretised to the same resolution of $d \times d$ sites. We assume that the grids, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K$, are all independent and identically distributed and take their multi-sample joint likelihood to be the product of the individual K likelihoods.

$$L(\boldsymbol{\eta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_K, \boldsymbol{\eta}^*) = \prod_{k=1}^{K} L(\boldsymbol{\eta}; \boldsymbol{y}_k, \boldsymbol{\eta}^*). \tag{3.18}$$

Recalling Eq. 3.11, one can re-write the multi-sample joint likelihood as:

$$\prod_{k=1}^{K} L(\boldsymbol{\eta}; \boldsymbol{y}_k, \boldsymbol{\eta}^*) = \prod_{k=1}^{K} \frac{h_\eta(\boldsymbol{y}_{obs,k}) / h_{\eta^*}(\boldsymbol{y}_{obs,k})}{z(\boldsymbol{\eta}) / z(\boldsymbol{\eta}^*)}.$$

When approximating the ratio of the normalising constants $z(\boldsymbol{\eta})/z(\boldsymbol{\eta}^*)$ in Eq. 3.12, one estimates the integration over the entire sample space by making use of M generated pseudo-grids (each of dimension $d \times d$). It is important to note, however, that in the multi-sample case estimating the ratio of the normalising constants jointly requires an integration over M random vectors of K grids each. Therefore, in computing the multi-sample joint likelihood, the integration in the measure $\mu(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K)$ no longer corresponds to integrating over all possible configurations of a single grid $\boldsymbol{y}$ but rather to all possible configurations of a joint set of grids. This joint set, however, is itself a function of the possible configurations for each of the K individual grids $t(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K)$ so that

$$\frac{z(\boldsymbol{\eta})}{z(\boldsymbol{\eta}^*)} = \int \ldots \int \frac{h_\eta(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K)}{z(\boldsymbol{\eta}^*)} f_{\eta^*}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K) d\mu.$$

In particular, note that the function of sufficient statistics $t(\cdot)$ is applied to each of the K observed samples as well as the M pseudo-data samples generated.

Under our parametrisation of the likelihood, where the likelihood is proportional to an exponentiated linear term, the joint ratio of the normalising constants can be estimated as follows. First, we add the sufficient statistics across all K distinct grids for the m^{th} pseudo-sample. Next, each of these summed sufficient statistics is used to evaluate the ratio of the unnormalised likelihoods at a given $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$. Finally, we take the average of these evaluated ratios of the unnormalised

joint likelihood as an approximation for their expectation.

$$\frac{z(\boldsymbol{\eta})}{z(\boldsymbol{\eta}^*)} = \mathbb{E}_{\eta^*}\left[\frac{h_\eta(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K)}{h_{\eta^*}(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K)}\right] \approx \frac{1}{M}\sum_m \exp\left\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T \sum_k t(\boldsymbol{y}^*_{k,m})\right\}. \quad (3.19)$$

Having derived an estimate for the ratio of the normalising constants in the multi-sample case, the multi-sample joint likelihood can be expressed as:

$$\begin{aligned}\prod_{k=1}^K L(\boldsymbol{\eta};\boldsymbol{y}_k,\boldsymbol{\eta}^*) \approx & \frac{\prod_{k=1}^K \exp\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T t(\boldsymbol{y}_{k,obs})\}}{\frac{1}{M}\sum_m \exp\left\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T \sum_{k=1}^K t(\boldsymbol{y}^*_{k,m})\right\}} \\ = & \frac{\exp\left\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T \sum_{k=1}^K t(\boldsymbol{y}_{k,obs})\right\}}{\frac{1}{M}\sum_m \exp\left\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T \sum_{k=1}^K t(\boldsymbol{y}^*_{k,m})\right\}}.\end{aligned} \quad (3.20)$$

Naturally, the multi-sample joint log-likelihood is:

$$\begin{aligned}&\ell(\boldsymbol{\eta};\boldsymbol{y}_1,\ldots,\boldsymbol{y}_K,\boldsymbol{\eta}^*) \approx \\ &\quad (\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T \sum_{k=1}^K t(\boldsymbol{y}_{k,obs}) - \log\left(\frac{1}{M}\sum_m \exp\left\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T \sum_{k=1}^K t(\boldsymbol{y}^*_{k,m})\right\}\right).\end{aligned} \quad (3.21)$$

Since the multi-sample joint likelihood has the same form as the univariate joint likelihood, the gradient and Fisher information matrix necessary for maximum likelihood estimation using the multi-sample joint likelihood are almost identical to the univariate case except that the sufficient statistics $t(\boldsymbol{y}_{obs})$ are replaced by the sum across the K grids $\sum_{k=1}^K t(\boldsymbol{y}_{k,obs})$. Besides this difference, derivatives are identical to those shown in Eq. 3.15 and Eq. 3.17 and, hence, are not presented here.

Although the analytical form of the multi-sample joint likelihood derived above can be used during the estimation procedure, experimentation showed this particular analytical form to be numerically unstable given the magnitude of the terms being exponentiated. Consequently, forms of the multi-sample joint likelihood which are more numerically stable were used in our implementation. These forms

are presented in Section 4.2

3.6 Interpretation of Model Parameters

To close this section, the interpretation of the model parameters is discussed. There are three broad types of parameters in the model, namely, parameters that relate to the covariates $\boldsymbol{\beta}$, the spatial auto-correlation parameter for occupied sites surrounded by empty sites θ_1, and the spatial auto-correlation parameter for tracking empty sites surrounded by empty sites θ_2. We examine each of these below by way of examining their effect on the conditional likelihood in Eq. 3.8.

First, note that since each parameter has an inverse-logit relationship to the probability of observing presence at a site k, interpretation of all the model parameters can be made in terms of the log-odds as in logistic regression. More specifically, given the value of its neighbours, the probability of observing presence at the k^{th} site is an inverse-logit transform of the function $g_k(\boldsymbol{x}_k)$ shown in Eq. 3.9.

We can observe that the covariate parameters $\boldsymbol{\beta}$ are encompassed within the term $\boldsymbol{x}_k\boldsymbol{\beta}$ of the function $g_k(\cdot)$ in the conditional likelihood. Since $\boldsymbol{x}_k$ refers to the vector of covariate values observed at the k^{th} site, an individual parameter β is interpreted as the expected change in the log-odds of observing a presence at the k^{th} site due to a one-unit increase in the value of the covariate x accounting for the presence or absence of observations at locations neighbouring the k^{th} site.

Although interpretation of the spatial auto-correlation parameters follows similarly, parameter interpretation depends on the neighbouring structure. First, changing one of the neighbours of the k^{th} site from absent to present (i.e., changing one $y_j = 0$ to $y_j = 1$) denotes an increase of $2(\theta_2 - \theta_1)$ in the log-odds of observing a presence at the k^{th} site. On the other hand, changing one of the neighbours of the k^{th} site from present to absent denotes an increase of $2(\theta_1 - \theta_2)$ in the log-odds of observing a presence at the k^{th} site. Note that since changing the value of one neighbour affects the statistics associated with both spatial auto-correlation parameters θ_1 and θ_2, we must interpret a change in the log-odds as resulting from the difference between these two parameters at any one time. That is, we cannot interpret a change in the log-odds associated with one auto-correlation parameter without accounting for the change produced by the other.

Finally, we note that the spatial-auto correlation parameters can be interpreted as being auto-regressive [16]. That is, although two sites may not be neighbouring, the strength of the spatial auto-correlation among sites may lead the two non-neighbouring sites to affect one another by way of a diffusion process transferring from site to site. This diffusion is auto-regressive by way of the Markov process inherent in the model construction. Hence, the larger the magnitude of the difference between the parameters θ_1 and θ_2, the stronger the long-range dependence of the model. Consequently, when there is a strong underlying spatial auto-correlation, although the correlation between two sites may decrease as the distance between these sites increases, the correlation between the sites does not reach zero [16].

Table 3.1: Example of how the different configurations of occupied vs. non-occupied sites shown in Fig. 3.1 yield the same spatial auto-correlation parameter value for the auto-logistic model in Eq. 3.1.

Site	$\boldsymbol{y}_1$	$\sum_{j\in\mathcal{N}_i} y_i y_j$	$\sum_{j\in\mathcal{N}_i}(1-y_i)(1-y_j)$	$\boldsymbol{y}_2$	$\sum_{j\in\mathcal{N}_i} y_i y_j$	$\sum_{j\in\mathcal{N}_i}(1-y_i)(1-y_j)$
1	0	0	1	1	1	0
2	0	0	2	1	2	0
3	0	0	1	1	1	0
4	1	1	0	0	0	1
5	1	2	0	0	0	3
6	1	2	0	0	0	2
7	0	0	1	1	1	0
8	0	0	1	1	1	0
9	1	2	0	0	0	2
Total		8	6		6	8

Chapter 4

Implementation and Numerical Optimisation

In this section we discuss implementation details for our estimation procedure. First, we walk through the implementation for the Gibbs sampling algorithm used for the purpose of generating the pseudo-samples of observations. Afterwards, we discuss methods for maintaining numerical stability in estimating parameters for the multi-sample scenario.

4.1 Gibbs Sampler Implementation

The implementation of the Gibbs sampling algorithm for generating our pseudo-samples is shown below (Algo. 1).

4.2 Numerical Optimisation in the Multi-sample Scenario

In estimating parameters for the multi-sample scenario, a straight-forward implementation of the likelihood shown in Eq. 3.20 yielded numerical stability issues. Even with moderately sized data grids (e.g., lattices of dimensions 10×10 cells) and small parameter values ($\|\boldsymbol{\eta}\| \approx 10^{-1}$), the magnitude of the negative potential tended to be quite large. Consequently, calculation of the exponentiated negative potential function and the likelihood resulted in either over-flow or under-flow er-

Algorithm 1 One iteration of the Gibbs sampler algorithm for the auto-logistic model.

Input: Spatial locations $y_1, \ldots, y_N$, importance parameters $\boldsymbol{\eta}^*$, and (optionally) covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.

for $i = 1$ to N **do**
  Compute sufficient statistics $t(y_i) \leftarrow \left(y_i \boldsymbol{x}_i, \sum_{j \in \mathscr{N}_i} y_i(1-y_j), \sum_{j \in \mathscr{N}_i}(1-y_i)(1-y_j)\right)$.
  Compute negative potential $Q(y_i) \leftarrow t(y_i)\boldsymbol{\eta}^*$.
  Compute acceptance probability $\alpha_c \leftarrow \min\{1, \text{logit}^{-1}(Q(y_i))\}$.
  Sample $u_c \sim \text{Unif}(0,1)$.
  if $u_c < \alpha_c$ **then**
    $y_i^* \leftarrow 1$.
  else
    $y_i^* \leftarrow 0$.
  end if
  if $y_i^* \neq y_i$ **then**
    $t(y_i^*) \leftarrow \left(y_i^* \boldsymbol{x}_i, \sum_{j \in \mathscr{N}_i} y_i^*(1-y_j), \sum_{j \in \mathscr{N}_i}(1-y_i^*)(1-y_j)\right)$.
    for s in $\mathscr{N}_i$ in the proposed space $(y_1, \ldots, y_i^*, \ldots, y_N)$ **do**
      $t(y_s^*) \leftarrow \left(y_s \boldsymbol{x}_s, \sum_{j \in \mathscr{N}_s} y_s(1-y_j), \sum_{j \in \mathscr{N}_s}(1-y_s)(1-y_j)\right)$.
    end for
    for s in $\mathscr{N}_i$ the current space $(y_1, \ldots, y_i, \ldots, y_N)$ **do**
      $t(y_s) \leftarrow \left(y_s \boldsymbol{x}_s, \sum_{j \in \mathscr{N}_s} y_s(1-y_j), \sum_{j \in \mathscr{N}_s}(1-y_s)(1-y_j)\right)$.
    end for
    Let $t(\boldsymbol{y}_{i,s}) \leftarrow (t(y_i), t(y_{s \in \mathscr{N}_i}))$ and $t(\boldsymbol{y}_{i,s}^*) \leftarrow (t(y_i^*), t(y_{s \in \mathscr{N}_i}^*))$.
    Current neighbourhood-extended negative potential $Q(\boldsymbol{y}_{i,s}) \leftarrow t(\boldsymbol{y}_{i,s})\boldsymbol{\eta}^*$.
    Proposed neighbourhood-extended negative potential $Q(\boldsymbol{y}_{i,s}^*) \leftarrow t(\boldsymbol{y}_{i,s}^*)\boldsymbol{\eta}^*$.
    Compute joint acceptance ratio $r \leftarrow \exp\{Q(\boldsymbol{y}_{i,s}^*)\} / \exp\{Q(\boldsymbol{y}_{i,s})\}$.
    Compute acceptance probability $\alpha \leftarrow \min\{1, r\}$.
    Sample $u \sim \text{Unif}(0,1)$.
    if $u < \alpha$ **then**
      Accept proposal $y_i \leftarrow y_i*$.
    end if
  end if
end for

Output: Posterior pseudo-samples $\boldsymbol{y}^*$.

rors. It is worth noting that these issues could also have arised from phase transition as has been observed in the Markov random field [11], network modelling [1], and statistical physics literature [6]. We used the methods described in this section to alleviate some of these issues. Although the methods discussed aid numerical stability to an extent, scaling the multi-sample estimation to very large or very fine spatial areas or to parameters which are greater than 1 in magnitude remains an issue.

First, since maximising the likelihood with respect to the parameters $\boldsymbol{\eta}$ is equivalent whether or not the importance $\boldsymbol{\eta}^*$ parameters are included, we stabilise the multi-sample joint log-likelihood Eq. 3.21 by dropping the importance parameters in the un-normalised likelihood. Although not beneficial in all cases, dropping the importance parameters in the un-normalised likelihood can alleviate overflow issues arising when the importance parameters are far away from the true parameters. The stabilised likelihood then has the form:

$$\begin{aligned}&\ell(\boldsymbol{\eta};\boldsymbol{y}_1,\ldots,\boldsymbol{y}_K,\boldsymbol{\eta}^*)\\&\quad\approx \boldsymbol{\eta}^T\sum_{k=1}^{K} t(\boldsymbol{y}_{k,obs}) - \log\left(\frac{1}{M}\sum_{m}\exp\left\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T\sum_{k=1}^{K} t(\boldsymbol{y}^*_{k,m})\right\}\right).\end{aligned} \tag{4.1}$$

Next, rather than optimising the log-likelihood directly, we opt to use Newton-Raphson's method to minimise the gradient of the numerically stable version of the negative log-likelihood. That is, we find a point where the gradient is zero and the Hessian is positive definite. The Newton-Raphson algorithm is shown in Algo. 2. In order to further stabilise the optimisation procedure at this point, we further approximate the log-likelihood as follows. Since we have K distinct data grids, which we have assumed to be i.i.d., we exponentiate the ratio of the normalising constants for a single grid, denoted as $z_1(\boldsymbol{\eta})/z_1(\boldsymbol{\eta}^*)$ below, rather than approximating the ra-

tio of the normalising constants across the entire K-dimensional space.

$$\begin{aligned}\ell(\boldsymbol{\eta};\boldsymbol{y}_1,\ldots,\boldsymbol{y}_K,\boldsymbol{\eta}^*) &= \log\left(\prod_{k=1}^{K}\frac{h_\eta(\boldsymbol{y}_{obs,k})/h_{\eta^*}(\boldsymbol{y}_{obs,k})}{z(\boldsymbol{\eta})/z(\boldsymbol{\eta}^*)}\right)\\ &\approx \log\left(\frac{\prod_{k=1}^{K}h_\eta(\boldsymbol{y}_{obs,k})/h_{\eta^*}(\boldsymbol{y}_{obs,k})}{[z_1(\boldsymbol{\eta})/z_1(\boldsymbol{\eta}^*)]^K}\right)\\ &= \boldsymbol{\eta}^T\sum_{k=1}^{K}t(\boldsymbol{y}_{k,obs}) - K\log\left(\frac{1}{M}\sum_{m}\exp\left\{(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^T t(\boldsymbol{y}_m^*)\right\}\right).\end{aligned} \tag{4.2}$$

Hence, within the optimisation procedure, the first-order partial derivative of the multi-sample log-likelihood can be computed as:

$$\frac{\partial\ell(\boldsymbol{\eta};\boldsymbol{y}_1,\ldots,\boldsymbol{y}_K,\boldsymbol{\eta}^*)}{\partial\eta_j} = \boldsymbol{\eta}^T\frac{1}{K}\sum_{k=1}^{K}t_j(\boldsymbol{y}_{k,obs}) - \sum_{m=1}^{M}w_m t_j(\boldsymbol{y}_m^*),$$

We note here that since the accuracy of the estimated parameters depends heavily on the chosen importance parameters, we found that there is often a need to iterate through the entire estimation procedure multiple times. That is, using the parameter estimates obtained from the Newton-Raphson optimisation, an additional set of pseudo-samples may need to be generated via MCMC followed by another Newton-Raphson optimisation. This process may be repeated multiple times until convergence. Although there is no set method to determine convergence of the algorithm, we follow [12] and assume convergence after a few cycles of the MCMC and Newton-Raphson optimisation procedure. Our assumption is supported by the results of our computer experiments, where we see that, in general, the estimation of the MLE improves significantly between the first and second MCMC and Newton-Raphson optimisation cycles (see Fig. 5.3 – 5.8). Improvements in the estimation of the parameters after the second cycle of the MCMC and Newton-Raphson optimisation are much smaller with decreases in variance being accompanied with increases in bias.

Algorithm 2 Newton-Raphson algorithm used for optimisation. The threshold ε and step-size λ were set at 10^{-7} and 1, respectively.

Input: Data sets $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_K$, importance parameter vector $\boldsymbol{\eta}^*$ (same as $\boldsymbol{\eta}_{MPLE}$ at initialisation), threshold ε, and step-size λ.

1: **Initialize:**
   $\boldsymbol{\eta}^{(0)} \leftarrow \boldsymbol{\eta}^*$; $i \leftarrow 1$; $i_{max} \leftarrow 20$
2: **while** $i < i_{max}$ **and** $\delta^{(i-1)} > \varepsilon$ **do**
3: $\quad \boldsymbol{\eta}^{(i)} \leftarrow \boldsymbol{\eta}^{(i-1)} + \lambda \left(\left[\nabla^2 \ell(\boldsymbol{\eta}^{(i-1)}; \boldsymbol{y}_1,\ldots,\boldsymbol{y}_K,\boldsymbol{\eta}^*) \right]^{-1} \nabla \ell(\boldsymbol{\eta}^{(i-1)}; \boldsymbol{y}_1,\ldots,\boldsymbol{y}_K,\boldsymbol{\eta}^*) \right)$
4: $\quad \delta^{(i)} \leftarrow \| \nabla \ell(\boldsymbol{\eta}^{(i-1)}; \boldsymbol{y}_1,\ldots,\boldsymbol{y}_K,\boldsymbol{\eta}^*) \|$
5: $\quad i \leftarrow i+1$
6: **end while**
7: $\boldsymbol{\eta}_{MLE} \leftarrow \boldsymbol{\eta}^{(i)}$
8: $\boldsymbol{\eta}^* \leftarrow \boldsymbol{\eta}_{MLE}$, $i \leftarrow 1$, and repeat until the end of time.

Output: MLE estimate $\boldsymbol{\eta}_{MLE}$.

Chapter 5

Computer Experiments

We now discuss the computer experiments carried out to empirically validate the theoretical results presented thus far. For these experiments we set $\boldsymbol{\beta}$ to be a univariate parameter β_1 whose value is kept constant at 0.25. Next, for all possible combinations of three distinct values of the spatial auto-correlation parameters θ_1 and θ_0, a simulation experiment was carried out. Each experiment consisted of generating sample data sets under the true parameters and then estimating these parameters using the theory discussed in previous sections.

In each simulation, 150 samples each containing $K = 20$ data grids of dimensions 10×10 were generated via Gibbs sampling. An example of one realisation of the data generated is shown in Fig. 5.1. In each simulation the distribution of the single covariate x was generated as follows. The lower triangular elements of the data grid were assigned a value uniformly between 600 and 1200 while the upper triangular elements were assigned a value uniformly between 0 and 10 (i.e., the dimension of the data grid). Finally, all of the values in the data grid were divided by the maximum value to yield $x \in [0,1]$. An example of the distribution of the covariates is shown in Fig. 5.2. For each sample, the MCMC chain was allowed to run for 1020 iterations, the last 20 of which were taken to be the observed data. Afterwards, the MPLE for the 20 observed data sets taken jointly was estimated. Using the joint MPLE as the initial value for the importance parameter $\boldsymbol{\eta}^*$, Gibbs sampling was used to generate 3000 pseudo-samples. The first $M = 1000$ of these were discarded as burn-in while the latter $M = 2000$ were used in estimating the ratio

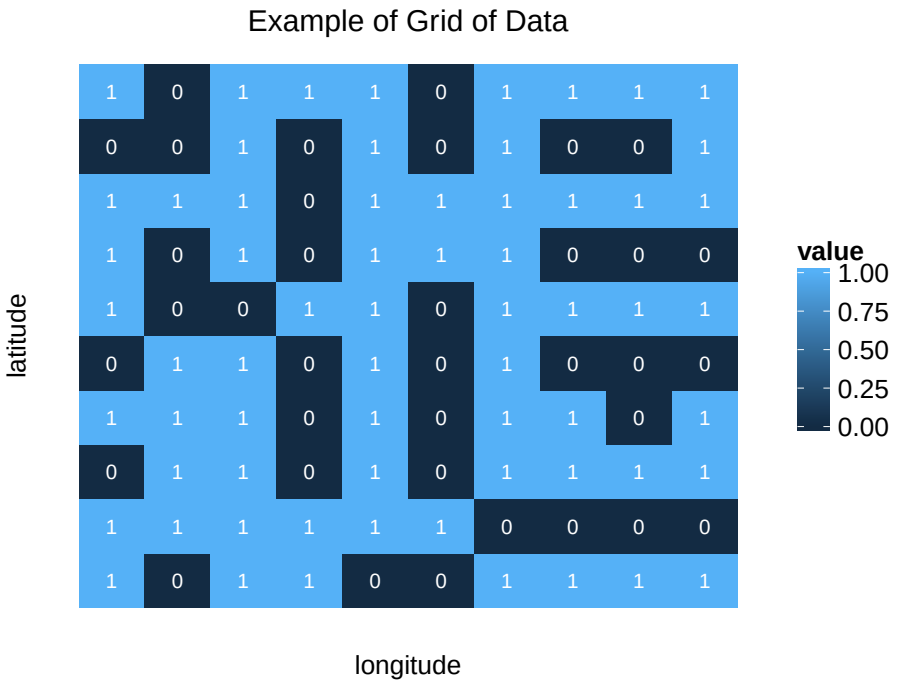

Figure 5.1: A realisation of an observed data grid generated in the simulations.

of the normalising constants as in Eq. 3.12. Using the pseudo-samples, the multi-sample joint log-likelihood was maximised via Newton-Raphson with a maximum of 20 optimisation iterations, where if the Newton-Raphson optimisation exceeded 20 iterations we denoted the MLE as non-convergent. Finally, the Gibbs sampling and Newton-Raphson steps were repeated four more times. The value at this final 5^{th} cycle of the optimisation procedure was taken to be the estimated MLE $\hat{\boldsymbol{\eta}}_{MLE}$.

Summary statistics for $\hat{\boldsymbol{\eta}}_{MLE}$ at the 5^{th} cycle of the optimisation procedure for six of the simulations are presented in Tables 5.1 – 5.3 and their distributions are shown in the last panel of Fig. 5.3 – Fig. 5.8. The distribution of the parameters during the first four cycles of the optimisation procedure are shown in these figures as well. For simulations 4, 6, and 9, only 3, 53, and 8 samples converged and, therefore, the results for these may not be reliable and are not shown here. We note that although the estimation improves considerably between the first and second cycle

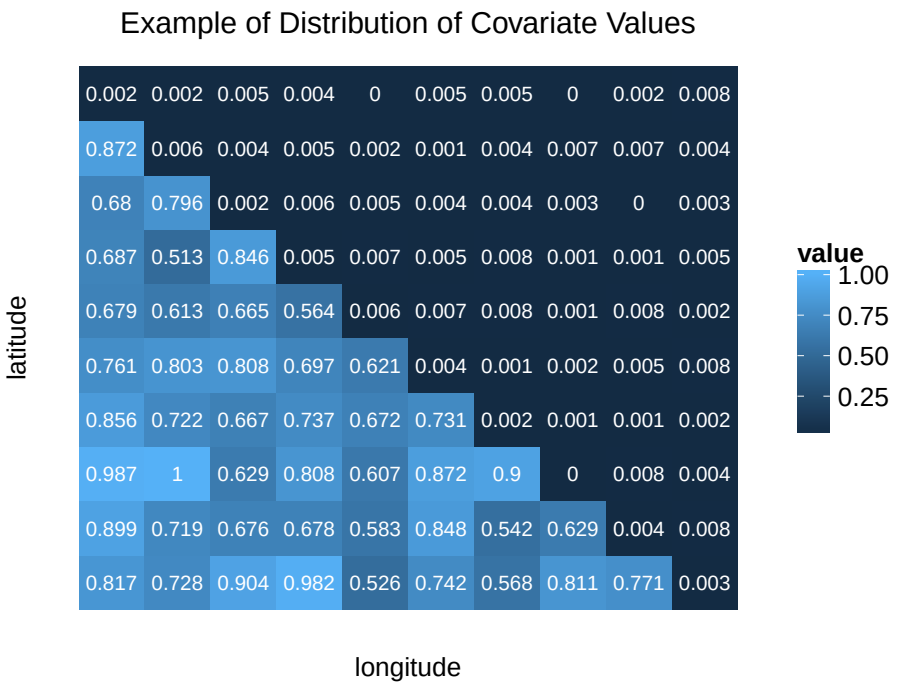

Figure 5.2: A typical realisation of the distribution of the covariate used in the simulations.

of the MCMC and Newton-Raphson optimisation procedure, later cycles provide much more moderate improvements to the estimation. Furthermore, the improvements in these later cycles are driven by a bias-variance trade-off where a decrease in variance is accompanied by an increase in bias. This behaviour is perhaps most clearly illustrated in Fig. 5.3 and Fig. 5.4 below. In the tables, *Asymptotic SD* denotes the standard deviation obtained using the the observed Fisher information averaged across the 150 samples while *MCMC SD* denotes the standard deviation of the 150 estimated parameters obtained in the simulation. In some of the simulations, the Newton-Raphson optimisation reached the maximum allowed 20 iterations for optimisation or the likelihood yielded a singular Hessian. If this occurred, we denoted the algorithm as not converging and labelled the MLE of the sample as unknown. The column *Num. Converged* denotes the number of samples that converged and yielded an MLE out of the possible 150 in each simulation.

Although some of the simulations yielded quite reasonable results, we observed that in some simulations the number of samples that converged was drastically small. We attribute this mainly to poor pseudo-samples being generated and to numerical issues remained even after implementing the methods discussed in Section 4.2.

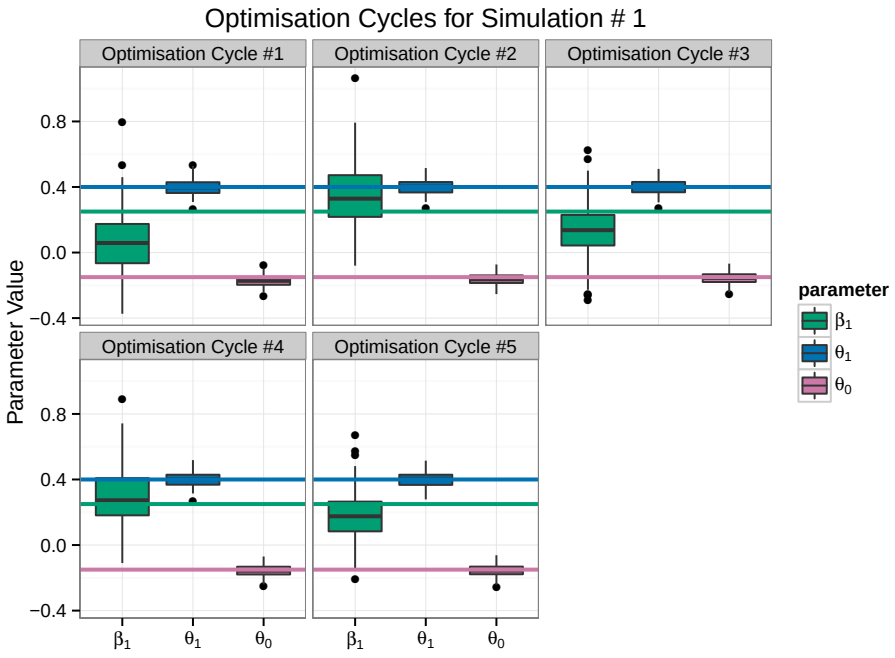

Figure 5.3: Distribution of estimated parameter vector $\hat{\boldsymbol{\eta}}$ throughout five cycles of the Gibbs (MCMC) and Newton-Raphson optimisation for the 1^{st} simulation experiment. Horizontal line indicates the true value of the parameter boxplot of the corresponding colour.

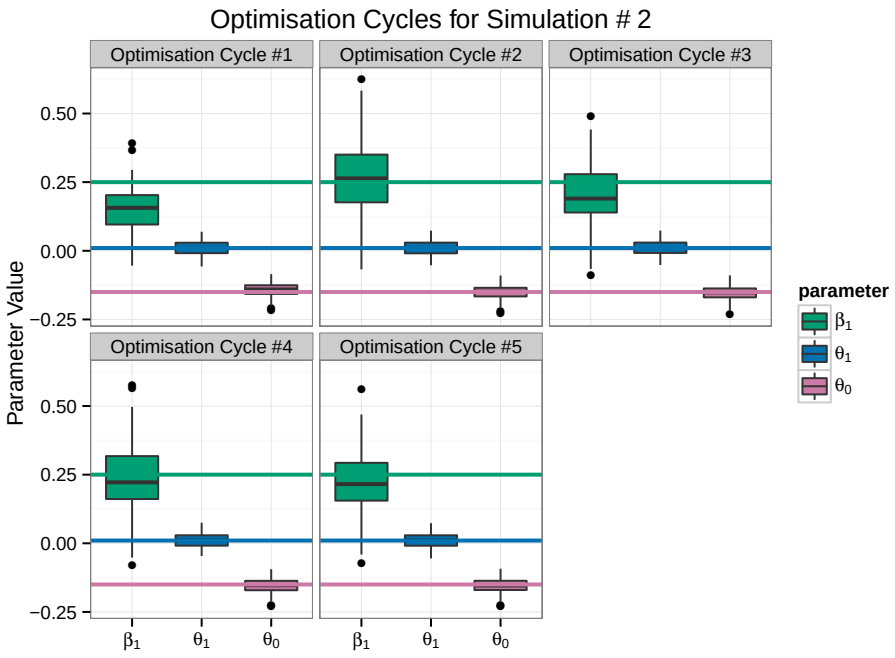

Figure 5.4: Distribution of estimated parameter vector $\hat{\boldsymbol{\eta}}$ throughout five cycles of the Gibbs (MCMC) and Newton-Raphson optimisation for the 2^{nd} simulation experiment. Horizontal line indicates the true value of the parameter boxplot of the corresponding colour.

Table 5.1: Summary statistics for estimated parameter $\hat{\beta}_1$ for six of the nine simulations conducted. Since only 3, 53, and 8 samples converged in the 4^{th}, 6^{th}, and 9^{th} simulations, respectively, the results are not shown here.

True $(\beta_1, \theta_1, \theta_0)$	Mean	Asymptotic SD	MCMC SD	Bias	MSE	Num. Converged
$(0.25, 0.4, -0.15)$	0.17	0.17	0.15	-0.08	0.03	150
$(0.25, 0.01, -0.15)$	0.22	0.13	0.11	-0.03	0.01	150
$(0.25, -0.35, -0.15)$	0.28	0.15	0.35	0.03	0.12	86
$(0.25, 0.4, 0.3)$	–	–	–	–	–	–
$(0.25, 0.01, 0.3)$	0.18	0.24	0.23	-0.07	0.06	81
$(0.25, -0.35, 0.3)$	–	–	–	–	–	–
$(0.25, 0.4, -0.05)$	0.15	0.18	0.18	-0.10	0.04	148
$(0.25, 0.01, -0.05)$	0.23	0.12	0.08	-0.02	0.01	150
$(0.25, -0.35, -0.05)$	–	–	–	–	–	–

Table 5.2: Summary statistics for estimated parameter $\hat{\theta}_1$ for six of the nine simulations conducted. Since only 3, 53, and 8 samples converged in the 4^{th}, 6^{th}, and 9^{th} simulations, respectively, the results are not shown here.

True $(\beta_1, \theta_1, \theta_0)$	Mean	Asymptotic SD	MCMC SD	Bias	MSE	Num. Converged
$(0.25, 0.4, -0.15)$	0.40	0.03	0.05	-0.00	0.00	150
$(0.25, 0.01, -0.15)$	0.01	0.02	0.03	0.00	0.00	150
$(0.25, -0.35, -0.15)$	-0.34	0.02	0.05	0.01	0.00	86
$(0.25, 0.4, 0.3)$	–	–	–	–	–	–
$(0.25, 0.01, 0.3)$	-0.05	0.16	0.16	-0.06	0.03	81
$(0.25, -0.35, 0.3)$	–	–	–	–	–	–
$(0.25, 0.4, -0.05)$	0.40	0.03	0.05	-0.00	0.00	148
$(0.25, 0.01, -0.05)$	0.01	0.02	0.02	0.00	0.00	150
$(0.25, -0.35, -0.05)$	–	–	–	–	–	–

Table 5.3: Summary statistics for estimated parameter $\hat{\theta}_0$ for six of the nine simulations conducted. Since only 3, 53, and 8 samples converged in the 4^{th}, 6^{th}, and 9^{th} simulations, respectively, the results are not shown here.

True $(\beta_1, \theta_1, \theta_0)$	Mean	Asymptotic SD	MCMC SD	Bias	MSE	Num. Converged
$(0.25, 0.4, -0.15)$	-0.15	0.02	0.04	-0.00	0.00	150
$(0.25, 0.01, -0.15)$	-0.15	0.01	0.03	-0.00	0.00	150
$(0.25, -0.35, -0.15)$	-0.19	0.02	0.04	-0.04	0.00	86
$(0.25, 0.4, 0.3)$	–	–	–	–	–	–
$(0.25, 0.01, 0.3)$	0.26	0.07	0.12	-0.04	0.02	81
$(0.25, -0.35, 0.3)$	–	–	–	–	–	–
$(0.25, 0.4, -0.05)$	-0.06	0.02	0.04	-0.01	0.00	148
$(0.25, 0.01, -0.05)$	-0.05	0.01	0.02	-0.00	0.00	150
$(0.25, -0.35, -0.05)$	–	–	–	–	–	–

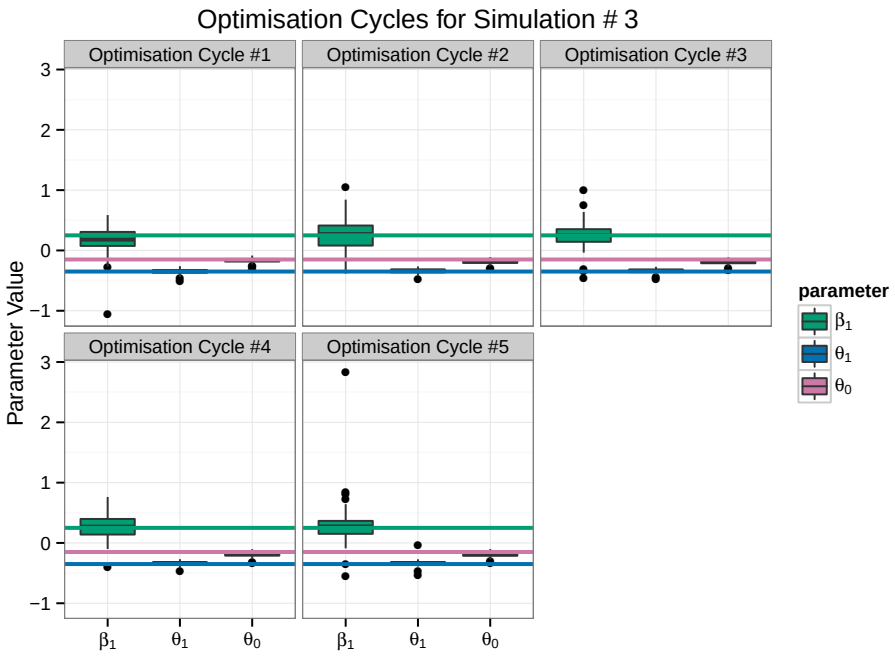

Figure 5.5: Distribution of estimated parameter vector $\hat{\boldsymbol{\eta}}$ throughout five cycles of the Gibbs (MCMC) and Newton-Raphson optimisation for the 3^{rd} simulation experiment. Horizontal line indicates the true value of the parameter boxplot of the corresponding colour.

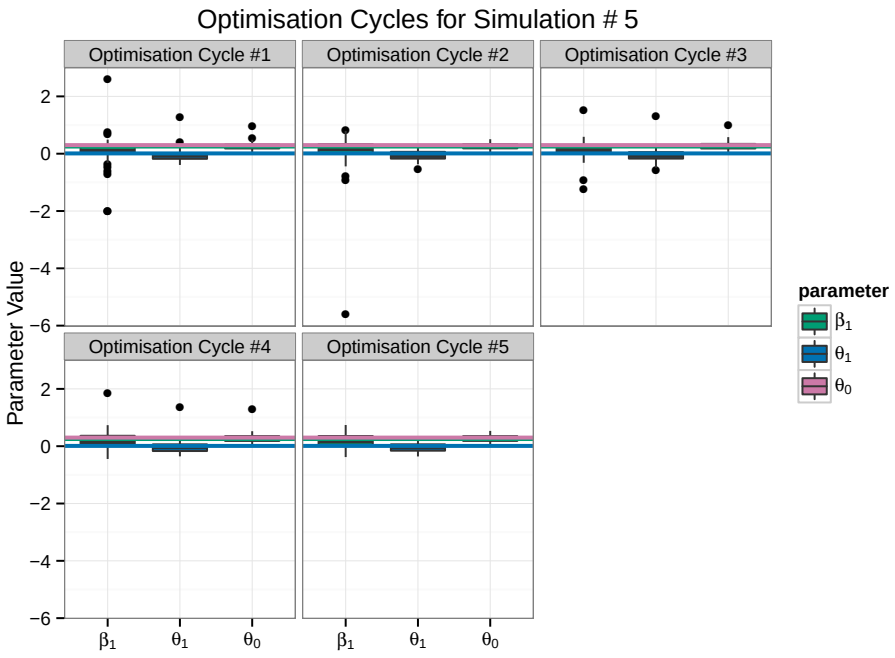

Figure 5.6: Distribution of estimated parameter vector $\hat{\boldsymbol{\eta}}$ throughout five cycles of the Gibbs (MCMC) and Newton-Raphson optimisation for the 5^{th} simulation experiment. Horizontal line indicates the true value of the parameter boxplot of the corresponding colour.

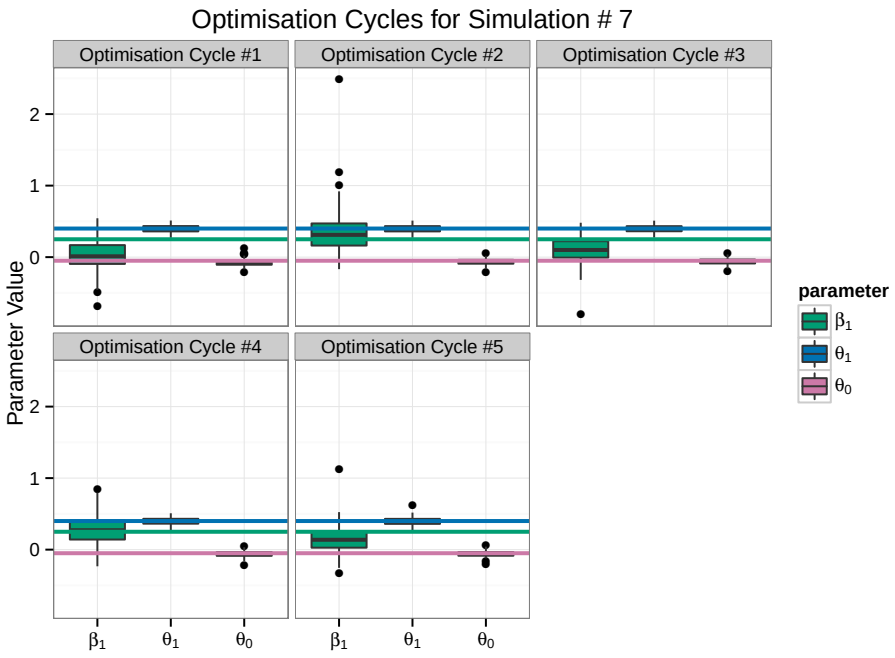

Figure 5.7: Distribution of estimated parameter vector $\hat{\boldsymbol{\eta}}$ throughout five cycles of the Gibbs (MCMC) and Newton-Raphson optimisation for the 7^{th} simulation experiment. Horizontal line indicates the true value of the parameter boxplot of the corresponding colour.

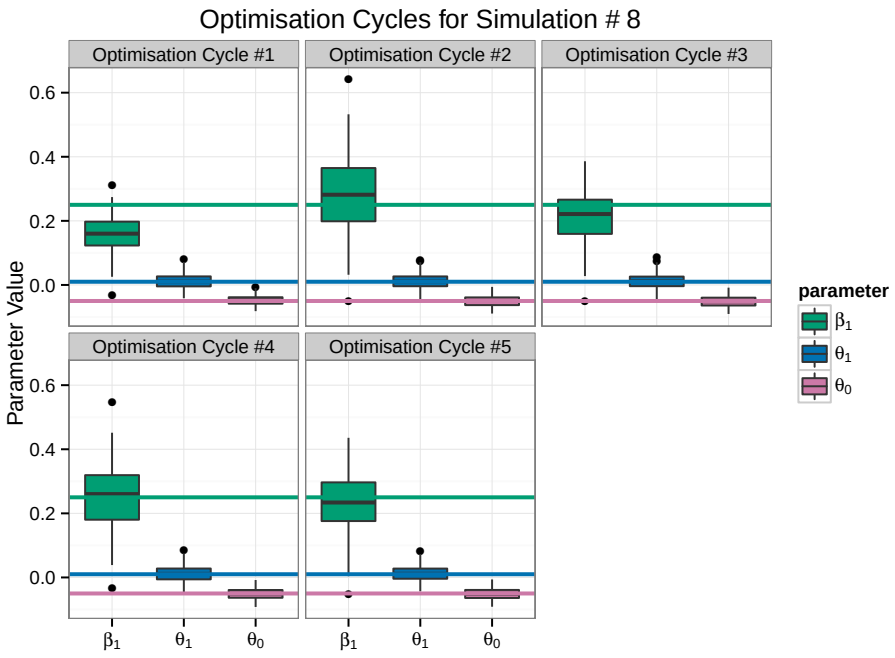

Figure 5.8: Distribution of estimated parameter vector $\hat{\boldsymbol{\eta}}$ throughout five cycles of the Gibbs (MCMC) and Newton-Raphson optimisation for the 8^{th} simulation experiment. Horizontal line indicates the true value of the parameter boxplot of the corresponding colour.

Chapter 6

Conclusion

The use of Markov random fields is a viable method for habitat modelling. By building an auto-logistic model within the Markov random fields framework, distinct types of ecological and biological data can be incorporated into a mathematical model for inference and habitat prediction. Although this model can be applied to the data from a single animal, it also generalisable for use with the data from multiple animals. However, particularly in the multi-sample case, there exist some computational limitations with the method which future work may explore from an implementation or a theory approach. This may include further work in optimisation procedures which are more cost effective or yield better convergence results than the MCMC-MLE procedure presented. Despite these issues, the current implementation of the model has the advantages of relying on inference techniques that are understandable by ecologists, of being intuitively interpretable via the model parameters, and of allowing a compromise among the accuracy and variety of the ecological and environmental data that may be used to build the model.

Bibliography

[1] A. Agaskar and Y. M. Lu. Alarm: A logistic auto-regressive model for binary processes on networks. In *Global Conference on Signal and Information Processing (GlobalSIP) IEEE*, pages 305 – 308, 2013. → pages 32

[2] S. Ban and A. W. Trites. Quantification of terrestial haul-out and rookery characteristics of steller sea lions. *Marine Mammal Science*, 23(3):496 – 507, 2007. → pages 1

[3] K. J. Benoit-Bird et al. Prey patch patterns predict habitat use by top marine predators with diverse foraging strategies. *PLoS ONE*, 8(1):e53348, 2013. → pages 3

[4] K. J. Benoit-Bird et al. Foraging behavior of northern fur seals closely matches the hierarchical patch scales of prey. *Marine Ecology Progress Series*, 479:283 – 302, 2013. → pages 3

[5] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36(2):192 – 236, 1974. → pages 5, 6, 10, 11

[6] R. J. Birgeneau, R. A. Cowley, G. Shirane, and H. Yoshizawa. Phase transitions in diluted magnets: Critical behavior, percolation, and random fields. *Journal of Statistical Physics*, 34(5 - 6):817 – 848, 1984. → pages 32

[7] G. K. H. Boor and R. J. Small. Steller sea lion spatial-use patterns derived from a bayesian model of opportunistic observations. *Marine Mammal Science*, 28(4):375 – 403, October 2012. → pages 2

[8] C. J. A. Bradshaw et al. At-sea distribution of female southern elephant seals relative to variation in ocean surface properties. *Journal of Marine Science*, 61:1014 – 1027, 2004. → pages 3

[9] N. Cressie. *Statistics for Spatial Data*. Wiley-Interscience, 1993. → pages 5, 7, 8, 10, 11

[10] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 4th edition, 2010. → pages 22

[11] A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes, editors. *Handbook of Spatial Statistics*. CRC Press, 2010. → pages 32

[12] C. J. Geyer and E. A. Thompson. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B*, 54(54):657 – 699, 1992. → pages 21, 33

[13] K. T. Goetz, R. A. Montgomery, J. M. V. Hoef, R. C. Hobbs, and D. S. Johnson. Identifying essential summer habitat of the endangered beluga whale delphinapterus leucas in cook inlet, alaska. *Endangered Species Research*, 16:135 – 147, 2012. → pages 1

[14] E. J. Gregr and A. W. Trites. A novel presence-only validation technique for improved steller sea lion eumetopias jubatus critical habitat descriptions. *Marine Ecology Progress Series*, 365:247 – 261, 2008. → pages 1, 2, 3

[15] C. Guinet et al. Spatial distribution of foraging in female antarctic fur seals arctocephalus gazella in relation to oceanographic variables: a scale-dependent approach using geographic information systems. *Marine Ecology Progress Series*, 219:251 – 264, 2001. → pages 3

[16] M. L. Gumpertz, J. M. Graham, and J. B. Ristaino. Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence. *Journal of Agricultural, Biological, and Environmental Statistics*, 2(2):131 – 156, 1997. → pages 28

[17] J. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. *Unpublished*, 1971. → pages 9

[18] F. Huffer and H. Wu. Markov chain monte carlo in spatial binary data models with application to the distribution of plant species. *Biometrics*, 54 (2):509 – 524, 1998. → pages 13, 21, 24

[19] D. S. Johnson, P. B. Conn, M. B. Hooten, J. C. Ray, and B. A. Pond. Spatial occupancy models for large data sets. *Ecology*, 94(4):801 – 808, 2013. → pages 1

[20] K. Kaschner, R. Watson, A. W. Trites, and D. Pauly. Mapping world-wide distributions of marine mammal species using a relative environmental suitability (res) model. *Marine Ecology Progress Series*, 316:285 – 310, 2006. → pages 2, 3

[21] C. A. Nordstrom, B. C. Battaile, C. Cott, and A. W. Trites. Foraging habitats of lactating northern fur seals are structured by thermocline depths and submesoscale fronts in the eastern bering sea. *Deep-Sea Research II*, 88 - 89:78 – 79, 2012. → pages 3, 4

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

[23] D. Warton and G. Arts. Advancing our thinking in presence-only and used-available analysis. *Journal of Animal Ecology*, 82:1125 – 1134, 2013. → pages 2, 3

Appendix A

Appendix

A.1 Standard Notation

This section serves as a reference for some of the notation commonly used throughout this document. Notation is listed in order of appearance.

Notation	Description
$\boldsymbol{s}$	Vector of sites indexed $i = 1, \ldots, N$
$y(\boldsymbol{s}_i)$	Observed response at the i^{th} site
$\mathcal{N}_i$	Neighbourhood around the i^{th} site.
$P(\cdot)$	Probability mass function with support ζ
$0(\boldsymbol{s}_i)$	Absence/zero at the i^{th} site
$\mathbb{I}(\cdot)$	Indicator function
$Q(\cdot)$	Negative potential function
y_i	Short-hand notation for $y(\boldsymbol{s}_i)$
$\boldsymbol{x}$	Model covariate(s)
$\boldsymbol{\beta}$	Model covariate parameter(s)
θ	Model spatial auto-correlation parameter(s)
$z(\boldsymbol{\eta})$	Partition function/normalising constant
$\boldsymbol{\eta}$	Model parameter vector
0_i	Short-hand notation for $0(\boldsymbol{s}_i)$
$h_\eta(\cdot)$	Un-normalised likelihood function
$\boldsymbol{\eta}^*$	Importance parameter vector
$f(\cdot)$	Likelihood function
$\boldsymbol{y}_{k,obs}$	Observed data of spatial locations for k^{th} animal
$\boldsymbol{y}^*_{k,m}$	m^{th} pseudo-data sample generated for k^{th} animal
$t(\cdot)$	Function of sufficient statistics
$L(\cdot)$	Likelihood function
$\ell(\cdot)$	Log-likelihood function