Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Thomas W. Hsiao Date

Statistical approaches for understanding and addressing preferential sampling in model-based geostatistics

By

Thomas W. Hsiao
Doctor of Philosophy

Biostatistics and Bioinformatics

Lance A. Waller, Ph.D.
Advisor

John Hanfelt, Ph.D.
Committee Member

Michael Kramer, Ph.D.
Committee Member

Robert Lyles, Ph.D.
Committee Member

Accepted:

Kimberly J. Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical approaches for understanding and addressing preferential sampling in model-based geostatistics

By

Thomas W. Hsiao
B.A., Rice University, TX, 2017
M.S.P.H., Emory University, GA, 2024

Advisor: Lance A. Waller, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2025

Abstract

Statistical approaches for understanding and addressing preferential sampling in model-based geostatistics
By Thomas W. Hsiao

Traditional geostatistical and point process methods assume observation locations are independent of the latent spatial process of interest. Violations of this assumption, known as preferential sampling (PS), can introduce bias in spatial inference and prediction. Current practice is to adjust for PS through a shared variable approach or shared latent process (SLP) model, which has been shown to improve prediction and inference. However, our understanding of how traditional methods perform under PS remains limited, alternatives to the SLP are sparse, and point process modeling approaches have yet to be implemented in several applied settings, including species distribution modeling of monarch butterflies.

In the first aim, we examined the large sample behavior of the maximum likelihood estimator (MLE) for the model-based geostatistics framework under PS, assuming a stationary Gaussian process with Matérn covariance. Surprisingly, we found that under general conditions, the fixed-domain asymptotic behavior of the MLE is unaffected by the sampling mechanism. Moreover, as sample size increases, the MLE corrects for PS-induced bias more effectively than common alternative methods like composite likelihood and the Vecchia approximation and attains performance similar to the SLP for PS adjustment.

In the second aim, we introduce inverse sampling intensity weighting (ISIW) as a novel alternative to the SLP. In ISIW, we first estimate the sampling intensity at each observation location and then incorporate these estimates as weights in a weighted likelihood adjustment. Our approach preserves kriging's linear predictor structure and leverages the Vecchia approximation for scalability. While ISIW performs poorly for inference, it dramatically improves prediction under PS, is computationally faster, and remains robust across different PS mechanisms, though estimating the weights in practice remains a key challenge.

Finally, we extend PS concepts to species distribution modeling, addressing varying sampling effort (VSE) in Journey North monarch butterfly data (2011–2020). Using presence-only citizen science data and distance sampling from the nearest road, we apply a thinned point process model approach with integrated nested Laplace approximation (INLA) to estimate the spatial distribution of monarchs in the western United States, improving model fit and revealing strong evidence of VSE in citizen science data. Our approach helps to quantify preferential sampling in passive wildlife surveillance.

Statistical approaches for understanding and addressing preferential sampling in model-based geostatistics

By

Thomas W. Hsiao
B.A., Rice University, TX, 2017
M.S.P.H., Emory University, GA, 2024

Advisor: Lance A. Waller, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2025

Acknowledgments

Some people begin their PhD with a clear vision of what they hope to achieve. Safe to say, I was not one of them. Through the help of everyone I would like to thank in this section, we were able to figure it out.

I would like to first thank Dr. Rudy Guerra from my time as a university statistics major. It's terrifying to realize how easily a professor can completely change the trajectory of one's life through just a simple word of encouragement. I'm fortunate that professor ended up being you, and I'm happy to say you were right all those years ago - I think I can be good at this.

A few professors outside of my dissertation committee require an acknowledgment. I would like to personally thank Dr. Mary Kelley for teaching me how to write a results section, Dr. Eugene Huang for inspiring me with the rigor and depth of thought you bring to your research, and Dr. Howard Chang for your invaluable guidance, which was instrumental in publishing my first first-author paper.

I truly enjoyed working with my dissertation committee—Dr. John Hanfelt, Dr. Bob Lyles, and Dr. Michael Kramer. Your engagement with my research not only challenged me but also deepened my appreciation for the importance of preparing and sharing research properly.

I made many incredible friends over the past few years. My first office - Teng, Xin, Lin , and Lindsey were a true powerhouse. I still curse the pandemic for sending us home mid-semester and cutting our time together short.

I also want to thank my 2019 cohort - Shiyu, Ye, Yuxuan, Yiheng, Jialu, Shijia, and Rachel. While busy schedules got in the way of our socials, I'm glad we entered the program and got to experience all the craziness of the 2020's together.

It took a while after COVID for me to make my way back to Emory, but my new office : Lindsey (again), Jacob, and Emily made going to school much more fun. Chopping up the latest NFL Sunday action and chomping down on infinite Jolly

Ranchers brought some much needed levity to my research life.

Thank you to Enoch, Preeti, Jithin, Tina, Radhika, Natalia, Ani and Nikol for being the strong interdisciplinary group of PhD student and PhD affiliated friends I needed to keep me sane during COVID and help me realize the world beyond the GCR 3rd floor.

On the same note, I want to thank the Emory Taiwanese Student Association for providing me with unbelievable experiences and in particular Eric, Katie, Allison, Chin-En, and Sean for being some of the closest friends I've made since leaving college.

My advisor, Dr. Lance Waller, has been essential to my growth. Thank you for showing me how it's done. I have undoubtedly become an improved person and researcher since I began the program thanks to your guidance.

Thank you to my parents for your love, support, and unwavering encouragement in pursuing my goals. I really am the luckiest son in the world.

Finally, thank you to my wife Christina. The past year we have been living life on speedrun but I am glad we are speeding together. I am grateful for every kind word, every supportive hug, and every warm meal we've shared over the course of this journey.

Contents

List of Figures

List of Tables

Chapter 1

Background and preliminaries

The research in this dissertation focuses on *preferential sampling*, an undesirable yet inevitable feature of spatial data analysis where the response of interest is dependent on the locations at which observations are made. Preferential sampling is studied under the framework of model-based geostatistics, a subfield of spatial statistics which borrows heavily from the theory of stochastic processes. Before diving into preferential sampling, we provide a brief overview of the preliminaries required.

1.1 Stochastic processes

Many problems can be solved by considering groups of random variables and the relationships between them. The stochastic process is the building block of these methods.

Definition 1.1.1 (Stochastic process)**.** *A stochastic process is a collection of random variables* $\{S_\alpha : \alpha \in \mathcal{I}\}$ *where each random variable* S_α *is defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *for some index set* $\mathcal{I}$*.*

Familiar stochastic processes include discrete and continuous time Markov Chains, martingales, and Poisson processes. While the index set $\mathcal{I}$ is flexible and can be

either countable or uncountable, we solely focus on *spatial* stochastic processes where $\mathcal{I} \subseteq \mathbb{R}^2$. Even with this restriction, the class of stochastic processes over Euclidean space is quite general - we will focus specifically on Gaussian random fields.

1.2 Gaussian random fields

We first introduce Gaussian processes (GP) and a subset of GP's known as Gaussian random fields (GRF).

Definition 1.2.1 (Gaussian process)**.** *A stochastic process* $\{S_\alpha : \alpha \in \mathcal{I}\}$ *is a Gaussian process if and only if for every finite set of indices* $\{\alpha_1, ..., \alpha_n\} \subseteq \mathcal{I}$*, the vector* $(S_{\alpha_1}, S_{\alpha_2}, ..., S_{\alpha_n})$ *follows a multivariate Gaussian distribution with mean vector* $\boldsymbol{\mu} = (\mathbb{E}(S_{\alpha_1}), \mathbb{E}(S_{\alpha_2}), ..., \mathbb{E}(S_{\alpha_n}))$ *and covariance matrix* $\boldsymbol{\Sigma}$ *with entries* $\boldsymbol{\Sigma}_{ij} = C(S_{\alpha_i}, S_{\alpha_j}) = \mathbb{E}[(S_{\alpha_i} - \mu)(S_{\alpha_j} - \mu)]$ *for a valid covariance function* C*.*

Definition 1.2.2 (Gaussian random field)**.** *A Gaussian random field is a Gaussian process where the index set* $\mathcal{I}$ *is a subset of* $\mathbb{R}^d$ *for* $d \geq 1$*.*

In practice, Gaussian random fields (GRFs) are often assumed to have additional properties that simplify computation and enhance mathematical tractability. A key challenge in spatial statistics is that observations represent an incomplete sampling of the underlying surface, and we typically observe only a single realization. Valid and attainable inference relies on the assumptions of stationarity and isotropy.

In focusing our attention to GRF's for spatial statistics, we now refer to the index set as the *study area* $\mathcal{D} \subseteq \mathbb{R}^2$ and *locations* in the study area as $\mathbf{x}$. The random variable of a GRF S associated with location $\mathbf{x}$ will now be referred to as $S(\mathbf{x})$.

Definition 1.2.3 (Weak (second-order) stationarity)**.** *A GRF is weakly (or second-order) stationary if and only if the mean is constant and for any lag vector* $\mathbf{h}$*, the covariance function can be written as* $\text{Cov}(S(\mathbf{x}), S(\mathbf{x} + \mathbf{h})) = C(\mathbf{h})$ *for all* $\mathbf{x} \in \mathcal{D}$*.*

Definition 1.2.4 (Intrinsic stationarity)**.** *A GRF is intrinsically stationary if and only if the mean is constant and* $\mathrm{Var}(S(\mathbf{x}) - S(\mathbf{x}+\mathbf{h})) = \mathrm{Var}(\mathbf{h})$ *for all* $\mathbf{x} \in \mathcal{D}$*.*

One can show that weak stationarity is the stronger condition and implies intrinsic stationarity. The converse is not necessarily true. The utility of intrinsically stationary GRF's is clear when working with GRF's where a covariance function C does not exist. We make further assumptions on covariance functions by isotropy.

Definition 1.2.5 (Isotropy)**.** *A weakly stationary GRF and its covariance function are isotropic if the covariance function depends on the spatial lag only through its Euclidean distance* $\|\mathbf{h}\|$*.*

While stationarity and isotropy are strong and potentially unrealistic assumptions, they transform an otherwise intractable $n = 1$ problem into one where parameters depend only on the distance between points, introducing a degree of replicability. Theoretical guarantees for statistical estimation typically require multiple observations of the same quantity, which becomes feasible only when assuming that covariances depend solely on distance. Without these assumptions, most spatial statistics problems with a single realization become nearly intractable.

Similar to finite-dimensional Gaussian distributions, GP's are fully characterized by their mean and covariance functions. A primary difficulty to applying GPs in practice is how to guarantee an admissible covariance function $C(\cdot, \cdot)$ that constructs a positive semi-definite covariance matrix $\mathbf{\Sigma}$ for any combination of indices. This challenge has lead to the use of several parametric families of covariance functions deemed flexible enough for application. Some well-known examples include the exponential, spherical, and Gaussian covariance functions. Perhaps the most widely used isotropic covariance function is the Matérn:

$$C(h;\nu,\sigma^2,\phi) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}h/\phi)^\nu K_\nu(\sqrt{2\nu}h/\phi), \sigma > 0, \phi > 0, \nu > 0 \qquad (1.1)$$

where h is the (scalar) Euclidean distance between points, ν is the smoothness parameter controlling the mean-squared differentiability of the GRF, σ^2 is the variance parameter, and ϕ is the range parameter. The Matérn has the attractive property of including the exponential covariance family ($\nu = 1/2$) and converging to the Gaussian class of covariance functions as $\nu \to \infty$. Additionally, the ν parameter provides greater flexibility in controlling smoothness compared to other covariance function families, which is one of the main reasons Stein (1999) strongly advocated for its use in applied spatial problems. For the remainder of this dissertation, we will primarily focus on the Matérn covariance.

Many other formulations of GP's exist, but the finite-dimensional representation given in Definition 1.2.1 is the simplest and most suitable for our purposes in model-based geostatistics.

1.3 Point processes

Point processes are mathematical objects suited for modeling spatiotemporal events. Their probability distribution is defined over collections of points rather than the points themselves. For some probability space $(\Omega, \mathcal{F}, P)$ and study area $\mathcal{D} \subseteq \mathbb{R}^2$, we equip $\mathcal{D}$ with the Borel σ-algebra $\mathcal{B}$ and denote the class of bounded Borel sets as $\mathcal{B}_0$. For any subset $x \subseteq \mathcal{D}$, let $n(x)$ equal the cardinality or number of events in x, where $n(x) = \infty$ indicates an infinite set. A set x is determined to be locally finite if for any bounded set $B \subseteq \mathcal{D}$, the set $x_B = x \cap B$ has cardinality $n(x_B) < \infty$. We can then formally define the space of locally finite point configurations $N_{lf} := \{x \subseteq \mathcal{D} : n(x_B) < \infty \text{ for all bounded } B \subseteq \mathcal{D}\}$ and equip it with the σ-algebra

$\mathcal{N}_{lf} := \sigma(\{x \in N_{lf} : n(x_B) = m\} : B \in \mathcal{B}_0, m \in \mathbb{N}_0)$. We are now ready to rigorously define a spatial point process.

Definition 1.3.1 (Spatial point process (Moller and Waagepetersen, 2003))**.** *A spatial point process X defined on a study area $\mathcal{D}$ is a random variable on a probability space $(\Omega, \mathcal{F}, P)$ mapping into the measurable space $(N_{lf}, \mathcal{N}_{lf})$. The term point process may refer either to X itself or the distribution of X, given by $P \circ X^{-1} : \mathcal{N}_{lf} \to [0, 1]$.*

Unlike typical random variables, point processes map into the measurable space of locally finite point configurations rather than a standard Euclidean space. This can make them somewhat awkward to work with mathematically for manipulation. A key lemma allows us to simplify the point process by uniquely defining their probability law through finite-dimensional distributions, not unlike our characterization of Gaussian processes in Definition 1.2.1. Let $N(B) = n(X_B)$ be the count function, which represents the number of points falling in a bounded set B from the point process X.

Lemma 1.3.1. *The distribution of a point process X is determined by the finite dimensional joint distribution of $N(B_1), N(B_2), ..., N(B_m)$, for any $B_1, ..., B_m \in \mathcal{B}_0$ and $m \in \mathbb{N}_0$.*

Lemma 1.3.1 allows us to work with familiar finite dimensional count distributions rather than directly with the infinite dimensional point process X. One straightforward notion is to allow each $N(B_i)$ to follow a Poisson distribution. Such a construction naturally leads us to the fundamental building block of point processes - the Poisson process. Before introducing the Poisson, we first define a few additional features of point processes known as the intensity function and the intensity measure.

Definition 1.3.2 (Intensity function and measure)**.** *The intensity function of a spatial point process is a non-negative function over the study region $\lambda : \mathcal{D} \to [0, \infty)$ that is locally integrable ($\int_B \lambda(\xi) d\xi < \infty$) for all bounded $B \subseteq \mathcal{D}$. The intensity measure is the measure given by $\mu(B) = \int_B \lambda(\xi) d\xi, B \subseteq \mathcal{D}$.*

Definition 1.3.3 (Poisson point process)**.** *A point process X is Poisson with intensity function λ if for any subset $B \subseteq \mathcal{D}$ with $\mu(B) < \infty$, the number of points in B, $N(B)$, follows a Poisson distribution with mean $\mu(B)$. In addition, for any collection of mutually exclusive sets $B_1, B_2, ...$ in $\mathcal{D}$ such that $B_i \cap B_j = \emptyset$ for $i \neq j$, the random variables $N(B_i)$ and $N(B_j)$ are independent.*

The second property in Definition 1.3.3 is known as complete spatial randomness (CSR). It describes a point pattern where the occurrence of any event has no bearing on the location of any other event in the point pattern. Poisson processes with a constant intensity function are referred to as *homogeneous* Poisson processes whereas those with λ that vary across the study region are known as *inhomogeneous* Poisson processes. Despite the added flexibility of an inhomogeneous intensity function, Poisson processes are often too simplistic to model the diverse point patterns encountered in applications, particularly those exhibiting second-order properties and clustering that violate the independence assumption. The Cox process addresses these limitations by incorporating the Poisson process into a hierarchical model with an additional layer where the intensity function is a realization of a random field. Because of this unique structure, Cox processes have also been referred to as doubly stochastic Poisson processes.

Definition 1.3.4 (Cox process)**.** *Let $S := \{S(\mathbf{x}) : \mathbf{x} \in \mathcal{D}\}$ be a non-negative random field (not necessarily Gaussian). If the conditional distribution of X given S is a Poisson process on $\mathcal{D}$ with intensity function S, then X is a Cox process driven by S.*

The clustering or attractiveness property of Cox processes can be proven by showing $\mathbb{V}\text{ar}[N(B)] \geq \mathbb{E}[N(B)]$ for any bounded set $B \in \mathcal{D}$, with equality only achieved if X is Poisson. Several types of Cox processes have been studied, including Neyman-Scott processes, shot-noise Cox processes (SNCPs), and sigmoidal point processes (Moller and Waagepetersen, 2003). A special type of Cox process called the log

Gaussian Cox process (LGCP) makes use of Gaussian random fields in its intensity function.

Definition 1.3.5 (Log Gaussian Cox process)**.** *Let S be a Gaussian random field. Then a Cox process X driven by* $\exp\{S\}$ *is a log Gaussian Cox process.*

The LGCP serves as the foundation for our analysis of preferential sampling. Traditional geostatistical methods typically ignore stochasticity in the observations X and focus solely on the underlying Gaussian random field S. In contrast, preferential sampling methods using the LGCP introduce dependence in the likelihood by modeling the observations X as being driven by $\exp\{S\}$. The choice of a LGCP is largely motivated by its mathematical tractability, as its moments have closed form expressions and the effect of preferential sampling can be exactly quantified. We first provide an overview of model-based geostatistics (MBG).

1.4 Model-based geostatistics

In a model-based geostatistical analysis, the main interest is in understanding a latent spatially continuous surface S over some study region $\mathcal{D} \subseteq \mathbb{R}^2$. Examples include air pollution monitoring (Shaddick and Zidek, 2012), heavy metal biomonitoring (Diggle et al., 2010), and abundance measuring of wildlife populations (Conn et al., 2017). A typical study will collect indirect measurements of S at a finite set of point-referenced locations within $\mathcal{D}$, but several sources of error exist that interfere with our ability to measure S directly. As first delineated by Diggle et al. (1998), let the observed measurements be a n-dimensional vector $\mathbf{Y} = (Y_1, ..., Y_n)$ located at points $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$. Then we assume for each Y_i,

$$Y_i = \boldsymbol{z}(\mathbf{x}_i)^\top \boldsymbol{\beta}_{\boldsymbol{z}} + S(\mathbf{x}_i) + \epsilon_i, \mathbf{x}_i \in \mathcal{D}, i = 1, \ldots, n. \tag{1.2}$$

where $\boldsymbol{z}(\mathbf{x}_i)$ are a set of measured covariates at location $\mathbf{x}_i$, $\mathbf{x}_i$ is the location at which Y_i is observed, ϵ_i is measurement error with mean 0 drawn from a normal distribution with "nugget variance" τ^2, and S is a zero mean GRF with covariance function dependent on parameters $\boldsymbol{\phi}$. It is often assumed that the covariance function is both *stationary* and *isotropic*. The power of the MBG framework is in its simplicity. What was originally a highly complex and nonparametric challenge of estimating a spatially continuous surface plus parameters has been transformed into a straightforward parametric statistics problem. The target parameters for inference are the mean coefficient vector, the nugget variance, and the covariance parameters. Once the parameters are estimated, the entire model is defined and other analyses of interest can be conducted include predicting values of S at unobserved locations, computing averages of S over a small region, or checking if S exceeds a given threshold value.

1.4.1 Estimation and inference

An important note is that the locations $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ are assumed to be fixed and independent of any other random element in (1.2). Therefore, the likelihood does not require the distribution of $\mathbf{X}$ and estimation can proceed as with any linear model.

Estimation of the model in (1.2) has several possibilites. Classic methods include two-stage or iterative estimation, where the covariance parameters and nugget are estimated in the first stage under intrinsic stationarity, commonly through variogram estimation. Their estimates are then plugged into the covariance function, and used to estimate the regression parameter vector by some form of least squares. Iterative methods will continue until estimates stabilize (Gelfand et al., 2010). For future reference, when referring to least squares estimation in the context of the regression parameters for spatial modeling, it is assumed the covariance parameters are known.

Least squares estimation of the fixed effects does not assume any distributional

form of the observations $\mathbf{Y}$. The least squares estimators follow the familiar form $\hat{\beta} = (Z^\top V Z)^{-1} Z^\top V \mathbf{Y}$ which can follow three different flavors.

1. Ordinary least squares (OLS). The matrix $V = I_n$ is the identity matrix.

2. Weighted least squares (WLS). The matrix V is a diagonal matrix, with entries equal to the weights $w = (w_1, ..., w_n)$ associated with each observation in $\mathbf{Y}$.

3. Generalized least squares (GLS). The matrix $V = \mathbf{\Sigma}$ equals the full covariance matrix, with entries corresponding to evaluations of the assumed covariance function C for the GRF S.

An alternative to the iterative procedure described is maximum likelihood estimation (MLE), which has the advantage of estimating covariance and regression parameters simultaneously at the cost of explicitly assuming the observed data follow a multivariate Gaussian distribution. Bayesian estimation also makes use of the full likelihood, but with the inclusion of prior distributions on each of the parameters.

We will shortly see in Chapter 2 that the GLS and MLE which make use of the full covariance matrix have much more desirable properties compared to simple OLS and WLS. However, unlike OLS and WLS, GLS and MLE estimation assuming a full covariance matrix are severely hampered by the necessary inversion of the covariance matrix $\mathbf{\Sigma}$, which is a cubic $O(n^3)$ operation despite the fast Cholesky decomposition available for symmetric covariance matrices. Cubic operations like matrix inversions quickly make estimation over 1000s of points computationally infeasible. Much research has been dedicated to approximate the full likelihood GLS/MLE including covariance tapering (Kaufman et al., 2008), composite likelihood (Bevilacqua et al., 2012), and sparse precision matrices (Lindgren et al., 2011; Datta et al., 2016) with some even reaching near linear complexity. A full survey of methods and a comparison can be found in Heaton et al. (2019).

1.4.2 Prediction

Once the parameters are estimated, prediction of S at new locations conventionally proceeds by kriging, also known as Gaussian process regression (GPR). Using properties of least square optimality (or conditional distribution identities for Gaussian random variables), the best linear unbiased predictor (BLUP) in terms of least squares error, and its variance are

$$\begin{aligned} \hat{S}(\mathbf{X}_0) &= \mu + C(\mathbf{X}_0, \mathbf{X}_n)^\top \boldsymbol{\Sigma}_n^{-1}(\mathbf{Y} - \mu \mathbf{1}_n), \\ \mathrm{Var}(\hat{S}(\mathbf{X}_0)) &= C(\mathbf{X}_0, \mathbf{X}_0) - C(\mathbf{X}_0, \mathbf{X}_n)^\top \boldsymbol{\Sigma}_n^{-1} C(\mathbf{X}_0, \mathbf{X}_n), \end{aligned} \tag{1.3}$$

where $\mathbf{X}_0$ are unobserved locations. As in the estimation procedure, the optimality of the BLUP relies on the assumption that the sampling locations are fixed and independent of $\mathbf{Y}$. However, under preferential sampling, this assumption no longer holds, and kriging may fail to retain the usual optimality properties that justify its use in standard geostatistics.

1.5 Preferential sampling

1.5.1 Failure of standard methods under PS

In this section, I outline the differences between non-preferential and preferential sampling scenarios. While the failure of standard statistical procedures under preferential sampling is often intuitive, it is useful to examine precisely when and how estimation and prediction break down.

Let the bracket notation $[\mathbf{Y}]$ denote the probability distribution of random vector $\mathbf{Y}$. In a non-preferential sampling (NPS) scenario, the critical assumption is that S and $\mathbf{X}$ are statistically independent, i.e. $[\mathbf{X}|S] = [\mathbf{X}]$. Let the subscript $\boldsymbol{\theta}$ indicate a probability distribution depends on some component of the target parameter vector,

under the model specified in 1.2. Then the associated likelihood is given by

$$\begin{aligned} L(\boldsymbol{\theta}) \propto \int [\mathbf{Y}, \mathbf{X}, S]_{\boldsymbol{\theta}} dS &= \int [\mathbf{Y}|\mathbf{X}, S]_{\boldsymbol{\theta}} [\mathbf{X}|S]_{\boldsymbol{\theta}} [S]_{\boldsymbol{\theta}} dS \\ &= \int [\mathbf{Y}|\mathbf{X}, S]_{\boldsymbol{\theta}} [\mathbf{X}] [S]_{\boldsymbol{\theta}} dS \\ &= [\mathbf{Y}|\mathbf{X}]_{\boldsymbol{\theta}} [\mathbf{X}] \end{aligned} \tag{1.4}$$

Therefore we have $L(\boldsymbol{\theta}) \propto [\mathbf{Y}|\mathbf{X}]_{\boldsymbol{\theta}}$ due to the absence of $\boldsymbol{\theta}$ in $[\mathbf{X}]$. Conditional on $\mathbf{X}$, standard geostatistical methods can proceed accordingly.

However, if $\mathbf{X}$ and S are dependent, then the factorization no longer guarantees the likelihood of the locations is free of the target parameters. The factorization would be

$$L(\boldsymbol{\theta}) \propto \int [\mathbf{Y}, \mathbf{X}, S]_{\boldsymbol{\theta}} dS = \int [\mathbf{Y}|\mathbf{X}, S]_{\boldsymbol{\theta}} [\mathbf{X}|S]_{\boldsymbol{\theta}} [S]_{\boldsymbol{\theta}} dS = [\mathbf{Y}, \mathbf{X}]_{\boldsymbol{\theta}} \tag{1.5}$$

Unlike $[\mathbf{Y}|\mathbf{X}]_{\boldsymbol{\theta}}$, the joint density is difficult to compute in closed form, as we can no longer ignore the stochasticity of $\mathbf{X}$. Dinsdale and Salibian-Barrera (2019a) provides an explanation of how PS affects prediction and interferes with the usual proof that the kriging predictor in (1.3) satisfies the BLUP property.

PS refers to scenarios where a stochastic dependence exists between the locations at which observations are made and the value of the response itself. An intuitive example is the design of an air pollution monitoring system. The goal is to monitor a study region to accurately measure pollution levels at any given location. However, resource constraints and land-use policies prevent monitors from being placed everywhere, requiring careful resource allocation. Rather than placing monitors in areas with consistently clean air, it is more practical to concentrate them in high-risk areas where pollution levels are likely to be elevated. As a result, observations are

preferentially collected in regions with higher pollution levels. If this dependence is ignored, severe bias may arise due to the restricted range of pairwise differences used in variogram estimation (Watson et al., 2019).

If monitor placement were completely independent of prior pollution knowledge, bias would not be a concern, but such a system would be less efficient and significantly more costly. Clearly, statistical methods that account for stochastic dependence in the sampling process are necessary. In the next section, we review the main developments in PS methodology over the past decade.

1.5.2 Marked point processes

Ho and Stoyan (2008) applied the theory of marked point processes to analyze data where observation locations are statistically dependent on the response. Their study focused on the Log-Gaussian Cox Process (LGCP) formulation of preferential sampling, where the log intensity of the observations is given by $\lambda(\mathbf{x}) = \exp\{\alpha + \beta S(\mathbf{x})\}$ and the marks $\mathbf{Y}$ follow a normal distribution with mean $\mu\mathbf{1}_n$ and a specified covariance structure. The authors derived the moments and various second-order functions for the marks, demonstrating that this type of preferential sampling induces a shift in the mean of the marks by an additive factor of $\beta\sigma^2$, where σ^2 is the variance of S.

1.5.3 The shared latent process model

Diggle et al. (2010) introduced the first model-based solution to the PS problem, which we will refer to as the shared latent process (SLP) model. The SLP makes use of the same marked point process described in Ho and Stoyan (2008), but now with a focus on estimation in the model-based geostatistics framework. The model rests on three main assumptions:

1. Each $Y_i|S(\mathbf{x}_i)$ is an independently distributed Gaussian random variable with

mean $\mu + S(\mathbf{x}_i)$ and variance τ^2.

2. $\mathbf{X}|S$ follows an inhomogeneous Poisson process with log intensity $\lambda(\mathbf{x}) = \exp\{\alpha + \beta S(\mathbf{x})\}$ (equivalent to saying $\mathbf{X}$ follows a LGCP driven by S).

3. S follows a weakly stationary and isotropic Gaussian process, with zero mean and covariance function chosen by the investigator (Matern, exponential, etc.)

Under these three assumptions, the SLP model can be written as

$$\begin{aligned} Y_i|S(\mathbf{x}_i) &\sim N(\mu + S(\mathbf{x}_i), \tau^2), \\ \mathbf{X}|S &\sim \mathcal{LGCP}(\lambda(\mathbf{x})), \\ \lambda(\mathbf{x}) &= \exp\{\alpha + \beta S(\mathbf{x})\}, \\ S &\sim \mathcal{GP}(0, \Sigma_\theta). \end{aligned} \tag{1.6}$$

The key feature of the model is the GRF S shared by both the intensity of the observations $\mathbf{X}$ and the mean of $[\mathbf{Y}|S]$. The parameter β measures the degree of PS. Positive values of $\beta > 0$ imply high rates of sampling in areas with large response values. A PS coefficient of $\beta = 0$ implies no PS present, since the sampling intensity would be completely independent of $\mathbf{Y}$. In essence, it is the full implementation of the marked point process model introduced in Ho and Stoyan (2008).

In the original paper, the model was estimated with a Monte Carlo maximum likelihood approximation. However, the authors recognized that the straightforward Monte Carlo estimate would produce many realizations of S that would be incompatible with $\mathbf{Y}$ and $\mathbf{X}$ leading to highly inefficient sampling. To circumvent this problem, they employed a clever adjustment to the likelihood, by breaking up the spatial process S into observed and unobserved locations. Let S_0 be the values of S where observations were made. Then

$$L(\boldsymbol{\theta}) = \int [\mathbf{X}|S] \frac{[\mathbf{Y}|S_0]}{[S_0|\mathbf{Y}]} [S_0][S|\mathbf{Y}] dS = E_{S|\mathbf{Y}} \left[[\mathbf{X}|S] \frac{[\mathbf{Y}|S_0]}{[S_0|\mathbf{Y}]} [S_0] \right] \tag{1.7}$$

By reframing the likelihood as a conditional expectation evaluated over $[S|\mathbf{Y}]$, a natural Monte Carlo estimate of m samples that only relies on simulating realizations of S conditional (or compatible) on $\mathbf{Y}$ can be derived as

$$L(\boldsymbol{\theta}) = \frac{1}{m} \sum_{j=1}^{m} [\mathbf{X}|S_j] \frac{[\mathbf{Y}|S_{0j}]}{[S_{0j}|\mathbf{Y}]} [S_{0j}] \tag{1.8}$$

The method was applied to heavy metal biomonitoring in Galicia, Spain, where the authors observed a stark difference between the parameter estimates and the risk surface obtained from the PS model versus the NPS model. This discrepancy suggests that PS was present and had a significant impact on estimation bias.

The two main limitations with the SLP method which have hindered its widespread adoption in practice, are:

1. **Computational speed**. While the authors mitigated the inefficiencies of the initial Monte Carlo approximation, their proposed solution still requires prohibitively long computation times for standard problems.

2. **Inaccurate estimation**. Dinsdale and Salibian-Barrera (2019a) proved that the approximation in (1.8) does not actually target the intended likelihood function from (1.6).

These issues make the DMS method challenging to implement and difficult to justify in applied settings. While all subsequent methods adhere to the SLP model, they improve upon it by refining the integration approach and entirely avoiding the issue of missing the target likelihood.

1.5.4 The Bayesian shared latent process model

Pati et al. (2011) extended the original SLP model from Diggle et al. (2010) by estimating it in a full Bayesian model and introducing both covariates (instead of a constant mean) and an extra Gaussian process η_r to the mean of $\mathbf{Y}$ (referred to as the "residual" process). The new Bayesian SLP (BSLP) is

$$\begin{aligned} Y_i|S(\mathbf{x}_i) &\sim N(\eta(\mathbf{x}_i) + \beta\lambda(\mathbf{x}_i), \tau^2) \\ \mathbf{X}|S &\sim \mathcal{LGCP}(\lambda(\mathbf{x})) \\ \lambda(\mathbf{x}) &= \exp\{z(\mathbf{x})^\top \beta_\lambda + S(\mathbf{x})\} \\ \eta(\mathbf{x}) &= z(\mathbf{x})^\top \beta_\eta + \eta_r(\mathbf{x}) \\ S &\sim \mathcal{GP}(0, \Sigma_{\theta_S}) \\ \eta_r &\sim \mathcal{GP}(0, \Sigma_{\theta_\eta}) \end{aligned} \tag{1.9}$$

At first glance, drawing direct parallels between (1.9) and the SLP model in (1.6) can be challenging. A useful starting point is the intensity function for $\mathbf{X}$. While $S(\mathbf{x})$ still drives the intensity, the coefficient β that quantifies the degree of PS is now absent. Instead, the log intensity of $\mathbf{X}$, denoted $\lambda(\mathbf{x})$, is incorporated as a covariate in the mean function, with β serving as its coefficient.

Additionally, the model introduces covariates $z(\mathbf{x})$ into the intensity function, providing greater explanatory power for the site selection mechanism. An "extra" Gaussian process, denoted as $\eta_r(\mathbf{x})$, is also included and is completely independent of $\mathbf{X}$. This "residual process" captures any remaining spatially correlated variation that cannot be explained by both $\lambda(\mathbf{x})$ and $z(\mathbf{x})$.

Despite these modifications to the original SLP model, the parameter β remains the key measure of PS, with higher values indicating stronger dependence. While the Bayesian implementation is a straightforward extension of SLP, the true value of the

paper was its establishment of key theoretical properties of the SLP model, albeit under increasing domain asymptotics. The authors demonstrated that the observed data $(\mathbf{Y}, \mathbf{X})$ are informative of the PS parameter β by proving that its posterior remains proper even when assigned an improper prior. They also established weak posterior consistency for all fixed effect, covariance, and PS parameters.

1.5.5 Estimation by TMB

A third estimation technique for fitting the SLP model alongside the Monte Carlo approximation for MLE in (1.8) and Markov chain Monte Carlo (MCMC) for full Bayesian estimation is the use of Laplace approximation to handle the latent Gaussian component S. Latent Gaussian models are particularly well suited for Laplace approximations, and the spatial model in (1.6) falls within this class, even with the additional log-Gaussian Cox process (LGCP) likelihood.

On top of proving that the approximation in (1.8) targeted the incorrect likelihood, Dinsdale and Salibian-Barrera (2019a) demonstrated that the SLP Laplace approximation (implemented via the Template Model Builder (TMB) R package (Kristensen et al., 2016)) exhibited superior bias reduction and efficiency compared to the original method in Diggle et al. (2010). Moreover, its performance was comparable to the stochastic partial differential equation (SPDE) approach (Lindgren et al., 2011) in R-INLA (Rue et al., 2009).

The main contribution of this work was to provide a reliable and computationally efficient method for fitting PS models while remaining fully consistent with the original SLP framework. While the SPDE approach in R-INLA also offers an alternative to the SLP, the authors advocate for TMB due to its greater flexibility, particularly in allowing users to specify the log-likelihood directly. Since the model structure remains identical to the original SLP, this approach enables a direct comparison based solely on estimation techniques.

1.5.6 Estimation by INLA (and a new PS framework)

The central tenet unifying the previous SLP methods is the inhomogeneous intensity function of $\mathbf{X}$. By allowing the same random field S to influence both the sampling intensity and the response mean, PS can be quantified through a single highly interpretable parameter β. However, Watson et al. (2019) argues that this LGCP is unrealistic for proper spatial inference.

Building on concerns first raised in Professor Dawid's discussion of the SLP (Diggle et al., 2010), Watson et al. (2019) asserts that the SLP model cannot plausibly represent a realistic site selection process. The designers of an initial site allocation cannot observe the true realization of S before data collection begins. Its structure is only partially revealed after the first measurements are obtained. The authors acknowledge the utility of β as a post-hoc measure of PS, but propose a new framework that incorporates latent effects capable of more realistically modeling the true site selection process.

Watson et al. (2019) also raises a critical issue in the existing PS literature: with the exception of the abundance study in Conn et al. (2017), prior work has largely assumed $\mathbf{Y}$ follows a continuous, approximately normal distribution. This assumption overlooks real-world scenarios where the response data exhibit skewness, heavy tails, overdispersion, or are inherently discrete.

Using the SPDE (Lindgren et al., 2011) approach as implemented in R-INLA (Rue et al., 2009), Watson et al. (2019), introduced a novel framework for spatial modeling under PS that deviated from the traditional SLP approach in five essential ways:

1. **Spatio-temporal extension**: the response is extended to the spatiotemporal case, not just spatial.

2. **Non-continuous data likelihood**: the response is extended to allow any distribution for the response (continuous, count, etc.).

3. **Sums of latent random effects**: the underlying random field S is now composed of a sum of independent latent random effects, rather than just defined as a realization of a single Gaussian process. A suite of spatially-correlated random effects including Gaussian Markov random fields (GMRFs), autoregressive terms, or even iid effects are available in R-INLA. The added flexibility of decomposing the latent field driving the stochastic dependence into a sum of latent effects may better capture the true data generating mechanism.

4. **Site-selection emulation**: temporal lags can be accommodated in the shared latent random effects driving the site selection process, that also impact the observation process. Therefore latent effect values from any time $t < t_0$ can inform the site selection at time t_0.

5. **Bernoulli site-selection**: the LGCP in the SLP model is discarded in favor of a selection process that only accounts for a finite number of potential sites, rather than entertaining any given point in the study region. In conjunction with the non-continuous data likelihood, this allows for greater flexibility in the variety of data that can be modelled (since a logistic likelihood for the site selection converges to a Poisson process as the discretized grid grows more granular).

Watson's work greatly expanded the types of models that could be fit under PS, repurposed PS models to actually describe site selection over time, and improved accessibility to applied researchers by taking advantage of an existing software in R-INLA.

1.5.7 Weighted composite likelihood for PS

Breaking further from the SLP tradition, both Schliep et al. (2023) and Vedensky et al. (2023) introduced weighted composite likelihood (CL) methods as an alter-

native approach to adjusting for PS. A composite likelihood is constructed as the product of marginal and conditional event likelihoods which are subsets of the full joint likelihood. While CL does not represent the true density of the observed data, therefore sacrificing some efficiency, it offers a practical computational alternative to the MLE when the full joint density is intractable (Varin et al., 2011).

Different CL formulations exist depending on context, with common choices for spatial data including pairwise marginal, pairwise difference, and pairwise conditional likelihoods. These approaches are computationally efficient alternatives to MLE and possess desirable large sample properties under mild assumptions. In particular, for CL-based variogram estimation using the pairwise difference likelihood, Curriero and Lele (1999) demonstrated significant computational improvements over MLE, with only a minimal increase in bias. The authors also showed that regardless of the true marginal distribution of the contrasts, as long as the first two moments are correctly specified, the estimator retains theoretical guarantees for consistency and asymptotic normality.

To adjust for PS, Schliep et al. (2023) extend the (unweighted) pairwise difference CL from Curriero and Lele (1999) by including weights inversely proportional to the sampling intensity for the respective location. The likelihood under a weakly stationary random field is:

$$\log L_{\mathcal{CW}}(\boldsymbol{\theta}) \propto \sum_{i=1}^{n-1} \sum_{j>i} \frac{1}{\lambda(\mathbf{x}_i)} \frac{1}{\lambda(\mathbf{x}_j)} \left\{ -\frac{(Y_i - Y_j)^2}{2\gamma(d_{ij};\phi)}) - \log(\gamma(d_{ij};\phi)) \right\} \tag{1.10}$$

Where $\gamma(d_{ij}) = \text{Var}(Y_i - Y_j)$ is the variogram evaluated for locations distance d_{ij} apart and the λ's are sampling intensities. Inference using the weighted composite likelihood (CL) requires estimating the inverse sampling weights beforehand. Schliep et al. (2023) employed nonparametric kernel density estimators to estimate the first-

order intensity, selecting the bandwidth following Diggle (1985). Their simulation studies suggest that using estimated sampling intensities produced results similar to those obtained with the true sampling intensity.

It is important to note that the weight adjustment in Schliep et al. (2023) was designed to recover estimates that would have been obtained if a finite set of population locations had been observed. In other words, the weights serve as a Horvitz-Thompson-style adjustment, accounting for the fact that only a subset of a finite population of locations (from a Gaussian random field) was sampled.

Our objective in this dissertation, however, is more aligned with traditional model-based geostatistics. Rather than estimating what would have been observed under a finite superset of points, we use a finite set of sampled locations to infer properties of the underlying continuous Gaussian random field itself. The weighted composite likelihood approach presented in the review by Vedensky et al. (2023) aligns more closely with this goal. However, their example serves more as a proof of concept rather than a fully developed methodological framework. They demonstrated that an inverse sampling intensity weighted univariate marginal composite likelihood based on (1.2), estimated in a fully Bayesian setting, can account for preferential sampling—even when the weights are estimated via the same kernel density approach. However, their results indicate that using the true weights still outperforms using estimated weights.

A key limitation of their approach is that, because it was based on a univariate model, only the mean parameters β and nugget variance τ^2 could be estimated. The covariance parameters could not be estimated in their approach. We seek to use the inverse sampling intensity approach to perform full estimation of mean, covariance, and nugget parameters in the MBG framework under preferential sampling.

1.5.8 Summary

Ho and Stoyan (2008) and Diggle et al. (2010) introduced the SLP model into geo-

statistics, marking the first statistical approach to explicitly address the stochastic dependence between site selection and the response. Pati et al. (2011) extended the SLP framework to full Bayesian estimation and incorporated an additional Gaussian residual process to enhance model flexibility. Their primary contribution was demonstrating that the observed data could be informative about the degree of preferential sampling under the SLP model. Dinsdale and Salibian-Barrera (2019a) re-estimated the SLP model using a Laplace approximation instead of the original Monte Carlo approximation, improving computational efficiency. Watson et al. (2019) further advanced the field by implementing INLA for the SLP and developing a more general PS framework. Their work pushed PS methodology beyond Gaussian data, introduced a wider range of random effect specifications within the SLP, and incorporated site selection dynamics over time.

Gelfand et al. (2012) did not propose a new PS model but instead presented an experimental framework for comparing prediction surfaces across multiple spatial interpolation methods. They emphasize that the ideal strategy for addressing PS should be the inclusion of appropriate covariates rather than compensatory modeling, effectively transforming a spatial missing not at random (MNAR) problem into a missing at random (MAR) problem. Under this condition, PS can theoretically be ignored. This perspective is also highlighted in Pacifici et al. (2016), who applied PS and adaptive sampling methods to estimate occupancy for rare species. However, both Gelfand et al. (2012) and Conn et al. (2017) demonstrate through simulation studies that the inclusion of an informative covariate does not fully correct for the bias introduced by PS, even when it is the sole covariate influencing the sampling intensity. These findings reinforce the necessity of explicit modeling for PS and help explain the continued adoption of the SLP approach in the geostatistics literature.

Preferential sampling has also gained significant attention in the field of ecology, with most applications focusing on species distribution models (SDMs). Unlike the

original SLP framework, SDMs often involve count or binary response data for counting the abundance or measuring the presence of wildlife. Additionally, observation locations correspond to areal units on a lattice rather than points within a continuous surface. Prior to the general framework introduced by Watson et al. (2021), several count-based models had been used to incorporate PS in abundance data (Chakraborty and Gelfand, 2010; Cecconi et al., 2016; Conn et al., 2017; Pennino et al., 2019). A key extension was introduced in Conroy et al. (2023), who extended the disease risk model from Cecconi et al. (2016) to scenarios where the denominator for the disease rate is unknown. Their approach involved two likelihoods: one for cases and one for controls, resulting in two distinct PS parameters (β^+ and β^-). A more general framework for PS in bivariate spatial data is provided in Shirota and Gelfand (2022). When count data are unavailable, an alternative and efficient strategy for species distribution modeling is to use presence-absence or presence-only data. These models treat observations as a thinned point process, integrating both the underlying species intensity and the probability of detection (Chakraborty et al., 2011; Manceur and Kühn, 2014; Fithian et al., 2015; Gelfand and Shirota, 2019; Sicacha-Parada et al., 2021).

Other discussions of preferential sampling include Ferreira and Gamerman (2015), who investigate optimal design strategies for geostatistical parameter estimation under PS. Ferreira (2020) extend the standard SLP model to account for local repulsion effects, where sampling effort around existing observation sites decreases to mitigate diminishing returns in information gain from closely spaced locations. This additional influence on site selection perturbs the underlying sampling intensity, necessitating adjustments for proper inference. Building on this idea, Gray and Evangelou (2023) propose a Bayesian method that relaxes the assumption of conditional independence in $[\mathbf{Y}|S]$ allowing for potential attraction or repulsion effects in observation locations. Beyond these developments, several other contributions have extended the SLP in

important directions and applications, including air pollution monitoring (Lee et al., 2011, 2015), phylodynamic inference (Karcher et al., 2016), hedonic modelling (Paci et al., 2020), spatially-varying PS (Amaral et al., 2023), a hypothesis test to detect PS in spatio-temporal data (Watson, 2021), and exact Bayesian inference for the LGCP component of the SLP (Moreira and Gamerman, 2022; Moreira et al., 2023). The predictive performance and computational tractability of the SLP has lead to the popularity of model-based approaches for accounting for dependence between spatial processes and their observation locations.

The current role of preferential sampling in spatial inference remains uncertain. It is evident that PS is prevalent in many geostatistical datasets, and it may well be the case that PS is more the norm than an exceptional circumstance in geospatial data collection. The SLP model as implemented in R-INLA remains the dominant approach for addressing PS, available for both Gaussian and non-Gaussian data. Additionally, a hypothesis test requiring no knowledge of hierarchical models has been proposed (Watson, 2021). However, many applied researchers have yet to incorporate checks for PS into their workflows.

Modern spatial statistics research has become highly computational. While most researchers adhere to the theoretical framework of a latent Gaussian field driving the response and use kriging for prediction, the $O(n^3)$ computational cost of inverting the covariance matrix has lead to an explosion of methods designed to perform spatial inference and prediction for massive spatial data. These methods include covariance tapering (Furrer et al., 2006), stochastic partial differential equations (Lindgren et al., 2011), nearest neighbor Gaussian processes (Datta et al., 2016), fixed rank kriging (Cressie and Johannesson, 2008), predictive processes (Banerjee et al., 2008), deep compositional models (Zammit-Mangion et al., 2022), and composite likelihood (Bevilacqua et al., 2012). Fundamentally, these advances have not altered the core geostatistical model in (1.2) but have introduced more scalable estimation techniques.

If we recognize PS as a critical factor in geospatial analysis, an important question arises: Do we need a PS-specific solution for every existing geostatistical method? Should separate studies be conducted to incorporate PS into NNGP, FRK, SPDE, and CL methods individually? Should separate R packages be developed for each method once PS is accounted for, in addition to those already available on CRAN?

Instead of treating PS methods as standalone models, a more generalized approach may be possible. PS could be incorporated as an interchangeable component applicable to any estimation procedure following the geostatistical model in (1.2), much like a prior or likelihood in Bayesian inference. Alternatively, we could seek a unified framework that generalizes across existing spatial methods and integrates PS mechanisms directly within this broader structure.

One counterargument against incorporating PS into the aforementioned spatial methods is that massive spatial datasets may not suffer from PS-related biases. As $n \to \infty$, it may become unlikely that observation locations remain dependent on the response, as large-scale datasets often feature gridded or regularly spaced sampling designs, such as those in remote sensing and satellite data. However, ecological applications, particularly those involving wildlife monitoring, provide a clear counterexample. Even when n large, PS remains a significant issue due to practical constraints in data collection. Determining whether data are sufficiently large or geographically extensive to mitigate PS remains an open and intriguing question.

This dissertation aims to address some of the key questions raised in this chapter. Chapters 2 and 3 examine the large-sample properties of standard geostatistical methods under PS to better understand the conditions under which PS may have minimal impact. Chapter 4 explores inverse sampling intensity weighting for geostatistical inference and prediction as an alternative to the SLP, introducing a novel integration into the Vecchia likelihood, as opposed to the pairwise difference and univariate marginal approaches used in previous studies. Finally, Chapter 5 investigates

PS methods for presence-only data, with a unique application to Citizen Science observations of monarch butterflies in the western United States. We wrap up with Chapter 6 detailing avenues for future work.

Chapter 2

The MLE under fixed domain asymptotics

To address the question of when preferential sampling (PS) matters in geostatistics, we examine the large-sample properties of the maximum likelihood estimator (MLE) when locations are observed under PS. While asymptotic results for independent data are well established, results for spatial data are more limited, particularly when locations are treated as stochastic rather than fixed. Theoretical guarantees for spatial estimators are especially challenging due to the ambiguity of the asymptotic framework due to a lack of natural ordering in $\mathbb{R}^2$ and the inherent dependence between observations. In this chapter, we review the main asymptotic results for the MLE and conduct an in-depth analysis of the asymptotic properties of each parameter in the model-based geostatistics (MBG) framework.

2.1 Asymptotic frameworks in spatial statistics

The most complete asymptotic results for spatial covariance and regression parameter estimators for stationary covariance functions relied on the increasing domain framework (Mardia and Marshall, 1984; Bachoc, 2014). Under the increasing domain

setting, the number of observation locations increases as $n \to \infty$, while the minimum distance between any two points is bounded below by some fixed constant. Consequently, the observation locations are not restricted to any bounded set but instead continually expand within the space. Although this assumption may be unrealistic in many practical spatial sampling designs due to natural boundaries and limited sampling areas, it offers significant mathematical tractability. The increasing separation between points leads to asymptotic independence, which controls the log-likelihood and its gradient. As a result, standard M-estimation proof techniques can be applied (Bachoc, 2020).

Fixed domain, or infill asymptotics are an alternative asymptotic regime for spatial statistics where the observation locations are confined to a bounded set in Euclidean space. Instead of assuming a minimum separation between all locations, infill asymptotics effectively imposes a maximum distance constraint, leading to increasingly dense sampling as $n \to \infty$. As the number of observations grows, the covariance function assumed for the GRF model induces higher correlations within the data. This growing dependence invalidates proof techniques used in the increasing domain setting, which rely on the eventual independence and vanishing covariance between points to take advantage of existing results for M-estimators (Boos and Stefanski, 2013; Vaart, 2000). A detailed discussion on the proper choice of asymptotic framework can be found in Zhang and Zimmerman (2005).

Covariance and regression parameter estimation under fixed domain asymptotics presents a counterintuitive scenario: some parameters of interest are identifiable but not estimable, while other parameters not necessarily of interest are estimable. Although increasing domain asymptotics permit a broad class of stationary covariance functions which lead to consistent and asymptotically normal estimators, estimation under fixed-domain asymptotics imposes far stricter conditions. For the remainder of our discussion on preferential sampling, we focus on the fixed-domain setting and

use the equivalence of Gaussian measures to clarify these concepts.

2.2 Equivalence of measures and microergodicity

The main proof for the consistency of the MLE for covariance parameters is similar in flavor to Wald's maximum proof of consistency (Wald, 1949). However, modifications are necessary because as $n \to \infty$ the data become increasingly dependent, preventing the use of the strong law of large numbers (SLLN). Instead, Zhang (2004) applied a martingale convergence theorem from Gikhman and Skorokhod (2004) where the limit of the likelihood ratio is a direct function of the equivalence between the two probability measures in question. Equivalence and orthogonality of measures play a central role in studying Gaussian processes under fixed-domain asymptotics (Ibragimov and Rozanov, 2012).

Definition 2.2.1 (Equivalent measures)**.** *Two probability measures P and Q defined on the same measurable space $(\Omega, \mathcal{F})$ are equivalent if they are absolutely continuous with respect to each other. For all $A \in \mathcal{F}$, $P(A) = 0$ if and only if $Q(A) = 0$.*

On the other end of the spectrum are orthogonal probability measures.

Definition 2.2.2 (Orthogonal measures)**.** *Two probability measures P and Q are orthogonal if there exists $A \in \mathcal{F}$ such that $P(A) = 1$ and $Q(A) = 0$.*

Equivalence and orthogonality are key concepts that determine whether two probability measures can be distinguished based on observed events. Two equivalent measures assign positive and zero mass to the same events. In contrast, two orthogonal measures describe completely disjoint outcomes. Events that occur under one have zero probability under the other. This concept is useful for Gaussian processes on infinite dimensional function spaces because any two infinite dimensional Gaussian processes are either equivalent or orthogonal (Ibragimov and Rozanov, 2012).

Knowing that any two Gaussian processes are either equivalent or orthogonal, we now seek sufficient conditions to determine which case holds. In particular, as discussed in Chapter 1, we focus on Gaussian random fields with stationary and isotropic covariance functions, which can be succinctly parameterized. The concept of microergodicity provides a powerful framework to determine equivalence or orthogonality of GRFs based on just a few key parameters, simplifying an otherwise complex problem. Consider a family of probability measures $\{P_\theta : \theta \in \Theta\}$, where each measure is indexed by covariance parameters from a parameter space Θ. We first define non-microergodicity in relation to equivalency of Gaussian measures.

Definition 2.2.3 (Non-microergodic parameter)**.** *Let g be a function from Θ to $\mathbb{R}^q$ for $q \in \mathbb{N}$. Then $g(\theta^*)$ is a non-microergodic parameter if there exists a $\theta \in \Theta$ such that $\Phi(\theta) \neq \Phi(\theta^*)$ and the measures P_{θ^*} and P_θ are equivalent.*

In other words, non-microergodic parameters are functions of the covariance parameters for which different values can still yield equivalent probability measures. As noted earlier, observed events cannot differentiate between equivalent probability measures. Therefore different values of non-microergodic parameters remain indistinguishable from one another regardless of how much data we observe. We can formalize this concept in the following result which tells us there exist no consistent estimators for non-microergodic parameters.

Lemma 2.2.1. *Let $(\boldsymbol{X}_i)_{i\in\mathbb{N}}$ be any sequence of points in study region $\mathcal{D} \subseteq \mathbb{R}^2$. If $g(\theta)$ is non-microergodic, then there does not exist a sequence of estimators $\hat{g}_n(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ that is weakly consistent under P_θ for $g(\theta)$ over all $\theta \in \Theta$.*

The proof can be found in both Zhang (2004) and Bachoc (2020). This result naturally raises the question: if non-microergodic parameters cannot be consistently estimated, which parameters can? We are now ready to discuss microergodic parameters.

Definition 2.2.4 (Microergodic parameter)**.** *Let g be a function from Θ to $\mathbb{R}^q$ for $q \in \mathbb{N}$. Then $g(\theta^*)$ is a microergodic parameter if for all $\theta \in \Theta$ such that $\Phi(\theta) \neq \Phi(\theta^*)$, the measures P_{θ^*} and P_θ are orthogonal.*

While it may be tempting to view microergodicity as a sufficient condition for consistent estimation, this is not the case. In fact, microergodicity is necessary but not sufficient. Additional conditions are required to ensure the existence of a consistent estimator. We focus our attention on Gaussian measures with the Matérn covariance function.

2.3 Asymptotics for the Matérn covariance

Recall the stationary and isotropic family of Matérn covariance functions in (1.1). Unfortunately, for mathematical and numerical convenience, we often assume the smoothness parameter ν is known and fixed, estimating only the variance σ^2 and range ρ parameters. However, recent work has shown that ν is also microergodic and consistent estimators for it can be constructed under various sampling designs (Loh et al., 2021; Loh and Sun, 2023). This research lies beyond the scope of this analysis and we focus on traditional estimation of $\theta = (\sigma^2, \phi)$. The following result identifies the non-microergodic and microergodic parameters in the covariance functions for Gaussian measures with fixed ν.

Proposition 2.3.1. *Let P_{θ_1} and P_{θ_2} be two probability measures such that both yield stationary Gaussian processes with mean 0 and isotropic Matérn covariances in $\mathcal{D} \subseteq \mathbb{R}^2$. Further assume that the processes share the same smoothness parameter ν but different variance and range parameters as in $\theta_i = (\sigma_i^2, \phi_i)$ for $i = 1, 2$. Then σ_i^2 and ϕ_i are non-microergodic parameters and $\kappa := \sigma_i^2/\phi_i^{2\nu}$ is a microergodic parameter.*

A proof using the spectral density of the Matérn covariogram can be found in Zhang (2004). By Lemma 2.2.1, it follows that σ^2 and ϕ cannot be consistently esti-

mated, unlike in the increasing domain case, where all parameters are estimable. In contrast, κ is microergodic and there is potential that we can estimate it consistently.

Theorem 3 in Zhang (2004) proved that the maximum likelihood estimator of the microergodic parameter, keeping ϕ fixed, is a consistent estimator for $\sigma^2/\phi^{2\nu}$. Kaufman and Shaby (2013) used the same result to extend consistency and asymptotic normality of the MLE maximizing over both σ^2 and ϕ, not just when ϕ is kept fixed. For a GRF with zero mean and zero nugget Matérn covariance, the log likelihood is

$$\log \mathcal{L}_n(\sigma^2, \phi; \mathbf{Y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_n|) - \frac{1}{2}\mathbf{Y}^\top \Sigma_n^{-1}\mathbf{Y}. \tag{2.1}$$

Let the MLE for the microergodic parameter be defined as $\hat{\kappa} := \hat{\sigma}^2/\hat{\phi}^{2\nu}$, where $\hat{\sigma}^2$ and $\hat{\phi}$ maximize (2.1) over $(0, \infty) \times [\phi_L, \phi_U]$ for any $0 < \phi_L < \phi_U < \infty$ where $\phi \in [\phi_L, \phi_U]$. We cite Theorem 2 from Kaufman and Shaby (2013) below.

Theorem 2.3.2. *Let S be a second order stationary Gaussian process with mean zero and isotropic Matérn covariance with parameter values ν, σ^2, ϕ, and $\kappa := \sigma^2/\phi^{2\nu}$. Let D_n be an increasing sequence of finite subsets of $\mathcal{D} \subseteq \mathbb{R}^2$ and $\hat{\kappa}$ be defined as the MLE for κ_0. Then as $n \to \infty$,*

1. *$\hat{\kappa} \to \kappa_0$ with probability 1,*
2. *$n^{1/2}(\hat{\kappa} - \kappa_0) \to N(0, 2\kappa_0^2)$ in distribution.*

A few remarks are in order. The proof of Theorem 2.3.2 builds upon the consistency argument in Theorem 3 of Zhang (2004), where ϕ is held fixed. The key distinction from standard M-estimator consistency proofs lies in the dependence structure of the data. Without independence, the SLLN and WLLN are unavailable. However, the final step to show that the continuously differentiable log-likelihood increases and decreases within an ϵ-neighborhood of the true parameter for any $\epsilon > 0$ remains the same. The concavity of the profile likelihood establishes the true consistency of the MLE, beyond merely proving the existence of a consistent root.

Instead of using LLNs to show log-likelihood convergence, Zhang (2004) applies a martingale convergence theorem from Gihman and Skorokhod (2004) to prove the convergence of the log-likelihood ratio. The limiting behavior depends on whether the underlying probability measures are equivalent or orthogonal: for equivalent measures, the log-likelihood ratio limit is 0, while for orthogonal measures, it converges to $-\infty$. This ensures that the log-likelihood increases and decreases around the true parameter.

Thus, beyond determining that non-microergodic parameters are inherently inconsistent, the concepts of equivalence and microergodicity are essential for establishing the consistency of microergodic parameters such as κ.

2.3.1 Extensions to nonzero nugget

The previous discussion assumed a zero nugget effect in (1.2). However, in applied geostatistics, it is far more common to estimate spatial models with a nonzero nugget, as its benefits for both estimation and prediction are well-documented (Diggle et al., 1998; Gramacy and Lee, 2012). While the consistency and asymptotic normality of the joint estimator for the nugget, covariance, and fixed effect parameters have been established under increasing-domain asymptotics (Mardia and Marshall, 1984; Bachoc, 2014, 2020), the spectral analysis techniques used in the previous section no longer apply due to the discontinuity in the covariogram introduced by a nonzero nugget.

Tang et al. (2021b) demonstrated that the joint MLE for τ^2 and σ^2, while keeping ϕ fixed—similar to the approach in Zhang (2004) rather than the joint estimation of σ^2 and ϕ as in Kaufman and Shaby (2013)—remains consistent and asymptotically normal. Their results indicate that introducing a nugget into the MLE causes the estimator for τ^2 to retain the expected $\sqrt{n}$ convergence rate, while the convergence rate of σ^2, which from Theorem 2.3.2 was also $\sqrt{n}$ (Du et al., 2009; Wang and Loh,

2011), slows to $n^{1/(2+2\nu)}$ in $\mathbb{R}^2$. This slowdown parallels findings for specific Ornstein-Uhlenbeck processes in Ying (1991) and Chen et al. (2000).

2.3.2 Extensions to fixed effects

Recent interest has focused on estimating the fixed effect parameter, particularly in the context of spatial confounding. Wang et al. (2020) showed that when the nugget is zero and the covariance structure is known, the GLS estimator for the fixed effect parameters β_z in (1.2) is inconsistent due to an asymptotically nonzero variance. However, results from Gilbert et al. (2024), Bolin and Wallin (2024), and an unpublished dissertation (Yu, 2022) indicate that fixed effect parameters can be consistently estimated under certain conditions. Specifically, a positive nugget term must be included in the estimated covariance matrix of the MLE, and the covariate surface associated with the fixed effect must exhibit some nonspatial variation (i.e., the variance of the covariate vector must be positive definite). In particular the standard case of smooth covariate surfaces, such as an intercept term, does not lead to an estimable fixed effect.

2.4 Summary

The purpose of asymptotics in spatial statistics is to derive approximations for finite-sample estimators. This has made the fixed-domain asymptotic framework far more prevalent than the increasing-domain framework, as the former provides better approximations in practice (Zhang and Zimmerman, 2005). However, fixed-domain asymptotics are inherently challenging, and while results exist for various scenarios, no general results are available for the simultaneous estimation of all parameters (Bachoc, 2020). Additionally, in most applications, the smoothness parameter ν is assumed to be known.

Despite the lack of theoretical results for the simultaneous estimation of all parameters, simulation studies and the combination of existing asymptotic results provide strong support for the use of MLE and GLS. While the variance, range, and fixed effect parameters are usually inconsistent, the nugget and microergodic parameter remain consistent and asymptotically normal, and fixed effects can be consistently estimated under certain conditions. Furthermore, kriging predictions using MLE-based parameter estimates exhibit desirable optimality properties (Stein, 1988; Wang et al., 2020).

Given this largely positive case for the MLE, an important question is how robust it and its approximations are when the data are preferentially sampled. We explore this question in the next chapter.

Chapter 3

When does geostatistical design matter? Insights into the effect of preferential sampling on the MLE

Preferential sampling (PS) occurs when the locations at which observations are made and the response of interest are statistically dependent. If ignored, PS can be detrimental to inference and prediction. The maximum likelihood estimator (MLE) assumes the locations are fixed and does not account for PS, leading to misspecification and bias. While it is well-established that the MLE is suboptimal under PS, the precise nature of the bias remains unclear. In this work, we study the finite and large sample behavior of the MLE under PS where the spatial process follows a stationary Gaussian process with Matérn covariance. We also conduct simulations to assess how PS affects the MLE, two of its popular approximations: pairwise composite likelihood and the Vecchia approximation, and the classic shared latent process model to adjust for PS. Our results show that although PS induces substantial bias to the MLE in small samples, its impact diminishes in large samples, particularly for spatially smooth random fields with long-range spatial dependence. These findings

provide valuable insights for applied spatial data analysis, offering guidance on when PS adjustments are necessary - or can be safely omitted.

3.1 Introduction

Geostatistics comprises a set of methods to infer properties and predict unknown values (or functions of values) of a spatially continuous process $\{S(\mathbf{x}) : \mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^2\}$ from a discrete set of observations taken at locations $\mathbf{X}_n = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$. Model-based approaches estimated via maximum likelihood estimation (MLE) typically assume S follows a Gaussian process, and subsequent inference and prediction treat $\mathbf{X}_n$ as fixed (Diggle et al., 1998). This assumption implies that the sampled locations are independent of the spatial process of interest, or in symbols $[\mathbf{X}, S] = [\mathbf{X}][S]$, where square brackets indicate the probability distribution and $\mathbf{X}$ is the point process generating $\mathbf{X}_n$. Violations of this independence are referred to as *preferential sampling* (PS). We will refer to sampling designs where $[\mathbf{X}, S] = [\mathbf{X}][S]$ as non-preferential sampling (NPS). There is strong evidence to suggest neglecting PS can heavily bias both parameter estimation and prediction (Diggle et al., 2010; Pati et al., 2011; Gelfand et al., 2012).

Adjustment for PS is widely acknowledged as an important consideration in spatial data analysis, yet the precise mechanism by which PS impacts the MLE has not been thoroughly formalized. Previous studies, such as the simulation study by Gelfand et al. (2012), examined the effect of PS on spatial prediction, primarily noting the poor spatial coverage of PS as a detriment to kriging, but without addressing the underlying nature of the impact of PS. Discussions surrounding the bias in inference under PS have mainly just acknowledged it as an issue, and typically only provide limited insight into the exact effects on mean and covariance parameter estimation. Notably, Dr. Michael Stein, in his response to Diggle et al. (2010), conjectured that

the influence of PS on covariance parameter estimation may be minimal for sufficiently large sample sizes. However, this conjecture lacks formal theoretical backing and has not been rigorously tested. In this work, we fill this gap by providing both theoretical and simulation evidence for the effect of PS on MLE prediction and estimation in finite and large samples when the underlying random field follows a stationary Gaussian process with Matérn covariance.

Given the computational complexity of the MLE, which scales at $O(n^3)$, practical applications often rely on approximations and are an active area of research (Heaton et al., 2019). Despite the widespread use of such approximations, little is known about how PS impacts their performance relative to that of the MLE. To address this, we also compare the MLE to two of its most well-known approximations: pairwise marginal composite likelihood (Bevilacqua and Gaetan, 2015) and the Vecchia approximation (Vecchia, 1988). Additionally, unlike previous PS studies that conduct their simulations on a single random field specification at one fixed small sample size, our study explores the effect of PS across multiple random fields with varying levels of smoothness and spatial dependence at several values for n. This broader range of settings allows for a deeper understanding of how PS impacts the analysis of spatial processes.

The organization of the paper is as follows. In Section 3.2, we present the theoretical results on the impact of PS on MLE estimation. Section 3.3 outlines the design of our simulation study. Section 3.4 reports the simulation results, and finally, Section 3.5 and 3.6 provide a broader discussion of the implications of our findings.

3.2 Theoretical Results

3.2.1 Background

Let S be a Gaussian process on bounded domain $\mathcal{D} \subseteq \mathbb{R}^2$. Suppose at locations $\mathbf{X}_n = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subseteq \mathcal{D}$ we observe values $\mathbf{Y} = \{\mu + S(\mathbf{x}_1) + Z_1, ..., \mu + S(\mathbf{x}_n) + Z_n\}$ where $Z_i \sim^{iid} N(0, \tau^2)$. We will further assume S is a mean zero stationary Gaussian process with a Matérn covariance model, which follows

$$C(h; \sigma^2, \phi, \nu) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\phi} h \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\phi} h \right), \tag{3.1}$$

where $\sigma^2, \phi, \nu > 0$, $h \geq 0$ is the Euclidean distance between any two locations, and K_ν is the modified Bessel function of the second kind. Each parameter governs a specific property of the random field S. The variance parameter σ^2 controls the dispersion, the range parameter ϕ determines how fast the correlation of values between locations decays with increasing distance, and the smoothness parameter ν decides the mean-square differentiability of S. Consequently, these parameters also influence the properties of the observations $\mathbf{Y}$, which are a function of S. The parameter τ^2, often referred to as the "nugget" in geostatistics, controls the measurement error or non-spatial variation in $\mathbf{Y}$ that is unexplained by S. Typically ν is assumed to be known, and the goal is then to estimate $\mathbf{\Psi} = (\mu, \sigma^2, \phi, \tau^2)$ and use the estimated parameters to predict $\mu + S(\mathbf{X}_0)$ at new locations $\mathbf{X}_0$ not in $\mathbf{X}_n$.

Let the MLE be defined as $\hat{\mathbf{\Psi}}_n := (\hat{\mu}_n, \hat{\sigma}_n^2, \hat{\phi}_n, \tau_n^2)$, which treats $\mathbf{X}_n$ as fixed and maximizes

$$\log \mathcal{L}_n(\mu, \sigma^2, \phi, \tau^2; \mathbf{Y}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_n|) - \frac{1}{2} (\mathbf{Y} - \mu \mathbf{1}_n)^\top \Sigma_n^{-1} (\mathbf{Y} - \mu \mathbf{1}_n), \tag{3.2}$$

where Σ_n is the covariance matrix with entries $C(\|\mathbf{x}_i - \mathbf{x}_j\|_2; \sigma^2, \phi, \nu) + \tau^2 \chi_{\{i=j\}}$ for C from (3.1). While the MLE is often considered the gold standard for geostatis-

tics, evaluation of the likelihood requires inversion of Σ_n which is a $O(n^3)$ operation, making it infeasible for many spatial applications.

Two key approximations of the MLE are the pairwise marginal composite likelihood (PMLE) and the Vecchia approximation. The PMLE is a composite likelihood (CL) method (Varin et al., 2011) that maximizes the product of all $\binom{n}{2}$ bivariate densities between elements of $\mathbf{Y}$ instead of the full likelihood in (3.2), trading off estimation efficiency for computational speed (Bevilacqua and Gaetan, 2015). By only considering second-order correlations in $\mathbf{Y}$, the PMLE achieves $O(n^2)$ computational complexity. The Vecchia approximation can also be considered a CL and is rooted in the simple fact that any joint distribution of $\mathbf{y} = (y_1, ..., y_n)$ can be factorized as a product of conditional distributions, or $f(\mathbf{y}) = f(y_1)f(y_2|y_1)...f(y_n|y_1, ..., y_{n-1})$ (Vecchia, 1988; Katzfuss and Guinness, 2021). Vecchia observed that much of the information in the conditioning sets with higher indices was likely to be redundant. One could attain a good tradeoff of efficiency for computational gain by setting a maximum size m for each conditioning set and being judicious about which variables to include. The hyperparameter m is typically chosen to be much smaller than n, and the resulting approximation to the joint likelihood can then be written as $f(\mathbf{y}; \mathbf{\Psi}) = f(y_{p(1)}; \mathbf{\Psi}) \prod_{i=2}^{n} f(y_{p(i)}|\mathbf{y}_{h(i)}; \mathbf{\Psi})$, where $p : \{1, ..., n\} \to \{1, ..., n\}$ is a permutation mapping which reorders $\mathbf{y}$ and $h(j) = \{l \in \{1, ..., n\} : p(l) < p(j)\}$ is the "history" or set of indices in the conditioning set of y_j. The Vecchia approximation achieves $O(m^3 n)$ computational complexity and factors in up to mth-order correlations within $\mathbf{Y}$, making it both more scalable and more accurate (more information used) compared to the PMLE.

The MLE and its approximations work well when the sampling design $\mathbf{X}$ is independent of S. Under PS, when there exists non-ignorable dependence between the observation locations and S, the MLE is misspecified and can result in bias. The classic model to adjust for PS induces dependence between the spatial process of interest

and the sampling locations through a shared variable approach (Ho and Stoyan, 2008; Diggle et al., 2010). We will refer to this model as the *shared latent process* model (SLP). Under the SLP, the sampling locations $\mathbf{X}_n$ are no longer assumed to be fixed and are instead a realization from a log Gaussian Cox process (LGCP) $\mathbf{X}$. Specifically, $[\mathbf{X}|S]$ follows an inhomogeneous Poisson process with intensity function equal to $\lambda(\mathbf{x}) = \exp\{\alpha + \beta S(\mathbf{x})\}$. The SLP takes its name from the presence of S in both the mean of $\mathbf{Y}$ and the intensity of $\mathbf{X}$. The parameter β can be interpreted as a single PS coefficient controlling the preferentiality of the SLP. The sign of β determines the direction of the preference, while its magnitude determines the likelihood of sampling extreme values.

While the SLP has been the focus of PS research due to its computational tractability, we can actually prove results for a much more general class of PS sampling designs. We now give some results on the estimation properties of the MLE when $\mathbf{X}_n$ may be sampled preferentially.

3.2.2 Parameter Estimation

We make the following additional assumption about the observation locations $\mathbf{X}_n$.

Assumption 1 (Point Process Conditions)**.** *The observation locations $\mathbf{X}_n$ are an increasing sequence of subsets of study region $\mathcal{D}$ sampled from point process $\mathbf{X}$. Additionally, the intensity function of $\mathbf{X}$ conditional on S, denoted as $\lambda(\cdot|S)$, satisfies $\lambda(\mathbf{x}|S) > 0$ for all $\mathbf{x} \in \mathcal{D}$.*

We now provide the following results on parameter estimation in preferentially sampled designs:

Theorem 1 (Inconsistency of $\hat{\sigma}^2$ and $\hat{\phi}$)**.** Let μ and ν be known. Given observations $\mathbf{Y}$ sampled as described in 3.2.1 and locations $\mathbf{X}_n$ satisfying Assumption 1, there do not exist estimators of σ^2 and ϕ that are consistent.

Proof. By Corollary 1 in Tang et al. (2021b), we need only check the observation locations represent an increasing sequence of subsets of $\mathcal{D}$, which follows by Assumption 1. □

Theorem 2 (Asymptotic normality of microergodic)**.** Suppose $(\sigma^2, \phi) \in (0, \infty) \times [\phi_L, \phi_U]$ for some $0 < \phi_L < \phi_U < \infty$. Let $\kappa := \sigma^2/\phi^{2\nu}$ and its MLE be $\hat{\kappa}_n := \hat{\sigma^2_n}/\hat{\phi}^{2\nu}_n$ where $(\hat{\sigma}^2_n, \hat{\phi}_n)$ maximize (3.2) over $(0, \infty) \times [\phi_L, \phi_U]$, assuming known μ, known ν and $\tau^2 = 0$. Then as $n \to \infty$

$$\text{(i) } \hat{\kappa}_n \to \kappa, \text{ almost surely, and}$$

$$\text{(ii) } n^{1/2}(\hat{\kappa}_n - \kappa) \to N(0, 2\kappa^2), \text{ in distribution.}$$

Proof. The sampling design satisfies the conditions needed for Theorem 2 in Kaufman and Shaby (2013) □

Theorem 4 and Theorem 5 from Tang et al. (2021b) and Theorem 3 from Wang et al. (2020) also extend to preferentially sampled $\mathbf{X}_n$ under Assumption 2.1. Since the reasoning is identical, we omit them here. These theorems provide conditions for the consistency and asymptotic normality of $\hat{\tau}^2_n$ and inconsistency of the generalized least squares (GLS) estimator for regression parameters for fixed covariates, respectively. It should be noted that recent work has demonstrated conditions on the covariate surface in which the GLS for regression parameters is consistent. In particular, Bolin and Wallin (2024), Gilbert et al. (2024) and Yu (2022) all show so long as the covariate surface is "rough" or noisy enough relative to the random field S, any slope parameters for these covariates are estimable by the GLS. In our work, we focus on the parameter associated with the smoothest possible covariate surface, the intercept μ, as the most challenging case of estimation.

These results should be of some comfort to the spatial data analyst relying on the MLE. Under Assumption 1, the asymptotic behavior of the MLE is unchanged

regardless of the sampling design. Therefore, if the main priority is to minimize differences in statistical inference between a NPS and PS design, a viable solution is to simply collect enough data while avoiding sampling designs which may exclude (but can still prefer) observations based on the value of S. Unfortunately, in many applications like air pollution monitoring or species distribution estimation, "collecting more data" is not always an option. The point at which n is large enough to reduce PS bias can also be difficult to identify in real applications. In that case, one must be aware of the substantial effect of PS on MLE performance in finite samples.

The most well-known and obvious impact of PS on the MLE is biased estimation of μ in finite samples (Diggle et al., 2010; Pati et al., 2011; Dinsdale and Salibian-Barrera, 2019a; Vedensky et al., 2023). Ho and Stoyan (2008) used point process theory to derive $E(\mathbf{Y}) = (\mu + \beta\sigma^2)\mathbf{1}_n$ when $(\mathbf{Y}, \mathbf{X}_n)$ are drawn from the SLP. Quantifying the same bias for the MLE is a much more difficult task. In finite samples, PS can also have an impact on the estimation of σ^2 and ϕ. Sampling from the SLP can lead to clusters of points with highly similar values. In that case, both σ^2 and ϕ are likely to be underestimated. This is because the sampling design results in less variability in $\mathbf{Y}$ and shorter distances between points in $\mathbf{X}_n$, making it more difficult to estimate the true dispersion of the response values and pinpoint the distance at which observations are no longer correlated.

3.2.3 Prediction

Once the parameters are estimated, prediction of values at new locations is achieved by means of *kriging*, also known as Gaussian process regression or spatial interpolation. Through application of the Gauss-Markov theorem and multivariate normal distribution theory, one can show that the best linear unbiased predictor (BLUP) conditional on $\mathbf{Y}$ for fixed $\mathbf{X}_n$ is equal to

$$\hat{S}(\mathbf{X}_0) = \mu + C(\mathbf{X}_0, \mathbf{X}_n)^\top \boldsymbol{\Sigma}_n^{-1}(\mathbf{Y} - \mu \mathbf{1}_n). \tag{3.3}$$

It is well-known that the least squares optimality property for kriging no longer applies under PS (Dinsdale and Salibian-Barrera, 2019a). Recall the best conditional least squares estimate of the value at an unobserved location $\mathbf{x}_0$ is the conditional expectation $\mathbb{E}(S(\mathbf{x}_0)|\mathbf{X}, \mathbf{Y})$. When $\mathbf{X}$ is fixed, the conditional expectation is equal to the linear predictor in (3.3), but is no longer true under PS. Therefore, kriging prediction with the true parameters will always be sub-optimal to $\mathbb{E}[S(\mathbf{x}_0)|\mathbf{X}, \mathbf{Y}]$ under a PS sampling design. This problem can be avoided by MCMC or Laplace approximation estimation approaches for correctly specified PS models to compute the mean or mode of $[S|\mathbf{X}, \mathbf{Y}]$ (Pati et al., 2011; Dinsdale and Salibian-Barrera, 2019a; Watson et al., 2019). It is important to note that predictions generated by these methods are not guaranteed to outperform kriging. If the estimate of the mean or mode of $[S|\mathbf{X}, \mathbf{Y}]$ is poor, kriging could still yield a better predictive surface.

3.3 Simulation Experiment

Theoretical study of the finite sample behavior and predictive performance of the MLE under PS is inherently challenging. The large sample results presented in the previous section also do not address the more practical scenario where all mean and covariance parameters are estimated simultaneously. To better understand the MLE in this context, we conducted a comprehensive simulation study evaluating the MLE fit over observations sampled from the SLP under various random field specifications. In addition to assessing the MLE, we compared its performance to the PMLE and Vecchia approximation. Since none of these methods directly adjust for PS, we included a "best-case scenario" benchmark - the SLP model fit in the `R-INLA` package (referred to here as INLA-SLP) as implemented in Dinsdale and Salibian-Barrera (2019a) and

Watson et al. (2019). This resulted in a comparison across four estimation methods.

For each of nine different Matérn random fields, we generated $B = 500$ realizations on the $[0, 1] \times [0, 1]$ unit square (Figure 3.4). Each random field shared the common parameters $\mu = 4$, $\sigma^2 = 1.5$, and $\tau^2 = 0.1$, but differed by the value of their range (ϕ) and smoothness (ν) parameters. We will refer to the different values of smoothness as rough ($\nu = 1/2$), neutral ($\nu = 1$), and smooth ($\nu = 3/2$) and that of the range by low ($\phi = 0.02$), medium ($\phi = 0.15$), and high ($\phi = 0.3$). Fields with low ϕ and low ν contain rapid changes and localized spatial dependence, whereas fields with high ϕ and high ν will tend to change very gradually and exhibit long range spatial dependence.

Each field was approximated by discretizing the study region into a 400×400 grid. We will refer to the values of S associated with the centroids of each discrete unit as $\mathbf{S}$. For every simulated realization, we collected the observations under both a NPS and PS sampling design, at different sample sizes of $N \in \{200, 800, 3200\}$. NPS locations were selected under complete spatial randomness (CSR) whereas PS locations were sampled according to the SLP model with $\beta = 1$, each cell with probability of inclusion equal to $p(\mathbf{x}) = \frac{\exp\{\beta \mathbf{S}(\mathbf{x})\}}{\sum_{s \in \mathbf{S}} \exp\{\beta s\}}$.

We then fit the PMLE, Vecchia approximation, MLE, and INLA-SLP over each point pattern to estimate the random field parameters in $\mathbf{\Psi}$ with the exception of ν which was fixed. The PMLE and MLE were both fit by a custom implementation and solved through the `L-BFGS-B` algorithm in `optim`. The Vecchia approximation was implemented in the `GpGp` package under the default arguments, with $m = 30$ neighbors and maxmin ordering (Guinness, 2018, 2021). The INLA-SLP was implemented with the SPDE approach in the `R-INLA` package and we computed the mesh with 32^2 points (Rue et al., 2009; Lindgren et al., 2011). We used the penalised complexity prior framework (Simpson et al., 2017) to define priors on our covariance parameters. For the variance and range, we set $P(\sigma^2 > 10) = 0.01$ and $P(\phi < 0.01) = 0.01$. The

prior for μ was set to the default Gaussian with precision equal to 0.001, and the prior for τ^{-2} was set to the default gamma prior.

Predictions for the PMLE, MLE and Vecchia approximation were calculated by plugging in each method's estimated parameters into the kriging equations (3.3). For the INLA-SLP, predictions were calculated from estimating the mode of $[S|\mathbf{X}, \mathbf{Y}]$ at each prediction location. Prediction error was evaluated by root mean squared prediction error (RMSPE) from predicting $\mathbf{S}$ over the entire 400×400 grid.

3.4 Results

Figure 3.1 compares covariance parameter estimation for all methods between NPS and PS sampling designs. Because the trends were largely similar between all nine of the random fields, for conciseness, we only show the results for the field with $\nu = 1/2$ and $\phi = 0.15$ and compare the estimates between the PS and NPS designs. The one exception was the random field with $\nu = 1$ and $\phi = 0.15$ (corresponding to the central square in Figure 3.4), which has been the random field of choice for comparing the MLE and SLP (Diggle et al., 2010; Dinsdale and Salibian-Barrera, 2019a). For this random field, the INLA-SLP estimates for all covariance parameters were approximately unbiased with impressive results (Figures 3.5, 3.6, 3.7). However we found this was a best-case scenario, and the INLA-SLP's estimates for σ^2, ϕ, and τ^2 were much more biased under the other eight random fields. We picked a different random field specification for Figure 3.1 from what has been traditionally used to illustrate how a slight change in the spatial process can lead to bias in the INLA-SLP's covariance parameter estimation.

The MLE estimates for the microergodic parameter κ and nugget parameter τ^2 were consistent under both sampling designs. Estimation for κ was similar between NPS and PS while estimation for τ^2 was more efficient under PS. On the other hand,

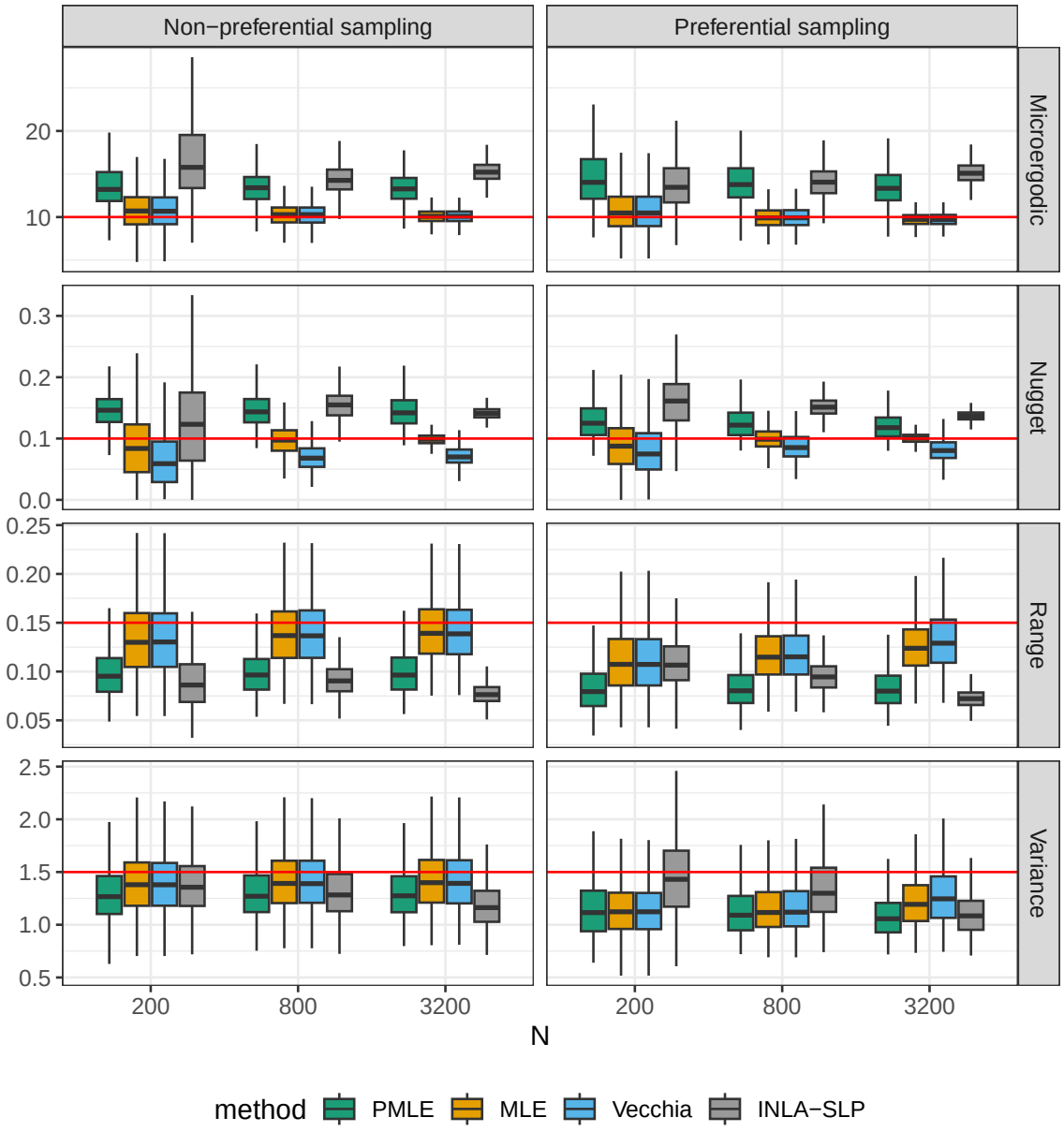

Figure 3.1: Simulation results for estimation of covariance parameters under the random field with medium range and rough smoothness ($\phi = 0.15$ and $\nu = 1/2$).

the MLE estimates for σ^2 and ϕ were underestimated and the variance of the estimates did not shrink with increasing n. The simulation implies the consistency of κ and τ^2 and inconsistency of σ^2 and ϕ hold in the case of full estimation of all parameters. As expected, MLE estimates for σ^2 and ϕ were lower under PS relative to NPS. Even though the MLE for σ^2 and ϕ are inconsistent, it does not mean the data contain

no information on these two parameters as the estimates did improve and begin to get closer to the truth with increasing n. The Vecchia approximation estimates of all covariance parameters were close to the MLE, with the exception of τ^2. Poor nugget estimation is a noted issue with `GpGp`'s Fisher scoring learning algorithm. We do not pursue the matter further since it did not affect subsequent results. Surprisingly, both the INLA-SLP and PMLE overestimated κ and τ^2 considerably. Their estimates of σ^2 and ϕ also tended to be lower compared to that of the MLE and Vecchia approximation.

While bias in μ estimation under NPS was low for all methods (Table 3.3), Figure 3.2 shows the estimation of μ using PS data under each of the nine random fields. As expected, all methods except the INLA-SLP overestimated the mean at smaller n. Interestingly, both the MLE and Vecchia approximation improved as sample size increased, with MLE's bias decreasing faster (Table 3.1). The PMLE, however, exhibited the highest bias and showed little improvement with increasing n. All methods were less biased for random fields with higher range and smoothness, though this came at the cost of increased variance in their estimates. The variance in estimates remained largely unchanged with increasing n, reinforcing the prior inconsistency result for $\hat{\mu}_n$, even when estimating all parameters simultaneously.

Figure 3.3 compares the predictive performance across the four methods. The PMLE had by far the highest error while the INLA-SLP consistently outperformed the others across all nine random fields. Despite the differences in parameter estimation noted earlier, the Vecchia approximation's RMSPE was nearly identical to that of the MLE, demonstrating its high approximation accuracy. As with μ estimation the RMSPE for the MLE/Vecchia approached the level of the INLA-SLP at $n = 3200$, becoming nearly indistinguishable for fields with $\nu \geq 1$ and $\phi \geq 0.15$. RMSPE decreased for all methods as ν and ϕ increased, with the MLE and INLA-SLP predictions differing the least for fields with high range ($\phi = 0.30$) and the most for

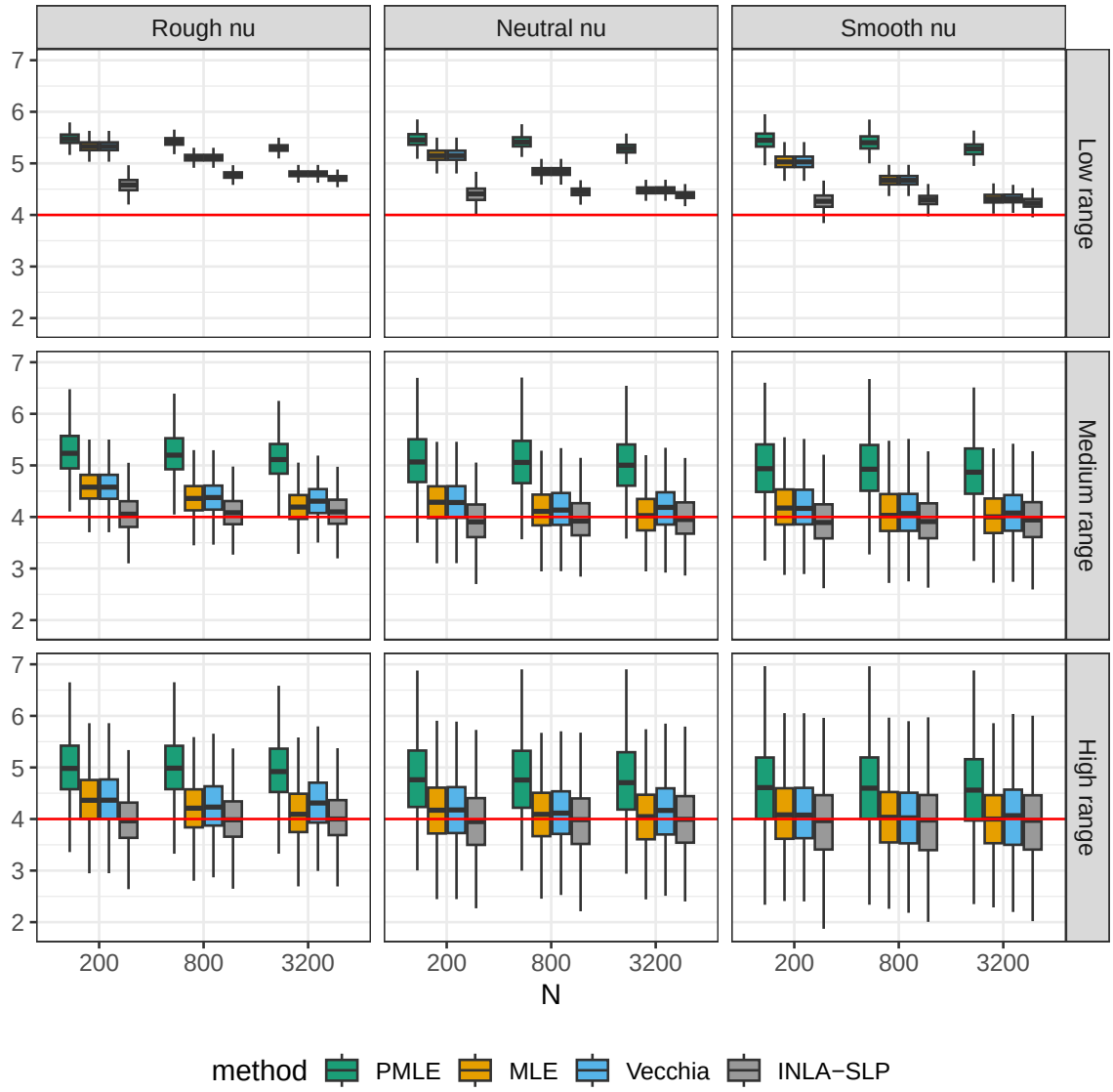

Figure 3.2: Simulation results for estimation of μ under all nine PS sampling designs. Rows indicate the value of the range (ϕ) while columns indicate the value of the smoothness (ν). The horizontal red line indicates the true value.

fields with low range ($\phi = 0.02$) (Table 3.2).

3.5 Discussion

In this work, we evaluated the impact of PS on the MLE, two of its approximations, and the SLP model. We found that while PS introduces substantial bias in inference

Table 3.1: Bias (RMSE) for estimation of μ under preferential sampling by the MLE for the simulation study.

	N	$\phi = 0.02$	$\phi = 0.15$	$\phi = 0.30$
$\nu = 1/2$	100	1.401 (1.41)	0.733 (0.83)	0.497 (0.76)
	200	1.327 (1.33)	0.585 (0.69)	0.380 (0.67)
	400	1.239 (1.24)	0.469 (0.59)	0.293 (0.62)
	800	1.111 (1.11)	0.359 (0.50)	0.215 (0.59)
	1600	0.967 (0.97)	0.271 (0.43)	0.155 (0.56)
	3200	0.798 (0.80)	0.194 (0.39)	0.115 (0.55)
$\nu = 1$	100	1.273 (1.28)	0.432 (0.65)	0.259 (0.74)
	200	1.156 (1.16)	0.297 (0.55)	0.173 (0.69)
	400	1.016 (1.02)	0.209 (0.51)	0.130 (0.66)
	800	0.840 (0.85)	0.136 (0.47)	0.084 (0.65)
	1600	0.660 (0.67)	0.086 (0.45)	0.053 (0.65)
	3200	0.476 (0.48)	0.049 (0.44)	0.030 (0.65)
$\nu = 3/2$	100	1.179 (1.19)	0.278 (0.62)	0.192 (0.79)
	200	1.032 (1.04)	0.193 (0.56)	0.123 (0.75)
	400	0.865 (0.87)	0.120 (0.54)	0.077 (0.72)
	800	0.671 (0.68)	0.071 (0.52)	0.040 (0.71)
	1600	0.485 (0.50)	0.037 (0.50)	0.033 (0.70)
	3200	0.316 (0.33)	0.019 (0.50)	0.010 (0.69)

and prediction of the MLE in finite samples, its effect diminishes in the long run, with the bias being less pronounced for observational point patterns sampled from spatial processes S exhibiting a high degree of smoothness (high ν) and strong long-range spatial dependence (high ϕ).

Specifically, we found that the MLE's asymptotic properties for covariance parameter estimation are unaffected by the sampling design. Under PS, the MLE's estimates of $\kappa := \sigma^2/\phi^{2\nu}$ and τ^2 are still consistent and asymptotically normal, while those for σ^2 and ϕ remain inconsistent, so long as every point in the study region has a positive sampling intensity regardless of the underlying random field. The effect of PS is most concerning in finite samples. Both σ^2 and ϕ are more heavily underestimated compared to when the data are NPS. This underestimation is a result of the clustering of similarly valued observations under the PS sampling design, leading to

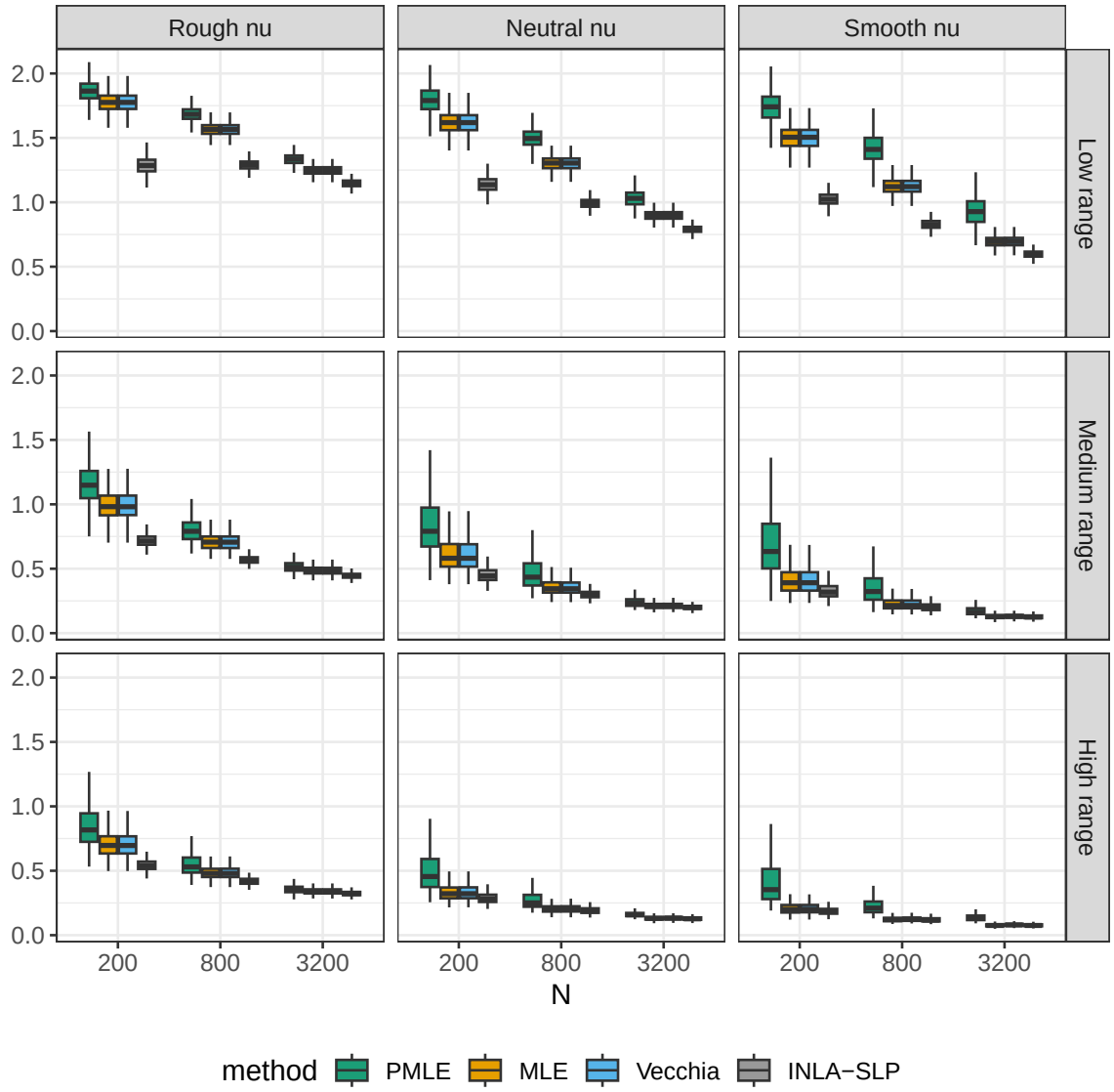

Figure 3.3: Simulation results for RMSPE over the entire grid under all nine PS sampling designs. Rows indicate the value of the range (ϕ) while columns indicate the value of the smoothness (ν).

lower variation in response values and shorter distances between points. The bias in estimation of μ is the most well-known in the PS literature. A key observation from our simulation was that the MLE bias for μ decreased with increasing sample size across all random fields, suggesting possible asymptotic unbiasedness irregardless of sampling design. This bias correction was most effective for random fields with high ϕ and ν. It is worth emphasizing that this property should not be taken for granted as

Table 3.2: Comparison of RMSPE for INLA-SLP and MLE for the simulation study. Values are shown as mean (SD).

	N	$\phi = 0.02$		$\phi = 0.15$		$\phi = 0.30$	
		INLA-SLP	MLE	INLA-SLP	MLE	INLA-SLP	MLE
$\nu = 1/2$	100	1.243 (0.14)	1.848 (0.11)	0.806 (0.07)	1.179 (0.16)	0.628 (0.07)	0.861 (0.17)
	200	1.287 (0.07)	1.776 (0.08)	0.720 (0.05)	1.004 (0.12)	0.547 (0.05)	0.712 (0.11)
	400	1.311 (0.06)	1.690 (0.06)	0.643 (0.04)	0.854 (0.09)	0.480 (0.04)	0.593 (0.09)
	800	1.291 (0.04)	1.565 (0.05)	0.569 (0.03)	0.712 (0.07)	0.421 (0.03)	0.490 (0.06)
	1600	1.240 (0.04)	1.421 (0.04)	0.505 (0.02)	0.591 (0.05)	0.371 (0.02)	0.411 (0.04)
	3200	1.147 (0.03)	1.247 (0.03)	0.446 (0.02)	0.488 (0.03)	0.325 (0.02)	0.344 (0.03)
$\nu = 1$	100	1.195 (0.08)	1.733 (0.11)	0.557 (0.11)	0.789 (0.19)	0.366 (0.08)	0.463 (0.16)
	200	1.139 (0.06)	1.619 (0.09)	0.458 (0.07)	0.612 (0.14)	0.292 (0.05)	0.348 (0.12)
	400	1.079 (0.05)	1.480 (0.07)	0.372 (0.05)	0.475 (0.10)	0.238 (0.05)	0.270 (0.06)
	800	0.994 (0.04)	1.304 (0.05)	0.306 (0.05)	0.362 (0.07)	0.196 (0.04)	0.209 (0.04)
	1600	0.904 (0.03)	1.111 (0.05)	0.251 (0.04)	0.278 (0.04)	0.159 (0.03)	0.166 (0.03)
	3200	0.790 (0.03)	0.899 (0.04)	0.203 (0.02)	0.216 (0.03)	0.130 (0.02)	0.133 (0.02)
$\nu = 3/2$	100	1.107 (0.07)	1.645 (0.12)	0.435 (0.11)	0.575 (0.19)	0.257 (0.08)	0.306 (0.14)
	200	1.026 (0.05)	1.500 (0.09)	0.335 (0.09)	0.420 (0.13)	0.192 (0.04)	0.217 (0.08)
	400	0.936 (0.04)	1.331 (0.07)	0.266 (0.06)	0.312 (0.09)	0.155 (0.04)	0.165 (0.05)
	800	0.829 (0.04)	1.127 (0.06)	0.208 (0.04)	0.232 (0.06)	0.122 (0.03)	0.127 (0.03)
	1600	0.719 (0.03)	0.914 (0.05)	0.167 (0.04)	0.174 (0.04)	0.096 (0.02)	0.097 (0.02)
	3200	0.599 (0.03)	0.696 (0.04)	0.129 (0.02)	0.132 (0.02)	0.078 (0.01)	0.077 (0.01)

the PMLE showed high bias and minimal improvement in μ estimation even at large n. Estimation of μ for the Vecchia approximation was much better compared to that of the PMLE, but still inferior to the MLE. This suggests the possible asymptotic unbiasedness of the MLE is directly related to the amount of dependence accounted for in the likelihood. Whereas the PMLE only accounts for 2nd order correlations and the Vecchia approximation accounts for at most mth-order correlations, the MLE operates over the entire joint likelihood of $\mathbf{Y}$. Lastly, while the INLA-SLP outpredicted the MLE in every scenario (as expected), at large n, RMSPE between the two methods were nearly identical, especially for smooth, highly correlated fields. With enough observation locations, the MLE appears to improve enough to achieve similar performance to the correctly specified INLA-SLP.

A byproduct of our analysis was the discovery that the PMLE is ill-suited for geostatistics, especially under PS. PMLE estimates of both mean and covariance parameters were heavily biased at all N, leading to poor predictions of the spatial field. Because the Vecchia approximation achieves better performance while also

being much more scalable, we strongly recommend its use over the PMLE when a MLE approximation is needed.

3.6 Conclusion

While preferential sampling is important to address in finite samples, our results suggest that the MLE and Vecchia approximation are likely robust to its effects in large samples, particularly when the underlying spatial process exhibits a high degree of spatial continuity and retains significant correlation over large distances. In addition, while the intercept term μ has been proven to be non-estimable, we provide simulation evidence that the MLE for this parameter may be asymptotically unbiased with finite variance even under preferential sampling, providing additional justification for the use of spatial regression for estimating intercept and slope parameters. We believe these findings provide valuable insights for applied spatial data analysis, and can guide the choice of appropriate geostatistical methods when preferential sampling may be present.

3.7 Supplementary Material

Table 3.3: Bias (RMSE) for estimation of μ for the MLE under non-preferential sampling.

	N	$\phi = 0.02$	$\phi = 0.15$	$\phi = 0.30$
$\nu = 1/2$	100	0.003 (0.14)	-0.025 (0.37)	-0.046 (0.57)
	200	-0.001 (0.11)	-0.024 (0.36)	-0.026 (0.55)
	400	-0.004 (0.08)	-0.032 (0.34)	-0.032 (0.54)
	800	-0.002 (0.07)	-0.027 (0.34)	-0.026 (0.54)
	1600	-0.002 (0.07)	-0.032 (0.33)	-0.034 (0.53)
	3200	-0.002 (0.06)	-0.030 (0.33)	-0.040 (0.59)
$\nu = 1$	100	-0.008 (0.15)	-0.029 (0.48)	-0.004 (0.68)
	200	-0.010 (0.12)	-0.028 (0.46)	-0.008 (0.66)
	400	-0.009 (0.10)	-0.027 (0.45)	-0.005 (0.65)
	800	-0.006 (0.09)	-0.026 (0.45)	-0.009 (0.65)
	1600	-0.007 (0.09)	-0.031 (0.44)	-0.014 (0.65)
	3200	-0.008 (0.09)	-0.028 (0.44)	-0.013 (0.64)
$\nu = 3/2$	100	-0.004 (0.16)	-0.029 (0.54)	-0.006 (0.76)
	200	-0.012 (0.14)	-0.029 (0.52)	-0.008 (0.73)
	400	-0.006 (0.11)	-0.028 (0.51)	-0.008 (0.72)
	800	-0.007 (0.11)	-0.027 (0.51)	-0.012 (0.70)
	1600	-0.008 (0.10)	-0.029 (0.51)	-0.016 (0.69)
	3200	-0.008 (0.10)	-0.025 (0.50)	-0.005 (0.70)

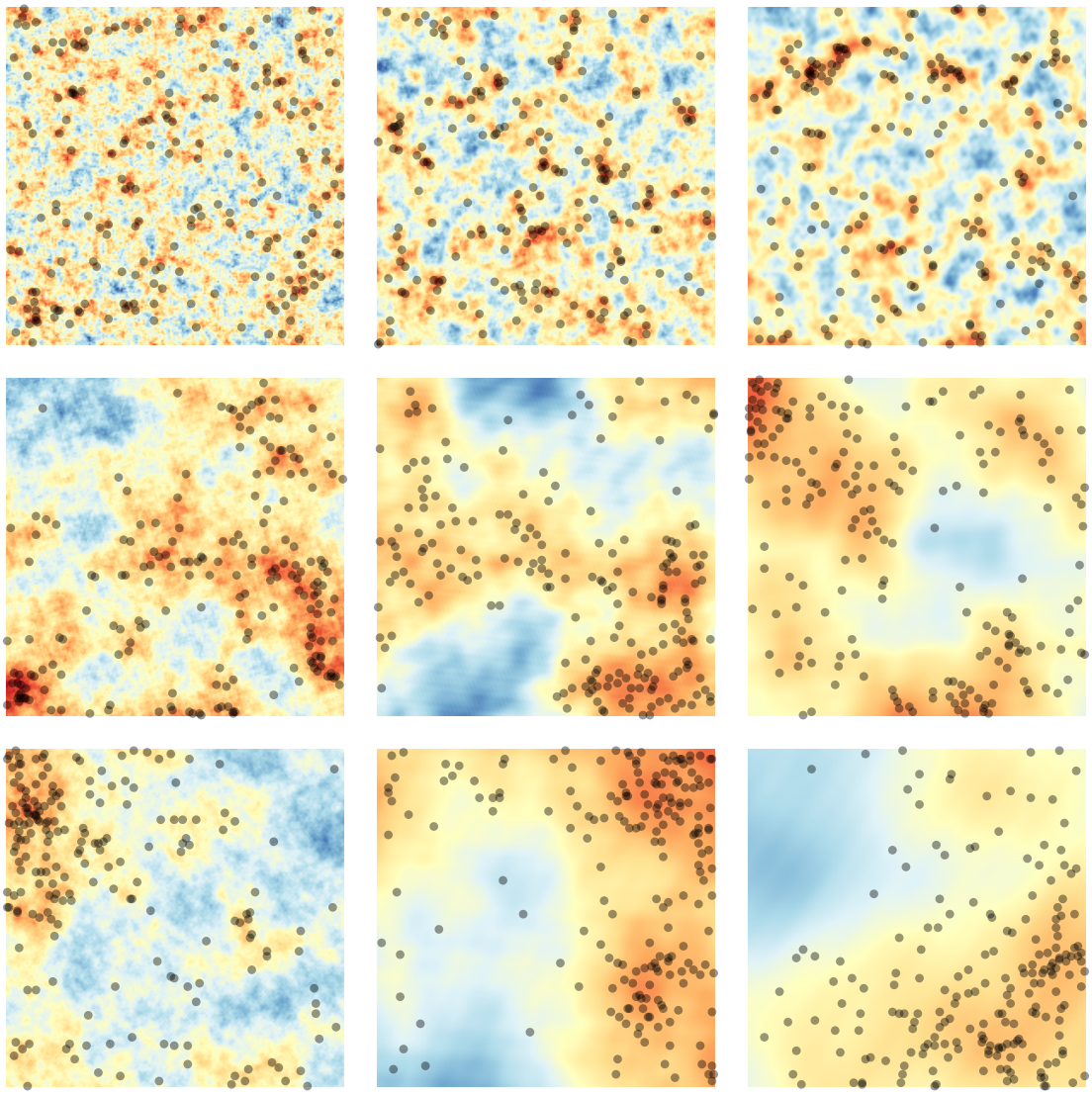

Figure 3.4: Realizations for the nine different specifications for S in the simulation experiment, with an example $n = 200$ point pattern sampled according to the SLP with $\beta = 1$. Rows represent the three different range parameters (top to bottom: $\phi \in \{0.02, 0.15, 0.30\}$) and columns represent the three different smoothness parameters (left to right: $\nu \in \{1/2, 1, 3/2\}$).

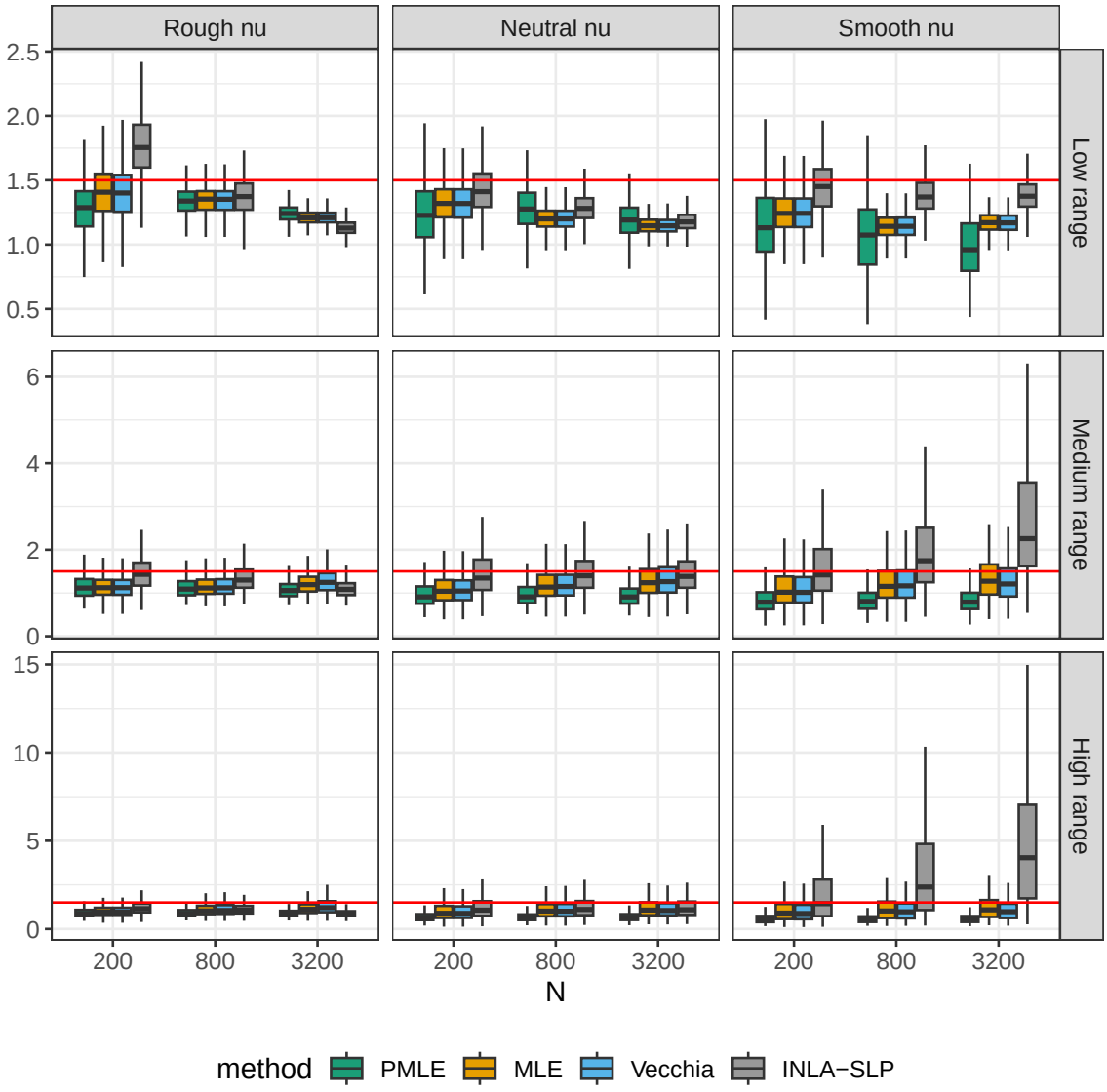

Figure 3.5: Simulation results for the variance (σ^2) over the entire grid under all nine PS sampling designs. Rows indicate the value of the range (ϕ) while columns indicate the value of the smoothness (ν).

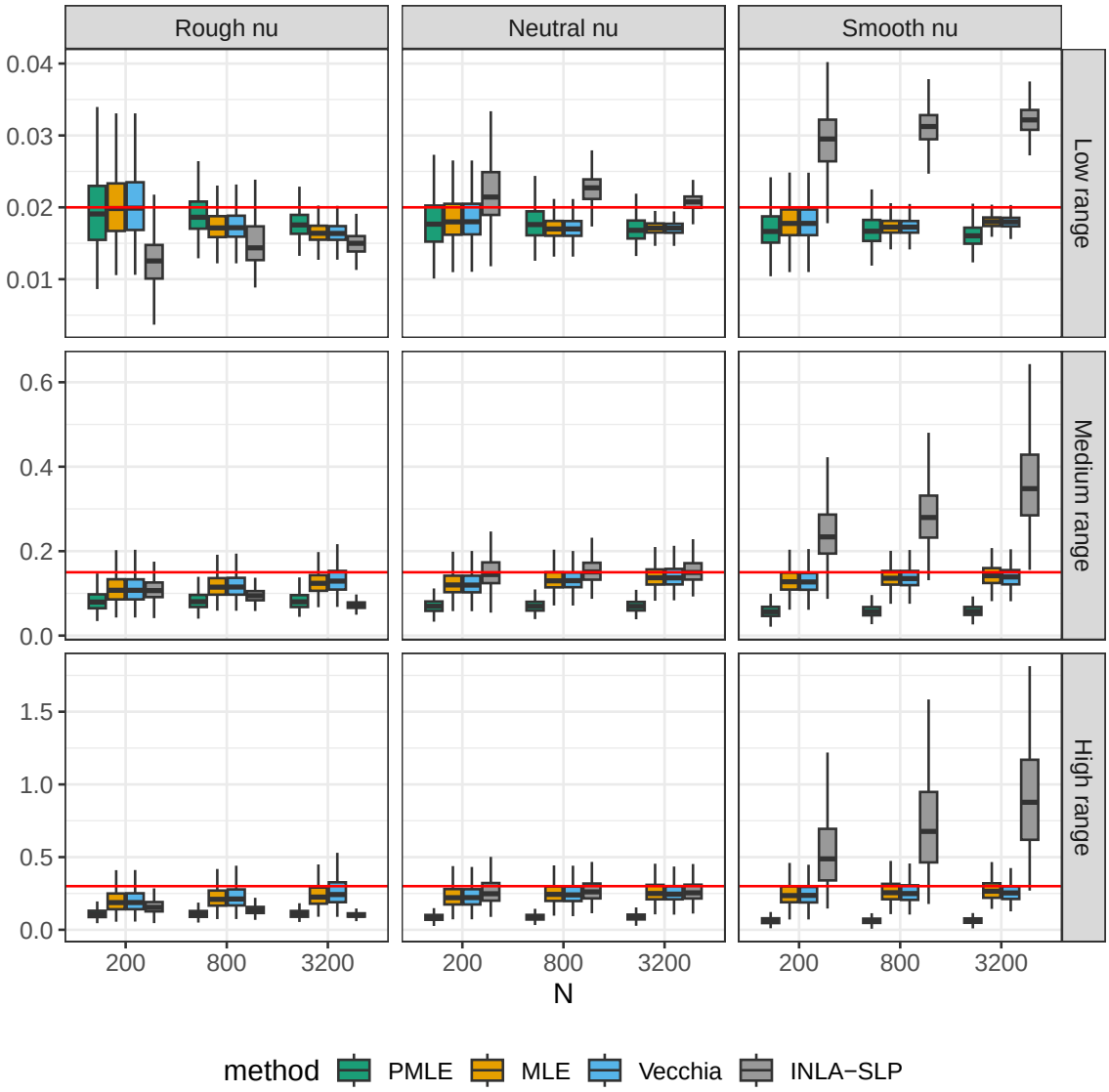

Figure 3.6: Simulation results for range (ϕ) over the entire grid under all nine PS sampling designs. Rows indicate the value of the range (ϕ) while columns indicate the value of the smoothness (ν).

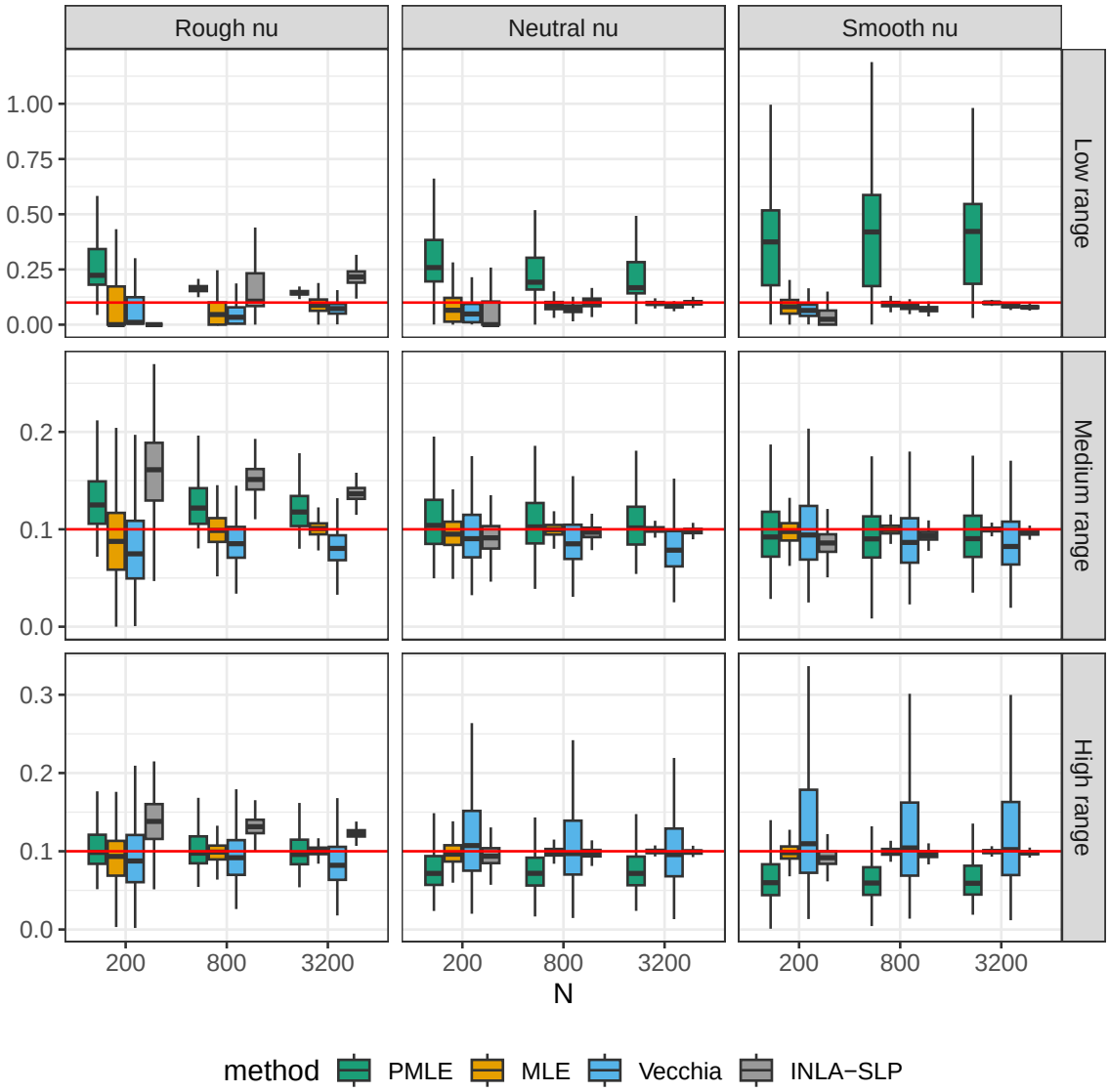

Figure 3.7: Simulation results for the nugget (τ^2) over the entire grid under all nine PS sampling designs. Rows indicate the value of the range (ϕ) while columns indicate the value of the smoothness (ν).

Chapter 4

Preferential sampling adjustment using inverse sampling intensity weights (ISIW)

Traditional geostatistical methods assume independence between observation locations and the spatial process of interest. Violations of this independence assumption are referred to as preferential sampling (PS). Standard methods to address PS rely on estimating complex shared latent variable models and can be difficult to apply in practice. We study the use of inverse sampling intensity weighting (ISIW) for PS adjustment in model-based geostatistics. ISIW is a two-stage approach wherein we estimate the sampling intensity of the observation locations then define intensity-based weights within a weighted likelihood adjustment. Prediction follows by substituting the adjusted parameter estimates within a kriging framework. A primary contribution was to implement ISIW by means of the Vecchia approximation, which provides large computational gains and improvements in predictive accuracy. Interestingly, we found that accurate parameter estimation had little correlation with predictive performance, raising questions about the conditions and parameter choices driving

optimal implementation of kriging-based predictors under PS. Our work highlights the potential of ISIW to adjust for PS in an intuitive, fast, and effective manner.

4.1 Introduction

The field of geostatistics comprises a set of methods to infer properties and predict unknown values (or functions) of a spatially continuous process $\{S(\mathbf{x}) : \mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^2\}$ from a discrete set of observation locations, denoted $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Model-based approaches estimated via maximum likelihood estimation (MLE) typically assume S is a realization from a random stochastic process, and subsequent inference and prediction treat $\mathbf{X}$ as fixed (Diggle et al., 1998). This assumption further implies that the sampled locations are independent of the spatial process of interest, or in symbols, $[\mathbf{X}, S] = [\mathbf{X}][S]$, where square brackets indicate the probability distribution. Violations of this independence assumption are referred to as *preferential sampling* (PS). There is strong evidence to suggest neglecting PS can negatively impact both geostatistical inference and prediction (Diggle et al., 2010; Pati et al., 2011; Gelfand et al., 2012).

Early works in PS examined $\mathbf{X}$ under the lens of a marked point process. Schlather et al. (2004) designed two Monte Carlo tests to detect dependence between marks and their locations. Ho and Stoyan (2008) introduced the classic PS adjustment model, referred to here as the *shared latent process* (SLP) model, and derived several of its properties. Full estimation and prediction procedures for the SLP model were formally established in the seminal work by Diggle et al. (2010).

The SLP framework induces dependence between the spatial process of interest and the sampling locations through a shared variable approach. Let μ, α, β, and τ^2 be scalars in $\mathbb{R}$, $\mathbf{Y}$ denote the $n \times 1$ random vector of observed values sampled at locations $\mathbf{X}$, S the shared latent process, $S(\mathbf{X})$ the values of S evaluated at locations $\mathbf{X}$,

IPP an inhomogeneous Poisson process, GP a Gaussian process, $C_\theta(\cdot,\cdot)$ a stationary covariance function indexed by parameter vector θ, and $\lambda(\cdot)$ the intensity function of $\mathbf{X}$ conditional on S. Then the SLP model is as follows:

$$\begin{aligned}
[\mathbf{Y}|\mathbf{X},S] &\sim N_n(\mu\mathbf{1}_n + S(\mathbf{X}), \tau^2 I_n),\\
[\mathbf{X}|S] &\sim IPP(\lambda),\\
\lambda(\mathbf{x}) &= \exp(\alpha + \beta S(\mathbf{x})),\\
S &\sim GP(0, C_\theta(\cdot,\cdot)).
\end{aligned} \tag{4.1}$$

While the SLP framework has been extended to handle covariates in the mean model, non-Gaussian likelihoods, and multiple shared latent processes (Pati et al., 2011; Watson et al., 2019), for the purposes of illustration, we focus on the original formulation of a constant mean, Gaussian likelihood, and one shared latent process. Under the SLP model, sampling locations are no longer assumed to be fixed and are instead a realization from a log Gaussian Cox process (LGCP). The latent process S is assumed to follow a zero mean second order stationary Gaussian process (GP) and for the remainder of this paper, we further assume C_θ to be an isotropic Matérn covariance function. The SLP takes its name from the presence of S in both the mean model of $\mathbf{Y}$ and the intensity function of $\mathbf{X}$. The parameter β can be interpreted as a single PS coefficient controlling the preferentiality of the SLP. The sign of β determines the direction of the preference, while its magnitude determines the likelihood of sampling extreme values.

Subsequent research on PS methods has remained largely faithful to the SLP model. Pati et al. (2011) extended the SLP model structure to a flexible Bayesian framework estimated by MCMC, and proved the posterior consistency of each of the mean, covariance and PS parameters under increasing domain asymptotics. Gelfand et al. (2012) also used Bayesian estimation, and introduced a framework to compare

prediction surfaces under PS between different methods. A surprising finding from their simulation analysis was that inclusion of an informative covariate was not sufficient to correct for predictive bias. This empirical finding highlights the importance of the SLP model even when informative covariates explaining the dependence between $\mathbf{X}$ and S are available. Dinsdale and Salibian-Barrera (2019a) significantly improved the computational efficiency of the SLP model relative to the previous simulation-based methods by means of the stochastic partial differential equations (SPDE) approach (Lindgren et al., 2011) combined with a Laplace approximation implemented in the Template Model Builder (TMB) `R` package (Kristensen et al., 2016). Watson et al. (2019) also fitted the SLP model using the SPDE approach but with an integrated nested Laplace approximation implemented in the `R-INLA` software (Rue et al., 2009), which enjoys comparable computational gains to the approach in Dinsdale and Salibian-Barrera (2019a). The authors further defined a framework to model PS spatio-temporal data and better emulate the evolution of $\mathbf{X}$ over time compared to the original SLP model defined in (4.1) above.

Several other contributions extend the SLP framework to explore a rich set of important applications, including air pollution monitoring (Lee et al., 2011, 2015), species distribution modeling (Manceur and Kühn, 2014; Fithian et al., 2015; Conn et al., 2017; Pennino et al., 2019; Gelfand and Shirota, 2019; Fandos et al., 2021), disease surveillance (Rinaldi et al., 2015; Cecconi et al., 2016; Conroy et al., 2023), phylodynamic inference (Karcher et al., 2016), hedonic modelling (Paci et al., 2020), bivariate spatial data (Shirota and Gelfand, 2022), spatially-varying PS (Amaral et al., 2023), optimal design under PS (Ferreira and Gamerman, 2015), space-filling designs (Ferreira, 2020; Gray and Evangelou, 2023), a hypothesis test to detect PS in spatio-temporal data (Watson, 2021), and exact Bayesian inference for the SLP model (Moreira and Gamerman, 2022; Moreira et al., 2023). The improved predictive ability and parameter estimation within the SLP framework contribute to the popularity

of model-based approaches accounting for dependence between spatial processes and their observation locations.

While the benefits of the SLP structure and the ubiquity of PS data support its widespread adoption for spatial data analysis, the SLP model still contains significant limitations. Few software packages and out-of-the-box implementations exist for SLP estimation, often coming in the form of custom INLA or MCMC sampler code. Furthermore, the SLP model is challenging to integrate with modern spatial approximation techniques, many of which have become the standard in spatial data analysis due to their scalability and feasibility, offering computationally efficient alternatives to the MLE while maintaining high accuracy (Heaton et al., 2019). Among these methods, only the SPDE approach has a well-documented implementation of a PS solution, and incorporating the SLP model into any other method would take a substantial effort. The SLP model requires joint evaluation of both the response and observation likelihood, not a trivial task. Prediction based on the SLP model also can suffer from high computational cost. While parameter estimation for the SLP model may scale well with the SPDE approach, prediction depends on summarizing the posterior distribution of $[S|\mathbf{X}, \mathbf{Y}]$ for each prediction point, which can be infeasible even on a moderately sized grid for both the Laplace approximation and MCMC.

An alternative strategy to the SLP is to incorporate dependence through the sampling intensities λ of the observations, rather than the entire likelihood of $\mathbf{X}$. These methods have been developed to much less fanfare compared to the SLP, and it remains unclear how well they work in geostatistical applications. The two main uses of sampling intensities are as a covariate in the mean model of $\mathbf{Y}$ or as inverse weights in a likelihood adjustment (Vedensky et al., 2023). A potential advantage of sampling intensity methods is their robustness to the form of PS relative to the SLP framework, which addresses a very specific type of clustering process.

The biggest impediment to the use of the sampling intensity is the need to esti-

mate this intensity from the observed locations, the driving reason such approaches have been sidelined in favor of full model-based approaches. Unlike in survey methodology where survey weights are fixed and known, sampling weights in geostatistical applications are largely unknown to the investigator and must be estimated. Unfortunately, nonparametric kernel smoothing estimators of the intensity (Diggle, 1985; Berman and Diggle, 1989) have few theoretical guarantees and are not consistent for the true intensity (Guan, 2008).

Even so, there is preliminary evidence that methods using nonparametrically estimated sampling intensities can still mitigate the effects of PS and, in some scenarios, can predict a surface better than the SLP approach in practice. Reich and Fuentes (2012) noted how the Bayesian SLP model parameters could be estimated without MCMC by replacing the shared latent process term in the mean model with some function of the sampling intensity $g(\lambda)$. Since the dependence between $\mathbf{X}$ and S is completely contained within $g(\lambda)$, the locations $\mathbf{X}$ can once again be treated as fixed, so long as λ is estimated well. They show improved prediction using the sampling intensity covariate in a 1-dimensional kriging example. In another example, Zidek et al. (2014) provide an approach using estimated weights for air pollution monitoring. Instead of working in a continuous study region, however, these authors considered a finite superpopulation of possible sampling sites and the probability of selecting any site over time was modeled by a logistic regression. The estimated inverse probabilities were then used as weights in a Horvitz-Thompson style design-based estimator for unbiased estimation of parameters under PS. In a third example, Schliep et al. (2023) also considered a finite superpopulation and similarly deviated from traditional model-based geostatistics by using estimated sampling intensity weights to recover the estimated parameter values had the model been fit on the superpopulation, rather than estimate the parameters of the spatial process S. These authors also derived a kriging variance estimate adjusted for PS. Finally, Vedensky et al. (2023) conducted

a simulation experiment comparing a univariate marginal composite likelihood (CL) weighted by inverse sampling intensities estimated by the `MASS::kde2d` function in `R` to the SLP and unweighted CL, and showed improved performance compared to no adjustment. We will refer to methods using inverse sampling intensities as weights in a weighted likelihood adjustment as *inverse sampling intensity weighting* (ISIW).

Despite preliminary investigations into ISIW for PS adjustment, specifics regarding the effectiveness of such approaches within model-based geostatistics remains largely unknown. In particular, it is unclear whether they reliably estimate both mean *and* covariance parameters for a latent spatial process on a 2D continuous surface. Composite likelihoods beyond pairwise difference and univariate marginal have also not been explored for ISIW. Finally, the robustness of ISIW and the SLP to misspecification has not been well-studied. In the sections below, we provide an expanded evaluation of ISIW methods, comparing the performance of MLE, SLP, and ISIW applied to the Vecchia and pairwise marginal CLs across multiple random fields and PS designs.

Our work proceeds as follows. We start off in Section 4.2 by providing background on model-based geostatistics and introduce the key methods we use as a basis for ISIW. In Section 4.3, we present the implementation of ISIW in detail. In Section 4.4, we conduct a comparison analysis of the MLE, SLP, and ISIW through a set of simulation studies. In Section 4.5 we apply the same methods to the famous Galicia moss dataset which has been the dataset of choice when investigating PS. In Section 4.6 we finish with final remarks and discuss the implications of our work, and outline future directions for continuing investigation.

4.2 Model-based geostatistics

4.2.1 Estimation

Maximum likelihood estimation

Under non-preferential sampling (NPS), we define a model for the Gaussian observations $\mathbf{Y}$ and underlying Gaussian process S following the SLP framework in (4.1), but we drop the point process likelihood for $\mathbf{X}$. The standard geostatistical model under NPS follows

$$[\mathbf{Y}|\mathbf{X}, S] \sim N_n(\mu\mathbf{1}_n + S(\mathbf{X}), \tau^2 I_n),$$
$$S \sim GP(0, C_{\boldsymbol{\theta}}(\cdot,\cdot)).$$

We further assume the covariance function C_θ follows a stationary isotropic Matérn covariance defined as

$$C_\theta(h) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{h}{\phi}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{h}{\phi}\right),$$

where $\theta := (\sigma^2, \nu, \phi)$ defines the covariance parameter vector containing the variance, smoothness, and range parameters respectively, h denotes the Euclidean distance between two points of observation, and K_ν is the modified Bessel function of the second kind. By convention, the smoothness parameter ν is assumed to be known and fixed *a priori* before estimation. Therefore, the parameters of interest are $\boldsymbol{\psi} := (\mu, \sigma^2, \phi, \tau^2)$.

Because any finite collection of random variables corresponding to point observations of S follows a multivariate Gaussian distribution and $\mathbf{X}$ is treated as fixed, the observed data likelihood is then defined by

$$[\mathbf{Y}, \mathbf{X}] \propto N_n(\mu\mathbf{1}_n, \Sigma_n(\theta) + \tau^2 I_n), \tag{4.2}$$

where Σ_n is the covariance matrix with ijth entry equal to $C_\theta(\|\mathbf{x}_i - \mathbf{x}_j\|)$. Estimation proceeds by optimization of the likelihood with respect to the parameter vector $\boldsymbol{\psi}$.

Justification for geostatistical inference via the MLE derives from spatial large sample theory. Two main frameworks have dominated asymptotics for geostatistical estimators: increasing domain asymptotics, and fixed domain (or infill) asymptotics. *Increasing domain asymptotics* assume that as $n \to \infty$, the study region $\mathcal{D}$ expands, ensuring a minimum separation distance between observation locations and a fixed density of observations per unit area. *Fixed domain asymptotics*, on the other hand, keep $\mathcal{D}$ fixed while increasing n, leading to a growing observation density and vanishing minimum distance between observations (i.e., observations occur closer together as the sample size increases within the fixed study area).

Under increasing domain asymptotics, the MLE for $\boldsymbol{\psi}$ is consistent and asymptotically normal (AN) (Mardia and Marshall, 1984; Bachoc, 2014, 2020). Results are more challenging under fixed domain asymptotics due to the inclusion of increasingly close observations of a spatially correlated process. Specifically, for a Matérn Gaussian process with known ν, only a subcomponent of the covariance parameters known as the *microergodic* parameter is consistently estimable and asymptotically normal (Zhang, 2004; Kaufman and Shaby, 2013). We refer to this parameter as κ, given by $\sigma^2/\phi^{2\nu}$ for Matérn covariance functions. The nugget τ^2 is also estimable (Tang et al., 2021b) while the variance σ^2 and range ϕ are not (Zhang, 2004). Although the smoothness parameter ν is theoretically estimable, it is numerically challenging to estimate and is conventionally treated as fixed (Loh et al., 2021). Fixed effect coefficients in the mean are typically non-estimable (Wang et al., 2020), except under specific smoothness conditions on the covariates (Yu, 2022; Bolin and Wallin, 2024; Gilbert et al., 2024).

The main drawback for practical use of the MLE is its computational burden. Evaluating the likelihood requires an inversion of the covariance matrix, which has

$O(n^3)$ time and $O(n^2)$ space complexity making it infeasible for applications for moderately sized n. Recent advancement in modern spatial statistics has focused on GP parameter estimation using approximations which massively reduce computational burden in exchange for marginal decreases in efficiency. We explore composite likelihood and the Vecchia approximation, two approaches which have been widely adopted in geostatistics and are particularly conducive to weighting.

Composite likelihood

Sticking with our previous notation, let $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^\top \in \mathbb{R}^n$ be a $n \times 1$ random vector with probability density $f(\mathbf{y}; \boldsymbol{\psi})$ for unknown r-dimensional parameter vector $\boldsymbol{\psi} \in \Psi \subseteq \mathbb{R}^r$. Define $\{\mathcal{A}_1, ..., \mathcal{A}_K\}$ to be a set of K marginal or conditional events with associated likelihood $\mathcal{L}_k(\theta; \mathbf{y}) \propto f(\mathbf{y} \in \mathcal{A}_k; \boldsymbol{\psi})$. The log composite likelihood (CL) is the weighted sum of the event-specific log likelihoods

$$\log \mathcal{L}_C(\boldsymbol{\psi}; \mathbf{y}) := \sum_{k=1}^{K} w_k \log \mathcal{L}_k(\boldsymbol{\psi}; \mathbf{y}),$$

where $\mathbf{w} := (w_1, ..., w_K)$ is a vector of weights, not necessarily non-negative and the density f follows that defined in (4.2). The maximum CL estimate is defined as $\hat{\boldsymbol{\psi}}_C := \text{argmax}_{\boldsymbol{\psi}} \log \mathcal{L}_C(\boldsymbol{\psi}; \mathbf{y})$. One common choice of CL is the univariate marginal where the log likelihood is the sum of the log marginal density of each observation.

$$\log \mathcal{L}_{UM}(\boldsymbol{\psi}; \mathbf{y}) = \sum_{i=1}^{n} w_i \log f(y_i; \boldsymbol{\psi}).$$

Typically, the use of this univariate marginal in geostatistics is limited because it ignores dependencies between observations and thus cannot estimate covariance parameters. For this reason, pairwise CLs have been much more popular (Varin et al., 2011). Bevilacqua and Gaetan (2015) compared the efficiency of three different pairwise CL estimators: pairwise marginal (PMLE), pairwise conditional (PCMLE),

and pairwise difference (PDMLE) CLs. These authors concluded that the PMLE outperformed all other pairwise CLs and recommended its use over the PCMLE and PDMLE. The likelihoods for each pairwise CL are

$$\begin{aligned} \log \mathcal{L}_{PM}(\boldsymbol{\psi}; \mathbf{y}) &= \sum_{i<j} w_{ij} \log f(y_i, y_j; \boldsymbol{\psi}), \\ \log \mathcal{L}_{PC}(\boldsymbol{\psi}; \mathbf{y}) &= \sum_{i \neq j} w_{ij} \log f(y_i | y_j; \boldsymbol{\psi}), \\ \log \mathcal{L}_{PD}(\boldsymbol{\psi}; \mathbf{y}) &= \sum_{i<j} w_{ij} \log f(y_i - y_j; \boldsymbol{\psi}). \end{aligned} \tag{4.3}$$

These authors also proved, under increasing domain asymptotics, the consistency and asymptotic normality of $\hat{\boldsymbol{\psi}}_C$ estimated from (4.3). Bachoc et al. (2019) proved the same properties for the PMLE and PCMLE in the one-dimensional setting under fixed domain asymptotics. However, the true utility of CL lies in its computational speed and robustness to misspecification. While the MLE requires a $O(n^3)$ matrix inversion and specification of the joint density, pairwise CL consists only of $O(n^2)$ terms and only requires correct specification of second order densities. The computational efficiency of pairwise CL can be further enhanced by using the weights $\mathbf{w}$ to only include pairwise observations within a certain distance d apart (Bevilacqua and Gaetan, 2015).

In later sections, we will use the PMLE as a basis for ISIW due to its superior performance over other pairwise likelihoods. In particular, we do not consider the univariate marginal or PDMLE because they do not directly estimate the entirety of $\boldsymbol{\psi}$. In the next section, we discuss the Vecchia approximation, an alternative likelihood approach that achieves even greater computational and statistical efficiency than the CL.

Vecchia approximation

The Vecchia approximation is a specific case of CL based on the observation that the joint distribution can be decomposed as the product of conditional distributions. Let $\mathbf{Y} = (Y_1, ..., Y_n)^\top \in \mathbb{R}^n$ and $p : \{1, ..., n\} \to \{1, ..., n\}$ be a permutation mapping which reorders $\mathbf{Y}$. We define the history of variable Y_j as a random subvector $\mathbf{Y}_{h(j)}$ where $h(j) = \{l \in \{1, ..., n\} : p(l) < p(j)\}$. The joint density of $\mathbf{Y}$ can then be refactored as

$$f(\mathbf{y}; \boldsymbol{\psi}) = f(y_{p(1)}; \boldsymbol{\psi}) \prod_{i=2}^{n} f(y_{p(i)} | \mathbf{y}_{h(i)}; \boldsymbol{\psi}).$$

Vecchia observed that much of the information in the conditioning sets with higher indices was likely to be redundant. One could attain a good tradeoff of efficiency for computational gain by decreasing the size of each conditioning set and being judicious about which variables to include. The density would then be approximated as

$$f(\mathbf{y}; \boldsymbol{\psi}) \approx f_V(\mathbf{y}; \boldsymbol{\psi}) = f(y_{p(1)}; \boldsymbol{\psi}) \prod_{i=2}^{n} f(y_{p(i)} | \mathbf{y}_{q(i)}; \boldsymbol{\psi}), \tag{4.4}$$

where $q(i) \subseteq \{p(1), p(2), ..., p(i-1)\}$ is the set of indices constituting the conditioning set of $y_{p(i)}$. The Vecchia estimate $\hat{\boldsymbol{\psi}}_V$ maximizes (4.4). The Vecchia approximation can also be written in a weighted CL form,

$$\begin{aligned} \log \mathcal{L}_V(\boldsymbol{\psi}; \mathbf{y}) = w_1 \log f(y_{p(1)}; \boldsymbol{\psi}) \\ + \sum_{i=2}^{n} w_{1i} f(y_{p(i)}, \mathbf{y}_{q(i)}; \boldsymbol{\psi}) \\ - \sum_{i=2}^{n} w_{2i} \log f(\mathbf{y}_{q(i)}; \boldsymbol{\psi}). \end{aligned} \tag{4.5}$$

While simple in concept, the Vecchia approximation requires careful selection of three key hyperparameters: 1) the size of the conditioning set, 2) which variables to include in each conditioning set, and 3) the ordering of the variables as determined by

p. We denote the maximum size of any $q(i)$ as m. In all later sections, we choose $m = 20$ based on a clear case of diminishing returns in inference and prediction for $m > 20$ observed empirically in Datta et al. (2016). Default settings in two implementations of the Vecchia approximation (`GPvecchia` and `GpGp` `R` packages) are $m = 20$ and $m = 30$, respectively (Katzfuss and Guinness, 2021; Guinness, 2021). For the choice of which variables to include in the neighborhood $q(i)$, we follow the recommendation given in Vecchia (1988) by choosing the nearest neighbors to $y_{p(i)}$ measured by Euclidean distance. Ordering for the Vecchia approximation requires additional specification in multidimensional settings due to the lack of natural ordering in observations of a spatial point process. We use the maxmin ordering scheme, which picks a location $p(i)$ sequentially by maximizing the distance to the nearest point in $\{y_{p(1)}, y_{p(2)}, ..., y_{p(i-1)}\}$ and has been shown to approximate the true distribution better than other ordering methods (Guinness, 2018; Katzfuss and Guinness, 2021).

Theoretical justification of the Vecchia approximation in spatial statistics is relatively light compared to that of the MLE and CL. Typical approaches to proving fixed domain asymptotics fail due to the loss of stationarity in the parent process introduced by the Vecchia approximation (Zhang et al., 2024). However, it is impossible to deny its excellent empirical performance in applications (Heaton et al., 2019). A key advantage the Vecchia approximation possesses over typical CL is that (4.4) and (4.5) correspond to a valid joint probability distribution for $\mathbf{Y}$. As a result, the Kullback-Leibler (KL) divergence of the Vecchia approximation with respect to the true distribution can be computed and has been shown to be a nonincreasing function of m when the conditioning sets $q(i)$ are chosen as the nearest neighbors (Katzfuss et al., 2020; Huser et al., 2023).

The Vecchia approximation achieves a good balance of computational complexity and efficiency. It clearly outperforms other CL methods in terms of what order correlations are being accounted for, scalability, and probabilistic validity. This is

the main reason for considering its use in ISIW methods compared to CL, a point we expand on in later sections.

4.2.2 Prediction

For many spatial analyses, the main goal is to predict values at unobserved locations. We focus on spatial prediction via *kriging*, or Gaussian process regression, a function approximation technique with point and variance estimate defined as

$$\begin{aligned} \hat{S}(\mathbf{X}_0) &= \mu + C_\theta(\mathbf{X}_0, \mathbf{X}_n)^\top \boldsymbol{\Sigma}_n(\theta)^{-1}(\mathbf{Y} - \mu \mathbf{1}_n), \\ \mathrm{Var}(\hat{S}(\mathbf{X}_0)) &= C_\theta(\mathbf{X}_0, \mathbf{X}_0) - C_\theta(\mathbf{X}_0, \mathbf{X}_n)^\top \boldsymbol{\Sigma}_n^{-1}(\theta) C_\theta(\mathbf{X}_0, \mathbf{X}_n), \end{aligned} \tag{4.6}$$

where $\mathbf{X}_0$ represents unobserved locations to be predicted. In practice, since the parameters are unknown, all parameters in (4.6) are replaced with estimates obtained from any of the aforementioned or other estimation methods.

Kriging has several properties that make it an effective predictor of spatial surfaces. As the best linear unbiased predictor (BLUP), it minimizes mean squared prediction error among all unbiased predictors (Cressie and Wikle, 2015). However, it is important to note that this optimality no longer holds under preferential sampling, where even the true values of $\boldsymbol{\psi}$ used in the kriging equations yield suboptimal predictions (Dinsdale and Salibian-Barrera, 2019a). Stein established the asymptotic efficiency of kriging under various robustness conditions (Stein, 1988, 1990a,b, 1993), while Wang et al. (2020) showed that kriging's prediction error vanishes under a uniform metric. Additionally, Putter and Young (2001) demonstrated that the difference between predictions using kriging with estimated and the true parameters is asymptotically negligible if the joint Gaussian distributions of the spatial process under the true and estimated covariance functions are contiguous almost surely. Given these results, one could argue that the geostatistical model in (4.2) is more appropriate for spatial interpolation rather than inference.

Table 4.1: Point process intensity function estimators considered in the simulation study.

Method	Reference	Bandwidth Selection
`diggle`	Diggle (1985)	Least-squares cross-validation
`scott`	Scott (1992)	Rule-of-thumb based on normal reference density
`ppl`	Loader (1999)	Likelihood cross-validation (leave-one-out)
`CvL`	Cronie and Van Lieshout (2018)	Maximum likelihood cross-validation
`CvL.adaptive`	van Lieshout (2021)	Adaptive bandwidth based on local point density
`R-INLA`	Simpson et al. (2016)	LGCP model (non-kernel-based)

4.3 Inverse sampling intensity weighting

We next describe estimation and prediction procedures of ISIW for PS adjustment. ISIW is a two-stage approach, wherein we first estimate the sampling intensity at each of the observation locations $\mathbf{X}$, and, in the second stage, we input the (estimated) inverse sampling intensities as weights into a weighted likelihood adjustment. The resulting adjusted parameter estimates are then substituted in the kriging equations for prediction.

4.3.1 Estimation of sampling intensity

The essence of ISIW is to account for the dependence between the response and observation locations by using the vector of estimated sampling intensities instead of the more complex full likelihood of the observation process. This approach assumes that the locations contain sufficient information about the spatial dependence structure of the observation locations to enable proper adjustment.

Methods to estimate a spatially varying intensity of a point process can be divided into parametric and nonparametric approaches. Domain knowledge can inform the parametric form of the intensity either through choice of model or covariates, but oftentimes this information is unavailable and nonparametric estimation is required. The nonparametric approaches follow the kernel smoothing approach discussed in

Diggle (1985). Let k be a d-dimensional kernel function from $\mathbb{R}^d \to \mathbb{R}^+$, which is a symmetric probability density function. Given a bandwidth size $h > 0$ and edge correction factor $w_h(\cdot, \cdot)$, the nonparametric kernel smoothing estimator of the intensity function is given by

$$\hat{\lambda}(\mathbf{x}; h) = h^{-d} \sum_{\mathbf{y} \in \mathbf{X} \cap \mathcal{D}} k\left(\frac{\|\mathbf{x} - \mathbf{s}\|}{h}\right) w_h(\mathbf{x}, \mathbf{s})^{-1}, \mathbf{x} \in \mathcal{D}.$$

The key hyperparameter for nonparametric intensity estimation is the bandwidth size h. Bandwidth selection is a well-studied problem, with methods ranging from high bandwidth, smooth intensity estimators to low bandwidth, flexible intensity estimators. Each has its place in the bias-variance tradeoff, with higher bandwidths exhibiting higher bias with lower variance and lower bandwidths vice versa. We experiment with several bandwidth selection strategies implemented in the `spatstat` `R` package (Table 4.1).

The effectiveness of any ISIW approach for PS should improve the closer estimated weights $\lambda(\mathbf{x})^{-1}$ are to the true inverse intensities. In theory, a parametric LGCP model should estimate the true intensity better than any nonparametric estimator if $\mathbf{X}$ follows a LGCP, especially with multiple realizations. However, it is known that it is impossible to distinguish *between* a single realization from an inhomogenous Poisson process with deterministic intensity $\lambda(\mathbf{x})$ and *one from* a stationary Cox process whose conditional intensity coincides with λ (Gelfand et al., 2010). We consider a LGCP method introduced in Simpson et al. (2016) estimated by `R-INLA` to verify whether a known parametric model shows improvement over a nonparametric estimator.

Once the sampling intensities are estimated at every observation location, we calculate weights as the inverse of the sampling intensities and then normalize these to sum up to the number of observations n. The estimated weight for the ith observation

becomes

$$\hat{w}_i = \hat{w}(\mathbf{x}_i) = n * \frac{\hat{\lambda}(\mathbf{x}_i)^{-1}}{\sum_{\mathbf{s}\in\mathbf{X}} \hat{\lambda}(\mathbf{s})^{-1}}. \tag{4.7}$$

4.3.2 Defining the likelihood

ISIW applies weights proportional to the inverse sampling intensity to each event in a likelihood factored as a product of densities. The only known examples of ISIW in the PS literature include a pairwise difference (Schliep et al., 2023) and univariate marginal CL (Vedensky et al., 2023). The pairwise difference CL cannot estimate μ directly, limiting ISIW's impact on PS adjustment of the mean, while the univariate marginal CL cannot estimate covariance parameters θ. To address these gaps, we propose two new ISIW estimators for simultaneous mean and covariance estimation: the ISIW pairwise marginal (ISIW-PM) and ISIW Vecchia (ISIW-V) estimator. Let the weights be defined as in (4.7). Then the ISIW-PM follows

$$\log \mathcal{L}_{WPM}(\boldsymbol{\psi}; \mathbf{y}) = \sum_{i<j} w_i w_j \log f(y_i, y_j; \boldsymbol{\psi}).$$

For the ISIW-V, we initially considered weighting the likelihood as follows

$$\begin{aligned}\log \mathcal{L}_{WV}(\boldsymbol{\psi}; \mathbf{y}) = {} & w_{p(1)} \log f(y_{p(1)}; \boldsymbol{\psi}) \\ & + \sum_{i=2}^{n} \left(\prod_{j\in\{p(i)\}\cup q(i)} w_j \right) \log f(y_{p(i)}, \mathbf{y}_{q(i)}; \boldsymbol{\psi}) \\ & - \sum_{i=2}^{n} \left(\prod_{j\in q(i)} w_j \right) \log f(\mathbf{y}_{q(i)}; \boldsymbol{\psi}),\end{aligned}$$

to maintain the probabilistic interpretation of the inverse weighting. However, numerical issues arise due to taking the product of several small weights, greatly increasing the chance of extreme values. As a more computationally efficient approach, we

approximate the true weight using the following weighted Vecchia approximation

$$\begin{aligned}\log \mathcal{L}_{WV}(\boldsymbol{\psi}; \mathbf{y}) = & \, w_{p(1)} \log f(y_{p(1)}; \boldsymbol{\psi}) \\ & + \sum_{i=2}^{n} w_{p(i)} \log f(y_{p(i)}, \mathbf{y}_{q(i)}; \boldsymbol{\psi}) \\ & - \sum_{i=2}^{n} w_{p(i)} \log f(\mathbf{y}_{q(i)}; \boldsymbol{\psi}).\end{aligned}$$

4.3.3 Numerical estimation

The ISIW for both the Vecchia and composite likelihood can be estimated by standard optimization procedures. In our study, we used the L-BFGS-B routine as implemented in the `optim` package in the `R` language. Initial values for parameters to be estimated were selected with general rules of thumb from the `GPVecchia` R package and all other optimization parameters were set to their default settings.

4.3.4 Winsorization of extreme weights

As with other inverse weighting procedures, a key concern for ISIW is extreme weights. Observation locations in sparsely sampled areas will have extremely high weights in the final likelihood and can cause numerical issues in the optimization. Two straightforward options are *trimming* and *Winsorization*. Trimming involves eliminating values considered as outliers, while Winsorizing enforces outliers to a threshold value. We chose Winsorization to handle extreme values because trimming observations can remove rare information not captured by less extreme weights, weakening the effectiveness of PS adjustment. For our Winsorization procedure, we first normalize the sampling intensities $\hat{\lambda}$ to sum to n. We then Winsorized the normalized intensities by choosing a lower threshold (to be elaborated on later), and normalized the new resultant weights to sum to n.

4.3.5 Prediction

ISIW prediction of unobserved location follows by plugging in the estimated parameters $\hat{\boldsymbol{\psi}}$ into the kriging equations in (4.6). This is a key distinction between ISIW and the SLP framework. Whereas prediction by ISIW adheres to kriging by substituting PS-adjusted parameters, the SLP model computes predictions from the estimated distribution of $[S|\mathbf{X}, \mathbf{Y}]$, making it much more computationally intensive.

4.4 Simulation analysis

4.4.1 Experiment

We conducted a simulation experiment to evaluate the inferential and predictive performance of ISIW-PM and ISIW-V versus the benchmark MLE and SLP approaches under preferential sampling. We generated $B = 500$ realizations for two separate Matérn random fields on a 200×200 grid on the unit square: a low-range ($\phi = 0.02$) and high-range ($\phi = 0.15$) surface with $\mu = 4, \sigma^2 = 1.5, \nu = 1$, and $\tau^2 = 0.1$. To test robustness of methods against a variety of observation patterns, we sampled three different point processes conditional on having $n = 100$ and $n = 800$ observations: a log Gaussian Cox process (LGCP), a sigmoidal Cox process (SCP), and a Thomas process. Traditionally, the SLP approach has only been tested against LGCP observation patterns where it is correctly specified.

The intensity function for the LGCP is $\exp\{\beta S(\mathbf{x})\}$. For the sigmoidal Cox process, the intensity function follows $\beta(1 + \exp\{-S(\mathbf{x})\})^{-1}$. For the Thomas process, a set of parent points was first generated from a homogeneous Poisson process. The number of offspring for each parent then followed a Poisson distribution with mean $\exp\{\beta S(\mathbf{x})\}$ and offspring points were sampled from a normal distribution centered at their respective parent locations, with a scale parameter of 0.1. We set $\beta = 1$ for all

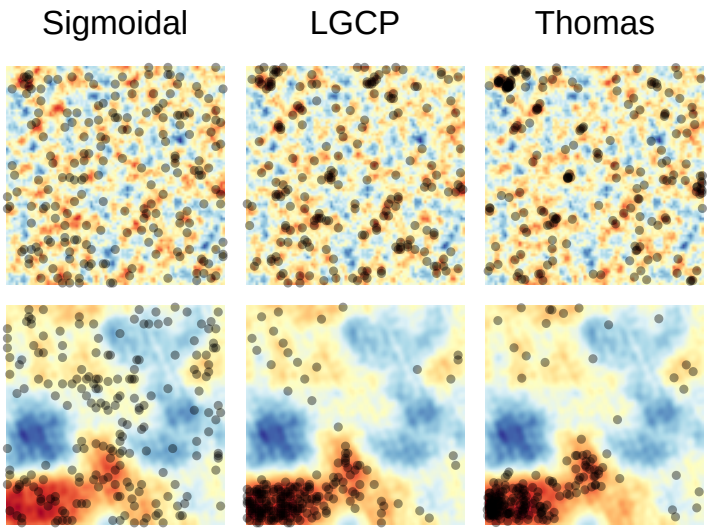

Figure 4.1: One realization of the point processes used in the simulation analysis for $n = 200$. The top and bottom row fields are low range ($\phi = 0.02$) and high range ($\phi = 0.15$) respectively.

point processes. The SCP, LGCP, and Thomas process were deliberately selected to represent visually similar clustered point processes, ordered by increasing clustering intensity, with more distinct and separated clusters in the Thomas process (Figure 4.1).

For each set of observations, we fit the MLE, SLP, ISIW-PM, and ISIW-V models and compared the results in terms of parameter estimation and prediction. The MLE, ISIW-PM, and ISIW-V were estimated using `optim` as described in Section 4.3.3. We estimated the SLP model parameters using the INLA-SPDE implementation as defined in Watson et al. (2019) and used penalized complexity (PC) priors on the variance and range parameters (Simpson et al., 2017; Fuglstad et al., 2019). For the range parameter we set $P(\phi < 0.01) = 0.01$ and for the variance parameter we set $P(\sigma > 10) = 0.01$. The default priors for INLA were used for μ and τ, set as $N(0, 1000)$ and Gamma$(1, 5 \times 10^{-5})$, respectively. Prediction for the MLE and all ISIW methods followed the kriging equations, with point estimates of the requisite parameters substituted into (4.6).

We evaluated several variants of the ISIW methods, distinguished by their first-stage sampling intensity weight estimation (Table 4.1). As a gold standard for assessing the optimal performance of ISIW, we also used the true weights extracted from the simulation surface, referred to as "Known" weights. The intensity estimators included multiple nonparametric kernel density estimators from `spatstat` and a parametric LGCP model, which used the same priors and mesh as the SLP. To determine a threshold for Winsorization, we conducted a small numerical experiment, simulating a Matérn random field with high range ($\phi = 0.15$) 100 times. For each replicate, we generated two LGCP point patterns with $\beta = 1$, one with $N = 100$ and the other with $N = 800$. ISIW-V was then fit under different weight thresholds using the `bw.diggle` bandwidth selection, as it was the most prone to numerical instability. The best-performing threshold, 1×10^{-2}, was selected based on yielding the best average root mean-squared prediction error (RMSPE). All subsequent ISIW methods implemented Winsorization using the 1×10^{-2} threshold.

Our approach to evaluating each method prioritized predictive performance due to the aforementioned challenges with standard statistical inference in fixed domain asymptotics and preferential sampling. The RMSPE was computed over each center point in a 200×200 grid, yielding 40,000 prediction points per method and simulation scenario. Parameter estimation was then assessed by computing relative bias and relative root mean-squared error (RMSE) for the parameter vector $(\mu, \sigma^2, \phi, \tau^2, \kappa)$.

4.4.2 Results

Across all simulation scenarios, INLA-SLP and ISIW-V Known were the top two predictors by RMSPE, with similar performance to each other and a clear margin over other methods. In general, INLA-SLP had better predictive performance in the high range scenario, whereas ISIW-V Known did so in the low range setting (Tables 4.4 and 4.5). The one exception was the high range scenario for the $N = 100$

Table 4.2: Predictive performance of all sixteen methods based on median rank, mean rank, and percentage of total simulations when RMSPE for the method was lower than that of MLE and SLP. Rank was determined by RMSPE.

Method	Median Rank of RMSPE	Mean Rank of RMSPE	% RMSPE lower than MLE	% lower RMSPE than SLP
INLA-SLP	2	5.29	73.3	NA
ISIW-V Known	2	5.75	70.0	44.0
ISIW-V CvL.adaptive	5	6.53	72.7	31.8
ISIW-V diggle	5	5.96	78.5	29.3
ISIW-V CvL	7	7.38	85.0	28.9
ISIW-V ppl	7	7.37	74.4	27.8
ISIW-V INLA	8	8.23	74.0	27.9
ISIW-PM CvL.adaptive	8	8.01	65.1	28.3
ISIW-PM diggle	8	7.98	61.3	28.1
ISIW-V scott	9	8.29	85.6	28.4
ISIW-PM Known	9	8.60	51.8	23.9
ISIW-PM ppl	10	9.84	50.4	27.9
MLE	11	10.2	NA	26.7
ISIW-PM CvL	12	11.3	30.9	26.5
ISIW-PM INLA	14	12.3	27.0	24.9
ISIW-PM scott	14	12.7	22.3	24.6

Thomas process, where ISIW-V Known had both a lower RMSPE (0.694) and a lower standard deviation (0.15) compared to INLA-SLP (0.749 and 0.27) (Table 4.6). However, there were instances of predictive instability in INLA-SLP, as indicated by the high standard deviation of RMSPE, particularly in the $N = 100$ low range SCP and high range Thomas process $N = 100$ scenarios where INLA-SLP was misspecified. The ISIW methods also tended to run faster than the INLA-SLP at both sample sizes and had similar runtimes to the MLE (Table 4.7). We did not experiment with larger sample sizes than $N = 800$ because PS is primarily a concern in small to moderate sample sizes; as sample sizes grow sufficiently large, its effects diminish.

While ISIW-V using estimated weights did not match the predictive performance of ISIW-V Known, weight estimation variants including `diggle` and `CvL.adaptive` showed improved predictive performance over the alternative of no PS adjustment (Table 4.2). They achieved a median rank of 5 out of all 16 methods, had a lower

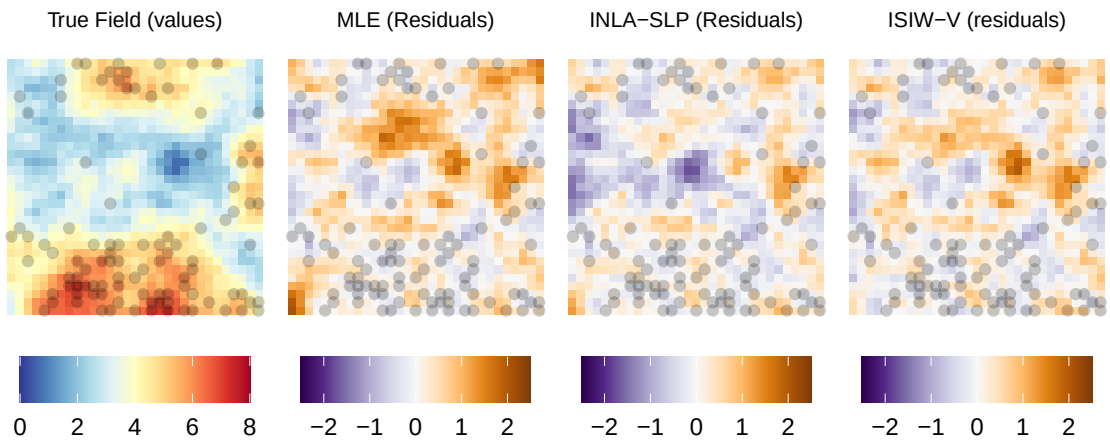

Figure 4.2: Example predictive surface for the MLE, INLA-SLP, and ISIW-V approaches. The first panel shows the true field values with observations denoted as points, while the remaining panels display residuals (observed minus predicted values) for each method.

RMSPE than MLE in over 70% of simulations, and even outperformed SLP nearly 30% of the time (for comparison, ISIW-V Known outperformed SLP in 44% of simulations). The `CvL` variant outperformed MLE in 85% of simulations, surpassing even SLP, which did so in 73%. Surprisingly, ISIW methods using weights estimated by the parametric LGCP model fitted in INLA performed relatively poorly, with a median rank of 8, though they still consistently outperformed MLE. All ISIW-V variants outperformed MLE in over 70% of simulations, whereas ISIW-PM lagged far behind, with some variants on average ranking worse than the MLE.

The variants that predicted best tended to estimate lower bandwidths which introduced greater noise in prediction, as reflected in the higher standard deviation of RMSPE for these variants. In contrast, `scott` and `ppl` traded variance for bias, which increased numerical stability but decreased the impact of weighting, making them nearly indistinguishable from prediction based on the MLE.

Figure 4.2 provides an illustrative example comparing the predictive tendencies of MLE, INLA-SLP, and ISIW-V Known. In the data-rich area at the bottom, predictions from all three methods are nearly identical. The differences emerge in data-

Table 4.3: Relative bias and RMSE in parameter estimation for MLE, SLP, and ISIW methods across all simulation scenarios. Bolded values indicate the method with the smallest bias or RMSE for a given parameter.

Method	Variant	μ		σ^2		ϕ		τ^2		κ	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
MLE	-	0.106	0.173	0.046	0.589	**0.013**	0.429	-0.057	**0.897**	**0.113**	**0.538**
INLA-SLP	-	**0.016**	**0.116**	0.241	1.32	0.110	0.434	2.13×10^2	1.19×10^4	0.546	17.1
ISIW-V	Known	-0.259	0.409	0.035	20.8	90.9	4.52×10^3	0.053	1.72	0.712	1.58
	CvL	0.022	0.240	-0.057	1.10	65.1	2.71×10^3	**-0.007**	1.14	0.140	0.624
	CvL.adaptive	-0.018	0.268	0.362	51.9	1.25×10^3	3.08×10^4	0.441	2.44	2.44	1.14
	INLA	0.059	0.170	-0.068	0.807	-0.023	0.673	-0.013	1.15	0.166	0.804
ISIW-PM	Known	-0.161	0.309	-0.293	**0.408**	-0.352	0.489	0.713	1.09	1.900	7.65
	CvL	0.122	0.203	0.061	0.596	-0.168	**0.397**	0.871	1.50	0.885	2.05
	CvL.adaptive	0.042	0.150	-0.162	0.511	-0.184	0.630	0.782	1.35	1.550	11.0
	INLA	0.135	0.211	**-0.018**	0.570	-0.201	0.414	0.859	1.56	0.912	1.69

sparse regions where PS is known to have its greatest impact. While the MLE over-predicts the center of the grid, the INLA-SLP predicts a much lower value over the same region which decreases predictive error but leads to underestimation in a specific area indicated by the deep purple. ISIW-V, similar to INLA-SLP, lowers predictions in that region but not as drastically, avoiding the central underprediction seen in INLA-SLP but still overestimating in another area, shown as dark orange slightly to the right of the center. As shown by this example, both INLA-SLP and ISIW-V adjust for PS in similar ways, but INLA-SLP tends to apply a larger adjustment in data sparse areas compared to ISIW-V.

Table 4.3 compares the relative bias and relative RMSE in parameter estimation for the MLE, SLP, and ISIW methods, with a specific focus on the Known, `CvL`, `CvL.adaptive`, and `INLA` variants. While the MLE approach overestimated μ and INLA-SLP estimated μ accurately, ISIW-V Known significantly underestimated the mean by over 25% (averaging an estimate of three for a true mean of four), with a highly variable RMSE of 0.409. The estimated weight variants of ISIW-V had higher error in mean estimation compared to INLA-SLP but lower error than ISIW-V Known and MLE.

For the range parameter, ISIW-V estimates were consistently too high across all variants and exhibited potential numerical instability. In contrast, ISIW-PM underestimated the range but did so with much greater stability. MLE had relatively low estimation error for all covariance parameters, particularly κ and τ^2, aligning with theoretical expectations. Aside from the mean, INLA-SLP did not estimate any parameter particularly well, with τ^2 showing significant numerical instability.

There was no clear relationship in the simulation between the method of parameter estimation and resulting predictive performance. INLA-SLP estimated the nugget and microergodic parameter extremely poorly, despite these being theoretically among the easiest to estimate, while ISIW-V on average underestimated the mean by 25% and vastly overestimated the range by more than a factor of 90. Yet, both methods led in predictive performance. In contrast, MLE and the `CvL` and `INLA` variants of ISIW produced relatively good parameter estimates but consistently ranked lower in predictive performance, with prediction errors far exceeding those of INLA-SLP and ISIW-V.

4.5 Application to the Galicia moss data

The Galicia moss dataset has been a widely used example for illustrating the effects of preferential sampling on geostatistical inference and prediction (Diggle et al., 2010; Dinsdale and Salibian-Barrera, 2019a). It contains lead concentrations in moss samples, measured in micrograms per gram of dry weight, collected from Galicia, northern Spain. Figure 4.3 shows the spatial distribution of 63 measurements taken in 1997 and 132 measurements taken in 2000. Sampling in 1997 was preferential, with a bias toward locations in the north with lower lead concentrations, whereas sampling in 2000 was more regular and non-preferential. Further details on the dataset can be found in Diggle et al. (2010).

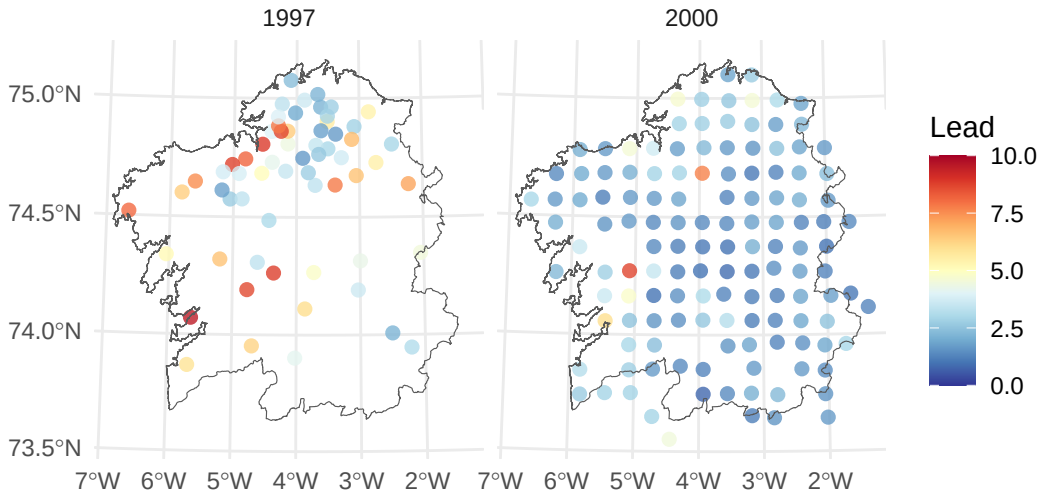

Figure 4.3: Observed lead concentrations at sampled locations in Galicia in 1997 and 2000.

We fit each of the MLE, INLA-SLP, ISIW-V, and ISIW-PM approaches using the 1997 and 2000 data separately and generated predictive surfaces over a 20×20 grid covering the Galicia area. Based on its good performance in the simulation analysis, we estimated the ISIW first stage intensities using `CvL.adaptive` bandwidth selection. We did encounter numerical difficulties when fitting the INLA-SPDE model to the 1997 data. While previous analyses of the Galicia dataset have used areal models, we elected to use the SPDE approach to be consistent with our simulation analysis. The model was particularly sensitive to the choice of prior for the range and initial value for the nugget. It would sometimes generate predictive surfaces with no variation, where the same value was predicted across the entire region including in locations with observations recording different values. Here we present the predictive surface that exhibited some variation and aligned with expectations from kriging, where predictions have similar values to the observations closeby.

Figure 4.4 displays the predictions for the lead concentration surface. There was little difference for the 2000 data, however, substantial differences emerged in the 1997 data. As expected, both the ISIW-V and INLA-SLP estimated higher values in

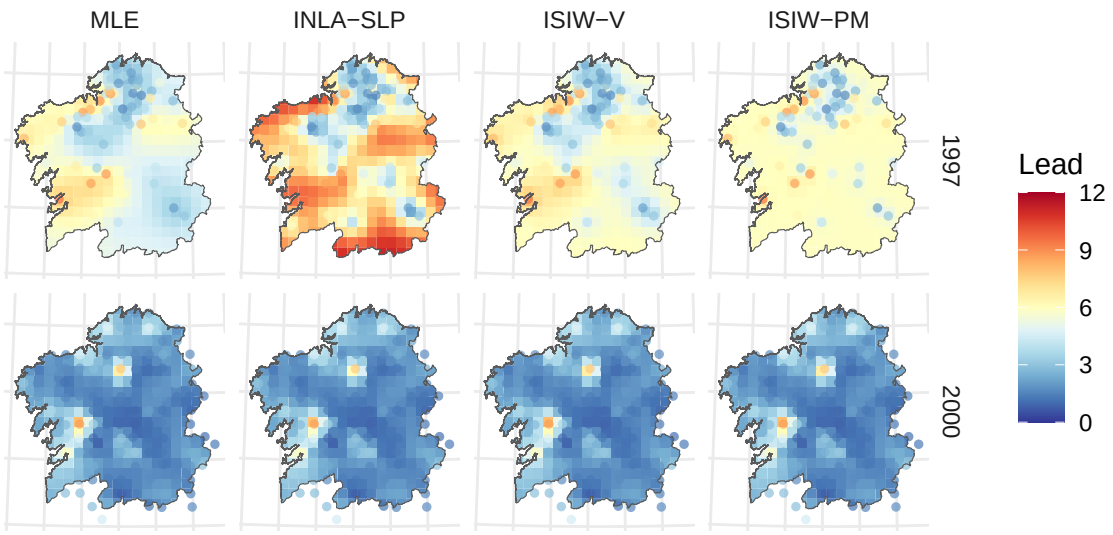

Figure 4.4: Spatial predictions of lead concentrations in Galicia in 1997 and 2000 using MLE, INLA-SLP, ISIW-V, and ISIW-PM. Points represent the observed data.

data sparse areas to correct for the preferential sampling of low values. The INLA-SLP applied a more extreme correction than the ISIW-V, similar to the example in Figure 4.2. It unexpectedly predicted values exceeding 10.5 in the south, where the nearest observations fall below 5.00. It also estimated a very high μ of 10.79 (Table 4.8). This was unusual considering the maximum observation was 9.51 and the mean of all 63 observations in 1997 was 4.72. The MLE, ISIW-V, and ISIW-PM only estimated a μ of 5.15, 6.07, and 5.95, respectively. Besides being a result of INLA-SLP's PS adjustment, the discrepancy could also be attributed to numerical instability in INLA-SLP estimation, and other priors or initial values might produce different results.

Similar to our simulation results, the ISIW-PM estimated a lower range compared to other methods. This low range likely explains why its predictive surface exhibits little spatial dependence with nearly uniform predictions across the region. ISIW-V appears to strike a balance between implementing a moderate adjustment for PS while preserving the spatial correlation evident in the dataset.

4.6 Discussion

In this work, we compared ISIW methods to the SLP and MLE for parameter estimation and prediction under preferential sampling. We implemented ISIW in a novel combination with the Vecchia approximation and found that, despite being less well-suited for geostatistical inference, ISIW with the Vecchia approximation nevertheless improves predictive accuracy over the MLE approach and remains competitive with the INLA-SLP approach using the true weights. We conducted a comprehensive simulation study that extended beyond previous analyses, evaluating the robustness of ISIW across various settings. Its benefits persisted across multiple observation point processes, different levels of spatial dependence in the underlying field, and alternative point process intensity estimators used in the first stage of ISIW. We also showed an application to the Galicia moss data where ISIW appears to give more reasonable predictions compared to the standard SLP approach.

Several factors make ISIW an attractive alternative to SLP for PS adjustment, namely its speed, simplicity, and strong predictive performance. Not only does it scale more efficiently with increasing n, but in terms of the practical implementation, we found that our version of ISIW-V ran faster than INLA-SLP for combined estimation and prediction, even though we did not implement an optimized version. ISIW is also simple to use and understand. INLA and TMB require specification of both the prior distributions for hyperparameters and mesh triangulation that can significantly alter results (Righetto et al., 2020). In contrast, ISIW is a straightforward inverse weighting method familiar to much of the research community that only requires an estimated vector of weights. Furthermore, since most effective Gaussian process approximation techniques are already based on the Vecchia approximation (Heaton et al., 2019), integrating ISIW into state-of-the-art methods is significantly simpler and more straightforward compared to the SLP. These computational and conceptual advantages come without sacrificing predictive performance. In our simulation

analysis, ISIW achieved improved predictive accuracy over the MLE with estimated weights and performed comparably to the gold standard SLP when the true weights were known. In scenarios where the SLP was misspecified, ISIW outperformed it, demonstrating greater robustness and fewer parametric assumptions.

Many improvements can still be made to ISIW. A significant gap remains between the predictive accuracy achieved by ISIW with estimated and true weights. The lack of accuracy and theoretical guarantees for the nonparametric intensity estimators used in this study may contribute to their inability to achieve similar performance. Moreover, Winsorization is necessary to prevent pathological numerical behavior but remains an undesirable ad hoc step. Intensity modelling incorporating the values of the observations rather than just their locations may also lead to improved weight estimation. Future work should explore more flexible and robust estimators for point process intensities beyond kernel density estimators and the INLA approach.

Additionally, ISIW does not yield reliable parameter estimates for the underlying geostatistical model. The weighted likelihood adjustment introduces substantial bias and error into parameter estimation, yet unexpectedly improves predictions through kriging. The apparent disconnect between parameter estimation and prediction warrants further investigation. It remains unclear why parameters obtained from the weighted likelihood adjustment consistently lead to better kriging predictions. While kriging is known to lose its predictive optimality under preferential sampling, the results of this study provide insight into conditions under which it can still serve as an effective function approximation tool under relaxed assumptions like PS.

Finally, we did not address uncertainty quantification for parameter estimates or predictions. While the kriging variance formula remains available and could be computed using the estimated parameters, its utility in the presence of preferential sampling is uncertain. In addition, geostatistical inference is already challenging and preferential sampling adjustment appears to exacerbate the difficulty. As a result,

any standard errors for parameter estimates obtained under ISIW are unlikely to have valid coverage properties. Future work will focus on developing principled approaches to quantifying uncertainty in both inference and prediction under ISIW.

4.7 Supplementary Material

Table 4.4: Average RMSPE (standard deviation) for methods under the LGCP simulation. Bolded values indicate the method with the lowest average RMSPE across all simulations for that setting.

		N = 100		N = 800	
Method	Weights	$\phi = 0.02$	$\phi = 0.15$	$\phi = 0.02$	$\phi = 0.15$
MLE		1.726 (0.11)	0.785 (0.17)	1.306 (0.06)	0.361 (0.07)
INLA-SLP		1.195 (0.08)	**0.553 (0.09)**	0.995 (0.04)	**0.305 (0.04)**
ISIW-V	Known	**1.193 (0.11)**	0.656 (0.13)	**0.908 (0.04)**	0.346 (0.06)
	CvL	1.660 (0.12)	0.728 (0.16)	1.238 (0.06)	0.353 (0.06)
	CvL.adaptive	1.639 (0.18)	0.714 (0.20)	1.098 (0.07)	0.362 (0.07)
	INLA	1.686 (0.11)	0.783 (0.17)	1.204 (0.05)	0.355 (0.06)
	diggle	1.616 (0.18)	0.712 (0.19)	1.112 (0.14)	0.355 (0.07)
	ppl	1.664 (0.13)	0.730 (0.23)	1.168 (0.16)	0.361 (0.13)
	scott	1.689 (0.11)	0.737 (0.20)	1.279 (0.06)	0.353 (0.06)
ISIW-PM	Known	1.210 (0.10)	0.773 (0.18)	1.066 (0.06)	0.384 (0.08)
	CvL	1.736 (0.12)	0.846 (0.21)	1.390 (0.07)	0.390 (0.09)
	CvL.adaptive	1.614 (0.16)	0.766 (0.18)	1.175 (0.06)	0.395 (0.09)
	INLA	1.764 (0.12)	1.006 (0.26)	1.331 (0.06)	0.392 (0.09)
	diggle	1.593 (0.16)	0.774 (0.19)	1.136 (0.06)	0.382 (0.08)
	ppl	1.723 (0.13)	0.804 (0.18)	1.238 (0.06)	0.380 (0.08)
	scott	1.788 (0.12)	0.845 (0.19)	1.448 (0.07)	0.393 (0.09)

Table 4.5: Average RMSPE (standard deviation) for methods under the SCP simulation. Bolded values indicate the method with the lowest average RMSPE across all simulations for that setting.

		N = 100		N = 800	
Method	Weights	$\phi = 0.02$	$\phi = 0.15$	$\phi = 0.02$	$\phi = 0.15$
MLE		1.289 (0.06)	0.579 (0.10)	0.999 (0.04)	0.271 (0.03)
INLA-SLP		1.374 (0.31)	**0.521 (0.08)**	0.935 (0.03)	**0.258 (0.02)**
	Known	**1.228 (0.08)**	0.551 (0.08)	**0.889 (0.03)**	0.270 (0.03)
	CvL	1.280 (0.07)	0.566 (0.09)	0.975 (0.04)	0.268 (0.03)
	CvL.adaptive	1.292 (0.09)	0.574 (0.09)	0.941 (0.05)	0.272 (0.03)
ISIW-V	INLA	1.289 (0.07)	0.581 (0.10)	0.984 (0.04)	0.270 (0.03)
	diggle	1.286 (0.08)	0.570 (0.10)	0.976 (0.04)	0.268 (0.03)
	ppl	1.288 (0.06)	0.569 (0.09)	0.987 (0.04)	0.269 (0.03)
	scott	1.285 (0.06)	0.568 (0.09)	0.995 (0.04)	0.269 (0.03)
	Known	1.232 (0.09)	0.596 (0.11)	0.900 (0.03)	0.281 (0.03)
	CvL	1.287 (0.07)	0.612 (0.12)	1.017 (0.04)	0.281 (0.04)
	CvL.adaptive	1.287 (0.10)	0.609 (0.12)	0.964 (0.04)	0.278 (0.03)
ISIW-PM	INLA	1.298 (0.07)	0.660 (0.14)	1.020 (0.04)	0.287 (0.04)
	diggle	1.282 (0.08)	0.601 (0.11)	1.017 (0.04)	0.281 (0.04)
	ppl	1.303 (0.06)	0.630 (0.12)	1.033 (0.04)	0.283 (0.04)
	scott	1.296 (0.06)	0.628 (0.12)	1.043 (0.04)	0.285 (0.04)

Table 4.6: Average RMSPE (standard deviation) for methods under the Thomas process simulation. Bolded values indicate the method with the lowest average RMSPE across all simulations for that setting.

		N = 100		N = 800	
Method	Weights	$\phi = 0.02$	$\phi = 0.15$	$\phi = 0.02$	$\phi = 0.15$
MLE		1.684 (0.11)	0.822 (0.20)	1.324 (0.06)	0.376 (0.07)
INLA-SLP		1.214 (0.11)	0.749 (0.27)	0.983 (0.04)	**0.324 (0.07)**
	Known	**1.207 (0.11)**	**0.694 (0.15)**	**0.918 (0.06)**	0.359 (0.06)
	CvL	1.641 (0.14)	0.765 (0.18)	1.265 (0.06)	0.367 (0.07)
	CvL.adaptive	1.627 (0.20)	0.752 (0.18)	1.121 (0.09)	0.375 (0.07)
ISIW-V	INLA	1.629 (0.12)	0.812 (0.20)	1.216 (0.06)	0.369 (0.07)
	diggle	1.580 (0.18)	0.766 (0.20)	1.137 (0.14)	0.367 (0.07)
	ppl	1.682 (0.26)	0.909 (0.41)	1.278 (0.28)	0.412 (0.25)
	scott	1.644 (0.13)	0.772 (0.24)	1.294 (0.06)	0.367 (0.07)
	Known	1.217 (0.08)	0.816 (0.21)	1.067 (0.06)	0.403 (0.09)
	CvL	1.744 (0.15)	0.918 (0.28)	1.435 (0.08)	0.414 (0.11)
	CvL.adaptive	1.603 (0.17)	0.816 (0.21)	1.182 (0.07)	0.409 (0.10)
ISIW-PM	INLA	1.720 (0.12)	1.040 (0.28)	1.337 (0.06)	0.411 (0.10)
	diggle	1.558 (0.16)	0.830 (0.22)	1.168 (0.07)	0.397 (0.09)
	ppl	1.617 (0.18)	0.816 (0.21)	1.239 (0.07)	0.397 (0.09)
	scott	1.762 (0.13)	0.886 (0.22)	1.482 (0.08)	0.413 (0.10)

Table 4.7: Mean (SD) and Median (IQR) runtime in seconds over all simulations for methods under different sample sizes.

		N = 100		N = 800	
Method	Variant	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
MLE		0.80 (0.26)	0.76 (0.34)	43.8 (13.8)	41.3 (17.2)
INLA-SLP		124.0 (71.8)	100.0 (87.3)	99.8 (39.3)	95.2 (37.3)
ILIW	Known	4.80 (1.81)	4.47 (2.20)	45.8 (16.5)	43.3 (20.9)
	CvL	4.14 (1.48)	3.85 (1.89)	34.4 (10.7)	32.7 (13.6)
	CvL.adaptive	4.97 (1.75)	4.63 (2.11)	41.6 (13.7)	39.4 (17.5)
	INLA	15.9 (4.51)	15.0 (5.39)	118.0 (41.4)	107.0 (51.0)
	diggle	4.83 (1.93)	4.44 (2.24)	42.5 (17.3)	39.1 (22.1)
	ppl	4.88 (1.89)	4.49 (2.16)	39.3 (16.2)	36.0 (16.8)
	scott	4.08 (1.44)	3.79 (1.83)	34.3 (11.0)	32.6 (13.8)
PMLE	Known	0.26 (0.06)	0.25 (0.06)	16.5 (2.4)	16.2 (3.3)
	CvL	0.25 (0.05)	0.24 (0.05)	14.9 (2.1)	14.8 (2.8)
	CvL.adaptive	0.59 (0.14)	0.55 (0.15)	16.6 (2.1)	16.4 (2.9)
	INLA	9.61 (3.90)	8.40 (3.41)	97.4 (36.4)	86.3 (41.8)
	diggle	0.28 (0.06)	0.27 (0.06)	16.4 (2.3)	16.2 (3.2)
	ppl	0.65 (0.13)	0.64 (0.18)	15.7 (2.1)	15.6 (2.8)
	scott	0.25 (0.05)	0.25 (0.05)	15.0 (2.3)	14.8 (3.3)

Table 4.8: Parameter estimates for the Galicia data.

Year	Method	μ	σ^2	ϕ (km)	τ^2
1997	MLE	5.15	2.85	15.35	2.24
	INLA-SLP	10.79	18.19	17.54	0.00007
	ISIW-V	6.07	2.18	8.45	1.04
	ISIW-PM	5.95	1.24	3.99	2.08
2000	MLE	2.24	1.40	8.90	0.000
	INLA-SLP	2.27	1.36	10.2	0.00013
	ISIW-V	2.24	1.40	8.91	0.000
	ISIW-PM	2.15	1.37	9.47	0.016

Chapter 5

Accounting for spatially varying sampling effort: A case study of monarch butterflies in North America

5.1 Introduction

Monarch butterflies are among the most widely recognized insects in North America. A striking feature of the monarch butterfly is its unique multi-generational migratory pattern between overwintering sites in Central Mexico and breeding grounds in the United States and Canada; no other butterfly species exhibits such long-distance migration patterns at a regular schedule. Estimating the species distribution of the monarch is critical for its conservation. Recent evidence shows the population of monarchs has been decreasing at an alarming rate as a result of habitat loss, climate change, pesticide use, and other anthropogenic factors (Davis et al., 2024). Significant resources and awareness have been raised to conserve the monarch butterfly species

and restore their habitats. Identifying monarch butterfly habitats through species distribution estimation is critical to guide conservation efforts and measure general environmental health.

Estimation of monarch species distributions presents many challenges. Monarchs travel vast distances within short time periods due to their migration and their multi-generational nature also causes problems for individual butterfly tracking. Traditional monarch butterfly tracking consisted of field surveys, mark-recapture studies, and counting monarchs at overwintering sites. While these methods are direct and target monarch abundance, they suffer from high labor costs and limited coverage across both space and time due to the necessity of tracking small population sizes across vast geographic areas. The advent of citizen science data has massively improved the coverage and availability of monarch data, providing researchers new opportunities to estimate a more comprehensive picture of the monarch's distribution. Unfortunately, citizen science data is not without its own issues, due to its underlying nonrandom sampling scheme. Ecological datasets often exhibit dependence between the species distribution and the observed locations reported by citizen scientists. The bias in both statistical inference and prediction from direct use of citizen science data is well-documented, and can sometimes lead to erroneous and damaging conclusions. The potential dependence between the response of interest and the observation locations goes by many names, including *preferential sampling* and *mark-point dependence* in the geostatistics literature. In this paper, we will refer to the dependence as *varying sampling effort* (VSE), as is commonly done in the ecology literature.

Several techniques have been proposed to account for VSE in data collected from complex, unknown sampling schemes. Tang et al. (2021a) use a Bayesian framework to estimate abundance of birding activity from the citizen science eBird database by simultaneously modeling the intensity of the desired point process and an effort variable, which is some function of the total number of visits and total time spent ob-

serving at a location. A shared underlying Gaussian process is included as a covariate in both the perceived sampling effort intensity function and the functional form of the effort variable.

A slightly different approach is taken by Sicacha-Parada et al. (2021), whom model sightings of moose in Norway as a thinned or degraded point process. The probability of detection is modeled as a log-linear function of covariates, and can be made more flexible by considering a set of basis functions (Yuan et al., 2017). Because they estimate their model with integrated nested Laplace approximation (INLA), keeping a log-linear structure with the latent parameters is paramount to ensuring a good approximation.

Despite existing approaches to adjust for VSE, studies estimating monarch species distributions using citizen science data have yet to incorporate proper adjustments (Flockhart et al., 2013, 2019; Kendrick and McCord, 2023; Erickson et al., 2023). The most common strategy is to include human population density as an adjustment covariate in the point process model. However, this only quantifies the effect of human population density rather than truly adjusting intensity estimates for increased sampling effort. Additionally, these approaches fail to account for residual spatial variation in intensity after covariate adjustment, implicitly assuming that all variation is explained by the selected covariates. To address these limitations, we use integrated nested Laplace approximation (INLA) and the SPDE approximation to Gaussian random fields to estimate Bayesian log-Gaussian Cox process models, incorporating varying sampling effort in monarch butterfly citizen science data. Our analysis uses Journey North data from 2011 to 2020.

5.2 Data

5.2.1 Adult monarch sightings

We used observations of adult monarch sightings from the Journey North project, a citizen science platform where users can report sightings of monarchs and other species, recording the time and location of each observation, as well as the number of monarchs estimated, comments and pictures. For this analysis, we used a curated Journey North dataset containing adult monarch sightings from 2011 to 2020. Our study focused on the United States at and west of the Rocky Mountains, including the states of Washington, Oregon, California, Nevada, Idaho, Montana, Wyoming, Colorado, Arizona, Utah, and New Mexico. While temporal patterns in monarch sightings provide valuable ecological insights, our primary interest was in species distribution across all years and seasons. To maximize spatial coverage and statistical power, we pooled observations across all years.

It is important to note that Journey North data fall into the category of presence-only ecological data (Renner et al., 2015). These observations indicate when and where monarchs were sighted but provide no information about locations where monarchs were absent. In contrast, presence-absence data include both confirmed sightings and explicit non-detections, offering more comprehensive information and allowing for a wider range of modeling approaches. The limitations of presence-only data constrain our methodological choices, making point process models a natural and appropriate framework for our analysis. By modeling the spatial intensity of observed sightings, we aim to infer monarch distribution patterns while acknowledging the inherent biases and challenges of opportunistic citizen science data.

5.2.2 Covariates

Explanatory variables

Based on previous literature (Flockhart et al., 2019; Davis et al., 2024), we expected monarch observations to decrease with colder temperatures and heavier precipitation due to increased difficulty of flight. As caterpillars, monarchs feed exclusively on milkweed, but as adults, they rely on nectar from various flowering plants. However, the exact distributions of specific nectar sources are difficult to obtain and estimate. To address this, we used the Normalized Difference Vegetation Index (NDVI) as a proxy for nectar availability. NDVI is a widely used remote sensing metric that measures vegetation greenness and biomass by comparing the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs). Higher NDVI values indicate greater vegetation density and productivity, making it a useful indicator of potential nectar sources for monarchs. Lower values indicate sparse vegetation or non-vegetated surfaces. The NDVI is a dimensionless index ranging from -1 to 1. Previous studies have shown a positive correlation between NDVI and monarch presence. Therefore, our primary covariates for modeling monarch species distribution were NDVI, 7-day average precipitation (millimeters), and 7-day average temperature (Celsius), as these factors are expected to influence both nectar availability and monarch activity.

The 7-day average weather data was computed as the average total precipitation and average mean temperature (sum of the daily minimum and maximum temperatures divided by two) over the seven days leading up to and including the sighting date. Weather data was obtained from the DAYMET database, which provides daily, gridded meteorological data at a spatial resolution of 1 km $\times$ 1 km across North America. DAYMET data are derived from a combination of ground-based weather station observations and interpolation techniques, making it a valuable resource for

high-resolution climate data in ecological studies. Each monarch sighting location was matched to the corresponding 1 km grid cell in the DAYMET dataset based on its recorded latitude and longitude.

For same-day NDVI data, we used the NASA MODIS (Moderate Resolution Imaging Spectroradiometer) database, which offers global remote sensing data derived from instruments aboard the Terra and Aqua satellites. MODIS provides multiple NDVI products at different resolutions and time scales. We used the MOD13Q1 NDVI product, which has the highest spatial resolution available for MODIS NDVI at 250 meters and provides 16-day composite images that account for cloud cover and atmospheric conditions. Each monarch sighting was matched to the 250-meter raster cell containing the recorded latitude and longitude to estimate the vegetation density at that location on the sighting date.

Effort metrics

Sampling effort can be measured in various ways. Tang et al. (2021a) proposed metrics such as observation duration, the number of site visits per location, and the percentage of weekend visits at a given site. However, a key limitation of these metrics is their reliance on sightings in existing locations (and times) in the dataset. If observation duration or the percentage of weekend visits are used as covariates to adjust for VSE, their values must also be known in locations where no sightings have been recorded. To address this, the authors impute effort metrics in unsampled locations using the average values observed in the data. However, we find this approach undesirable due to the well-documented bias issues associated with mean imputation in the missing data literature. Instead, we select effort metric covariates derived from pre-available raster datasets, ensuring that effort values can be computed for any location and time point within our study period.

Previous monarch distribution studies have relied on human population density as

the key adjustment variable for VSE (Flockhart et al., 2019; Momeni-Dehaghi et al., 2021). In a separate study on species distribution modelling of moose in Norway, Sicacha-Parada et al. (2021) hypothesized distance from the nearest road contains the key information for measuring sampling effort. Since both metrics are available as rasters, we assessed the utility of both as VSE adjustment factors. We collected human population density from the Gridded Population of the World, Version 4 (GPWv4), Revision 11 (v4.11) dataset which provides estimates of human population counts and densities on a global scale. It is produced by the Center for International Earth Science Information Network (CIESIN) at Columbia University.

We obtained human population density data from GPWv4.11 for the year 2015, cropping the dataset to the western United States. The dataset is gridded at a spatial resolution of 30 arc-seconds (approximately 1 km at the equator), providing population density estimates derived from national and subnational census data. The GPWv4 dataset does not apply ancillary data such as land cover or road networks in its population allocation, ensuring that the population density values strictly reflect administrative unit-based distributions.

For distance to the nearest road, we used data from OpenStreetMap (OSM) for the same region of the western United States. OpenStreetMap is a collaborative, open-source mapping project that provides detailed geospatial data on roads, infrastructure, and other geographic features worldwide. The dataset is continually updated by contributors, incorporating both official sources and user-generated edits. For our analysis, we filtered the OSM road network to include only major "motorways," which correspond to interstate highways and other high-capacity roads designed for fast, long-distance travel. Using this subset, we computed the Euclidean distance from the center of each 1km $\times$ 1km raster cell to the nearest identified highway. This approach provides a straightforward measure of accessibility to major transportation routes and human population centers. We assume that the most recent available

OSM update at the time of data extraction (2020) is representative of road infrastructure throughout the Journey North study period. While road networks may have undergone modifications over time, major highways tend to remain stable, making this assumption reasonable for our purposes.

We found that much of the information captured by the human population density variable was largely contained within the distance to the nearest road variable. Given this overlap, we selected distance to the nearest road as our primary VSE effort metric.

5.3 Methods

Our goal was to model the species distribution of the adult monarch across the western USA using Journey North data from 2011-2020 while adjusting for varying sampling effort. We assume a thinned point process model that allows for residual latent spatial variation through a Bayesian LGCP mechanism, following the model detailed in Chakraborty et al. (2011), Yuan et al. (2017), and Sicacha-Parada et al. (2021). Estimation proceeds by INLA in Rue et al. (2009), the SPDE method for approximating Gaussian processes in Lindgren et al. (2011), and the LGCP point process likelihood computation from Simpson et al. (2016).

5.3.1 Statistical model and priors

To model the monarch sighting data, we assume the observations $\mathcal{S} = (\mathbf{s}_1, ..., \mathbf{s}_n)$ follow a thinned log Gaussian Cox process with log intensity $\log(\lambda(\mathbf{s})q(\mathbf{s}))$ where the thinning probability is denoted by q. Let $\mathbf{s} \in \mathcal{D} \subseteq \mathbb{R}^2$ represent a spatial location $\mathbf{s}$ in the study region $\mathcal{D}$. Further, let $\lambda(\mathbf{s})$ define the true intensity of the monarch distribution intensity, ω be a stationary Gaussian random field with Matérn covariance and $q(\mathbf{s})$ be the thinning probability. Then our full model follows

$$\log(\lambda(\mathbf{s})q(\mathbf{s})) = \mathbf{X}^{\top}(\mathbf{s})\beta + \omega(\mathbf{s}) + \log(q(\mathbf{s})), \tag{5.1}$$

with half normal detection probability function $q(\mathbf{s})$ as defined in Yuan et al. (2017) as

$$q(\mathbf{s}) = \exp\left\{-\gamma \cdot d(\mathbf{s})^2/2\right\}; \gamma > 0, \tag{5.2}$$

where γ measures the magnitude of VSE and d represents the distance to the nearest road.

The parameter vector of interest to be estimated is $\theta = (\beta, \gamma, \sigma^2, \phi)$. We fixed the smoothness parameter ν at 1, which is the maximum allowable value in the INLA-SPDE model. However, recent advances in the rational SPDE approach have introduced methods for accommodating other values of ν within the INLA-SPDE framework (Bolin and Kirchner, 2020). In practice, ν is often fixed in applied settings. For the range and variance parameters, we use penalized complexity (PC) priors (Simpson et al., 2017; Fuglstad et al., 2019), which offer an alternative to traditional parametric probability distributions by providing a more interpretable way to encode prior beliefs. Specifically, we place PC priors on the range and variance parameters of the Gaussian random field ω. The range parameter ϕ represents the approximate distance beyond which points in the field are considered independent, while the variance σ^2 governs overall variability. We set $\mathbb{P}(\phi < 200) = 0.01$ and $\mathbb{P}(\sigma > 2) = 0.01$ as our PC priors, where the range is measured in kilometers. For the fixed effect parameters β associated with environmental covariates, we assume Gaussian priors with mean 0 and variance 100. The last parameter, the VSE effect, is log-transformed before estimation to ensure positivity on the natural scale. On the log scale, we impose a normal prior with mean 1 and variance 20.

5.3.2 Estimation and computation

The main implementation for estimating our Bayesian thinned log Gaussian Cox process (LGCP) model is the `R-INLA` software. The estimation procedure consists of three key components: (1) integrated nested Laplace approximation (INLA) for fitting Bayesian latent Gaussian models, (2) the stochastic partial differential equation (SPDE) approach for approximating the continuous spatial Gaussian random field, and (3) an LGCP modeling technique that avoids the traditional discretization of the study region.

Performing Bayesian inference poses a computational challenge due to the need to evaluate the posterior distribution. Markov Chain Monte Carlo (MCMC) methods provide a simulation-based approach to obtain exact posterior samples by constructing a Markov chain whose stationary distribution matches the target posterior. However, MCMC methods are computationally expensive, particularly for high-dimensional spatial models. INLA has emerged as a computationally efficient alternative for latent Gaussian models (Rue et al., 2009). Unlike standard Bayesian inference methods, INLA provides marginal posterior distributions for each parameter rather than the full joint posterior, significantly improving computational speed while maintaining accurate approximations.

Within the INLA framework, the SPDE approach can be used to approximate a Gaussian random field (GRF) when the GRF serves as the latent Gaussian model. Lindgren et al. (2011) demonstrated that a stationary GRF with Matérn covariance can be represented as the solution to a stochastic partial differential equation. This representation allows the GRF to be expressed as a linear combination of basis functions, which are defined over a triangulated mesh of the study region. This approximation introduces a few user-controlled hyperparameters, including the number of basis functions and the number of vertices in the triangulation, which influence computational efficiency and model accuracy. The `R-INLA` software abstracts much

of this process by constructing a mesh over the study region, which serves as the foundation for the approximation (Figure 5.1).

The final component of the estimation procedure is the LGCP modeling technique introduced by Simpson et al. (2016). The likelihood of an LGCP model involves evaluating an integral over the entire study region, which does not have a closed-form solution. Traditionally, this integral is approximated by discretizing the study region, but finer discretizations—while improving accuracy—can become computationally infeasible. To address this, Simpson et al. (2016) proposed an alternative approach that avoids traditional discretization. Instead of relying on increasingly fine grid-based approximations, their method leverages the finite-dimensional representation from the SPDE approach to approximate the underlying Gaussian random field (GRF). This allows for a more computationally efficient and stable evaluation of the integral without requiring excessive discretization.

5.3.3 Evaluation

To test our hypothesis regarding the importance of distance to the nearest road as a driver of sampling effort, we fit two models: a standard LGCP model without a VSE term, referred to as the unadjusted model (i.e., (5.1) without q), and the full model incorporating VSE, as specified in (5.1) which we call the VSE model. In addition to examining the estimated coefficients and standard errors, we assess global model fit using three widely used Bayesian model selection criteria: the Deviance Information Criterion (DIC), the Watanabe-Akaike Information Criterion (WAIC), and the Logarithm of the Pseudo Marginal Likelihood (LPML).

DIC is a generalization of AIC for Bayesian models, balancing model fit (via the deviance) and model complexity (by penalizing excessive parameters). It is computationally efficient and widely used, but it can be less reliable for hierarchical models or those with latent Gaussian fields, where the effective number of parameters is dif-

ficult to estimate accurately. WAIC, on the other hand, is a fully Bayesian criterion that evaluates model fit using pointwise log predictive density, averaging over the posterior distribution. Unlike DIC, WAIC does not rely on a point estimate and is asymptotically equivalent to leave-one-out cross-validation. While WAIC provides a more theoretically justified approach, it is computationally more intensive and can be sensitive to the method used for posterior approximation.

Lastly, LPML is a cross-validation based metric that directly measures predictive accuracy using the conditional predictive ordinate (CPO), effectively evaluating how well the model predicts each observation given the rest of the data. LPML is particularly useful for assessing real-world predictive performance, but it is computationally demanding and can be numerically unstable when extreme CPO values occur. By comparing these three metrics, we aim to obtain a comprehensive evaluation of model fit, balancing considerations of predictive accuracy, model complexity, and computational feasibility.

5.4 Results

During the study period from 2011 to 2020, a total of 5,400 adult monarch sightings were recorded with most heavily clustered in metropolitan areas. After filtering to unique observations by latitude and longitude, 3,515 unique observations remained. Sightings during the breeding months were concentrated primarily in California's coastal regions, the Denver area in Colorado, and the Albuquerque area in New Mexico (Figure 5.2). During the overwintering months, sightings were largely restricted to coastal California and the Phoenix area in Arizona (Figure 5.3). The median distance to the nearest road for monarch sightings was 4.54 km, compared to an average of 54.5 km across all grid cells in the study area. Figure 5.4 illustrates spatial rasters of the environmental covariates in 2020 used for predictions.

Figure 5.5 shows the log intensity estimates for the true monarch distribution (λ, excluding q) for both the naive and VSE models. Figure 5.6 illustrates the difference in log intensities between the two models. As expected, the log intensity at each point in the VSE model is equal to or greater than that in the naive model, with differences ranging from 0 to 3. Notable regions where the VSE model shows nontrivial log intensity (above 0) that was absent in the naive model include the Colorado-New Mexico border, the southeast corner of New Mexico, the Seattle area in the Pacific Northwest, and the Nevada-California border. The largest differences in log intensity occur in areas farthest from major roadways, particularly in regions with observed clusters of points, implying these differences are not merely an artifact of the introduced distance sampling effect. Uncertainty in regions with increased log intensity also increased, though it slightly decreased in certain areas near roadways (Figure 5.7). Despite these changes in log intensity, the intensity on the normal scale remains relatively low compared to higher-density regions, such as California. The estimated relationship between distance to the nearest road and the thinning probability follows a half-normal, monotonically decreasing function (Figure 5.8). The thinning probability is 50% at approximately 100 km, decreasing to near zero around 250 km.

Model fit metrics showed little difference between the naive and VSE models (Table 5.6). While LPML appeared to show a clear difference between the two, we found that this was due to extreme log values when the intensity was near zero, meaning the differences in LPML did not reflect meaningful differences in predictive performance. DIC indicated a slightly improved fit for the VSE model, whereas WAIC suggested similar overall fit between the two models.

The posterior marginal distributions for the coefficients of environmental covariates, including NDVI, precipitation, temperature, and the precipitation-temperature interaction, were also similar between the naive and VSE models (Figure 5.9, Ta-

ble 5.2). The direction of each estimated coefficient aligned with prior expectations, with NDVI and temperature having a positive effect and total precipitation having a negative effect. The interaction between precipitation and temperature was also positive, suggesting that the negative effect of precipitation diminishes as temperatures increase. For the distance sampling function, the γ coefficient was estimated with a median of 1.43 and a 95% credible interval of (1.03, 1.98), providing clear evidence of an effect of distance from the nearest road on the probability of observation.

A more significant difference between the models was observed in the parameters estimated for the Gaussian Random Field (GRF). In the naive model, the range was estimated at 287.8km (220.2, 375.5), and the standard deviation was 1.79 (1.48, 2.18). In contrast, the VSE model produced higher estimates for both parameters, with the range at 303.8km (228.3, 411.3) and the standard deviation at 1.88 (1.51, 2.36).

5.5 Discussion

Previous studies on monarch species distribution modeling have accounted for key environmental drivers but have often overlooked varying sampling effort and residual spatial variation. The primary contribution of this study lies in quantifying the extent of VSE in citizen science data for monarchs. To address this, we used INLA to estimate a Bayesian LGCP model that explicitly adjusts for VSE in adult monarch species distribution modeling across the western USA. Our results indicate that estimates of key environmental covariates align well with prior findings in the literature. Additionally, we found strong evidence that road-based distance sampling influences observed monarch locations in the Journey North dataset. Our thinned point process model provides clear evidence that the observed monarch distribution is partially shaped by sampling effort variations related to proximity to major roadways. This effect likely acts as a proxy for broader factors, such as increased human activity, park

accessibility, and observer behavior patterns, all of which contribute to an uneven spatial distribution of observations.

An interesting finding is that incorporating VSE had little to no impact on the statistical inference for environmental covariates. The estimated effects of NDVI, precipitation, temperature, and their interaction remained consistent in both sign and magnitude. Prior research has shown that ignoring VSE can significantly alter statistical inferences (Sicacha-Parada et al., 2021), so the fact that we did not observe this effect suggests either that sampling effort is largely independent of the key environmental drivers in our study area, or that our dataset is sufficiently large and well-distributed such that unaccounted VSE does not introduce substantial bias in these estimates. This may indicate that monarch presence, conditional on sampling effort, is still strongly driven by environmental factors, minimizing distortions in inference.

Global model fit was similar between the VSE and unadjusted models. The DIC slightly favored the VSE model, while the WAIC showed no meaningful difference between the two models. However, interpreting global fit metrics in point process models remains challenging, particularly due to the high prevalence of zero-intensity locations. Fit metrics like the LPML operate on a log scale, which can artificially amplify discrepancies at near-zero intensity locations, making direct comparisons less reliable. We believe further research on model evaluation techniques for point process models, especially those with extensive zero-intensity regions, is an important direction for future work.

Our study has several strengths. It is the first in monarch species distribution modeling to adjust for both residual spatial variation and VSE. Additionally, we incorporate fine-scale environmental covariates to improve intensity surface predictions and conduct statistical inference on the relationship between monarch distribution and environmental factors. However, several limitations remain. While we quantify

the relationship between sampling effort and distance to the nearest road, the true selection mechanism is likely more complex, with substantial information on sampling effort missing from the dataset. More flexible modeling approaches or additional data could help address these gaps. In particular, incorporating additional environmental data such as monarch roosts, caterpillar sightings, and milkweed presence could provide a more system-oriented and holistic model for monarch distribution. Additionally, while monarch sighting counts are unreliable, they could still offer valuable information beyond presence-only models, potentially enabling direct abundance modeling. Expanding the analysis to include other citizen science datasets beyond Journey North is also possible but would require careful consideration of the data-generating process before simply pooling datasets together.

Ultimately, our findings suggest that additional data is needed to refine the VSE approach and improve the utility of citizen science data in species distribution modeling. More detailed information on sampling effort metrics would be invaluable for estimating a more nuanced and flexible VSE function. Furthermore, regions with higher standard error and greater intensity under the VSE model could be identified as priority areas for additional data collection. By highlighting these regions, Journey North and similar citizen science platforms could actively encourage users to submit more observations in underrepresented areas, improving future models and species distribution estimates.

5.6 Figures and Tables

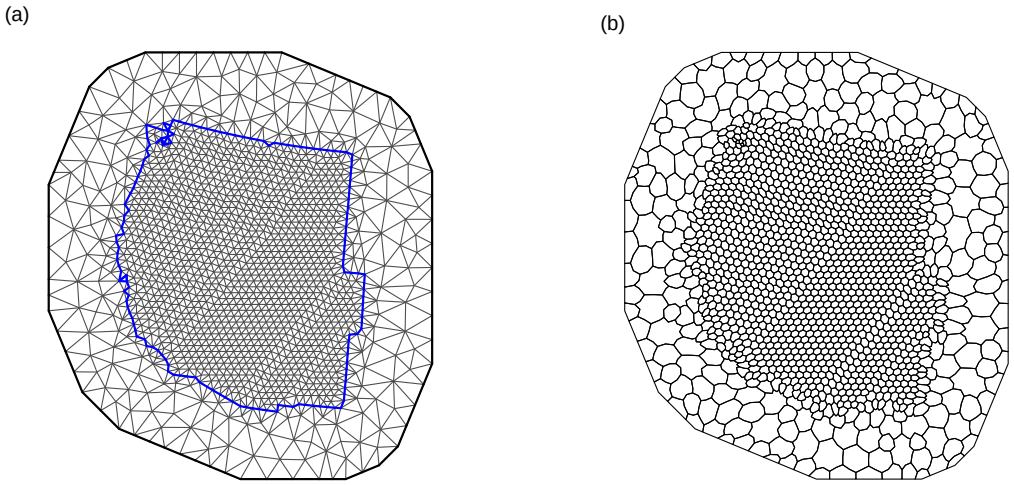

Figure 5.1: Mesh triangulation (left) for the western USA for the SPDE method and dual mesh (right) for the approximation to the LGCP likelihood.

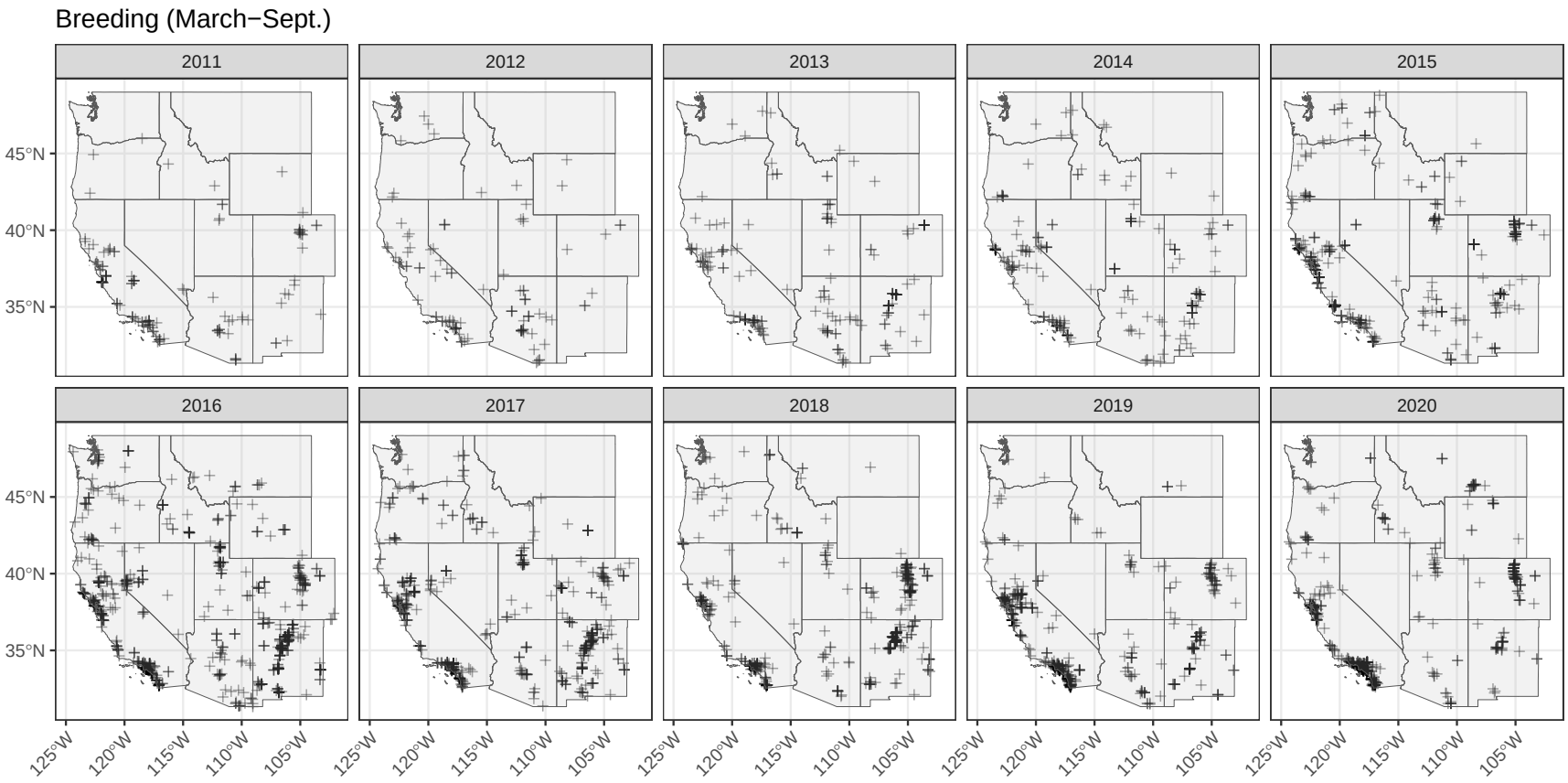

Figure 5.2: Adult monarch sightings in the Journey North Dataset during breeding months (October to February).

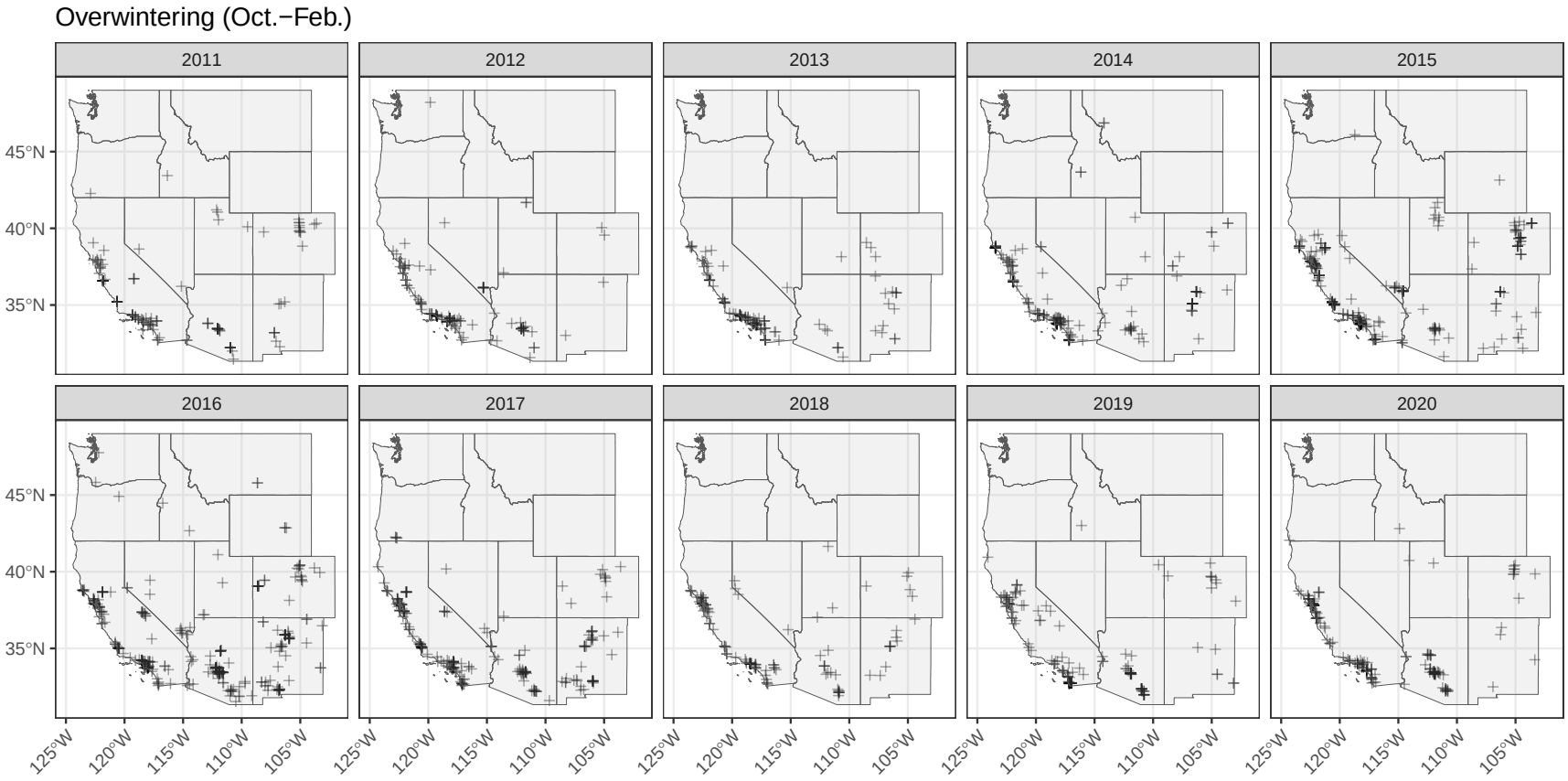

Figure 5.3: Adult monarch sightings in the Journey North Dataset during overwintering months (March to September).

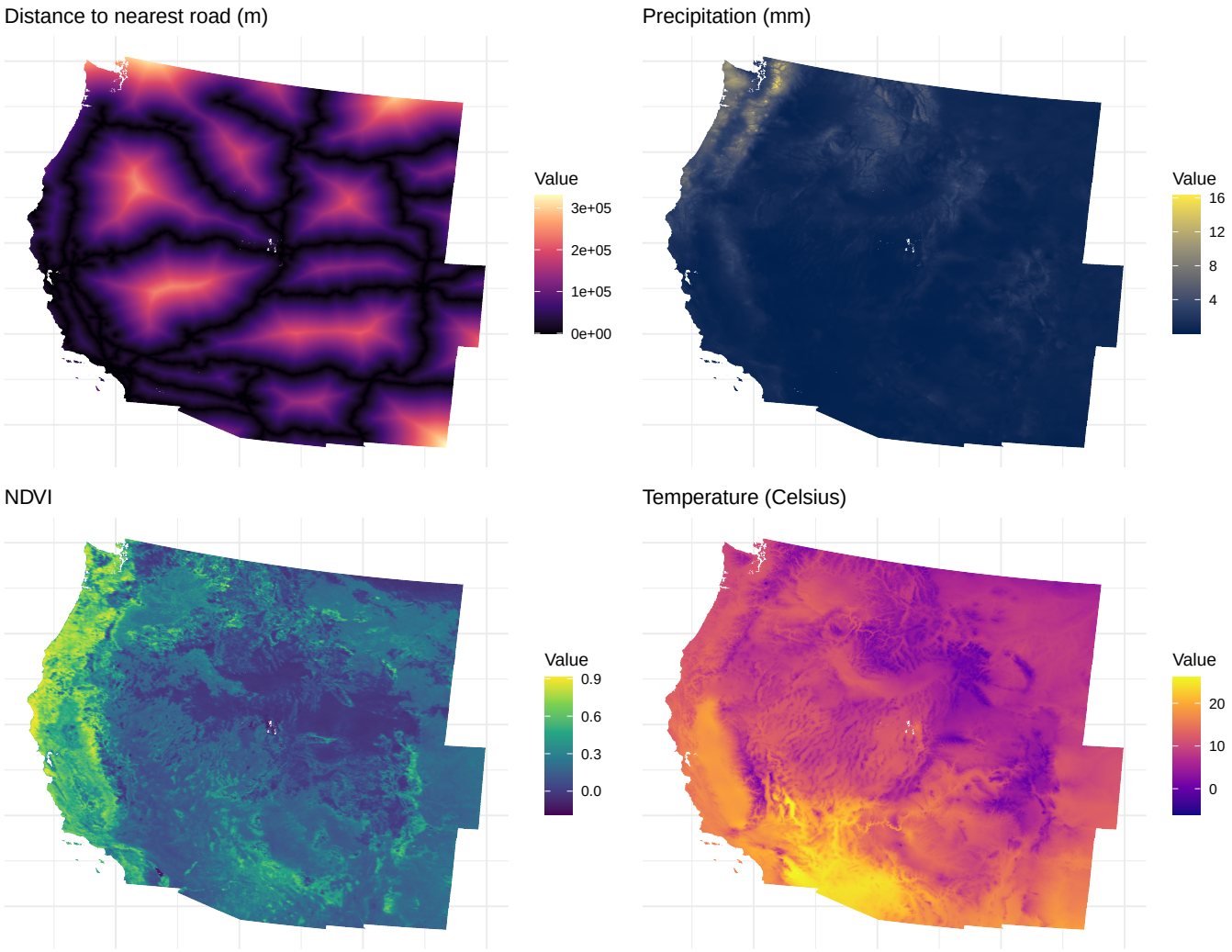

Figure 5.4: Spatial rasters of environmental covariates used for the naive and VSE model in the year 2020.

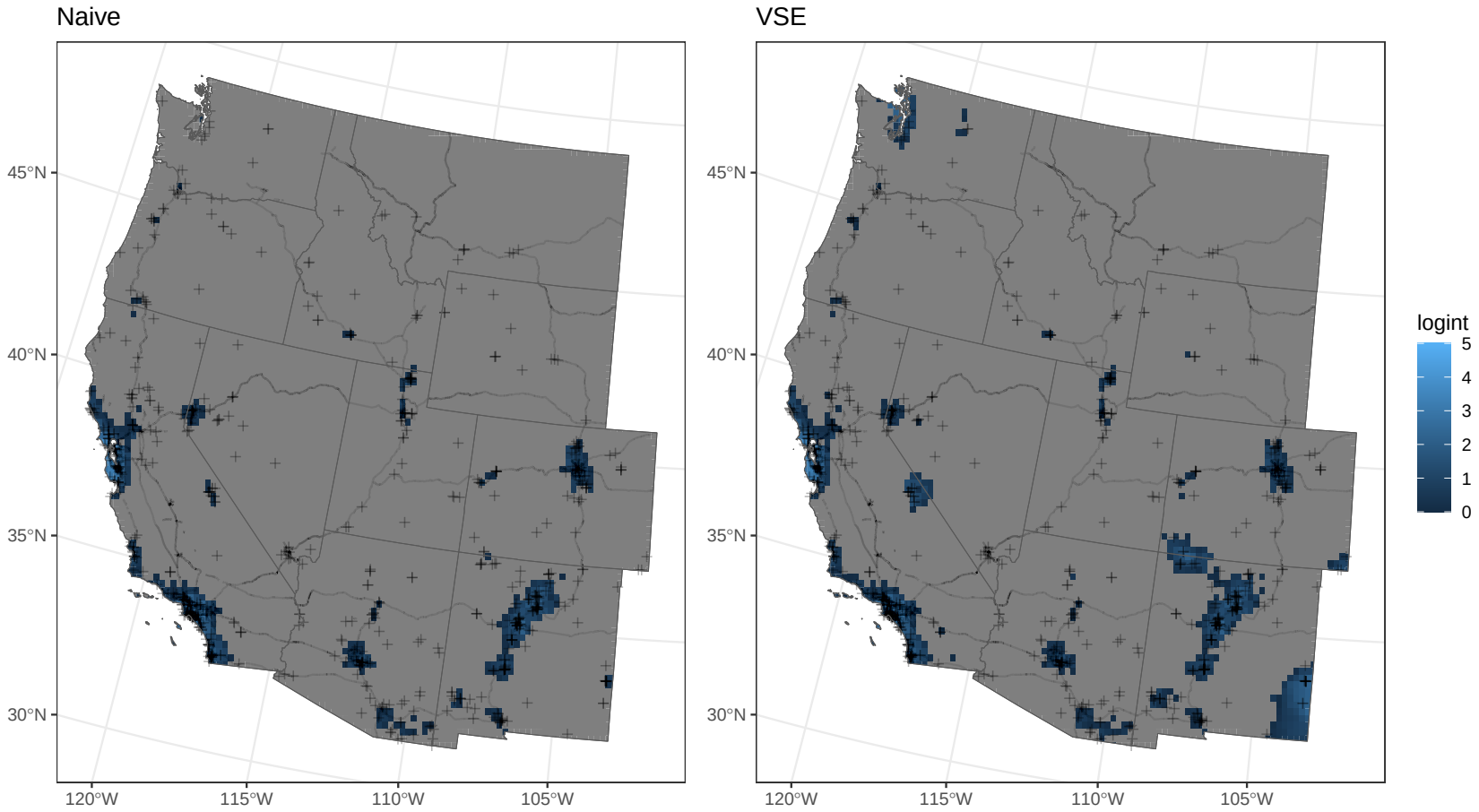

Figure 5.5: Estimated mean log intensity for the VSE and naive models with major motorways highlighted.

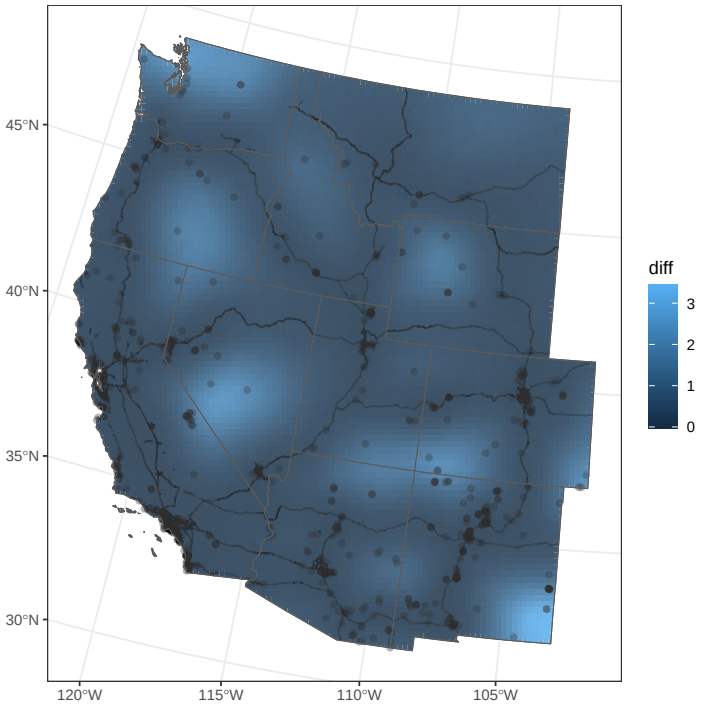

Figure 5.6: Difference in the mean log intensity between the VSE model and the naive model with major motorways highlighted (VSE - Naive).

Model	WAIC	DIC	LPML
Naïve	1909.4	773.0	19082
VSE	1910.1	756.3	27859

Table 5.1: Bayesian global model fit metrics between the naive and VSE models.

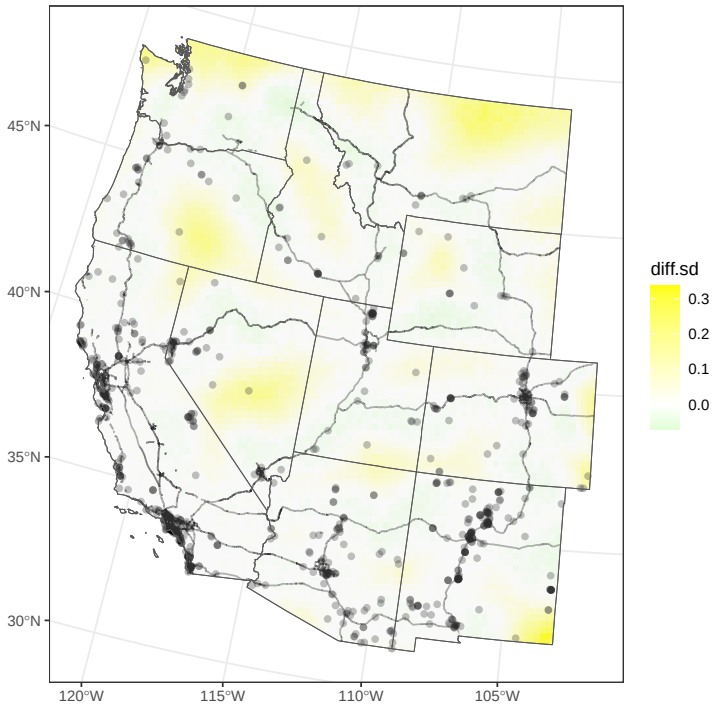

Figure 5.7: Difference in the standard deviation of log intensity between the VSE model and the naive model with major motorways highlighted (VSE - Naive).

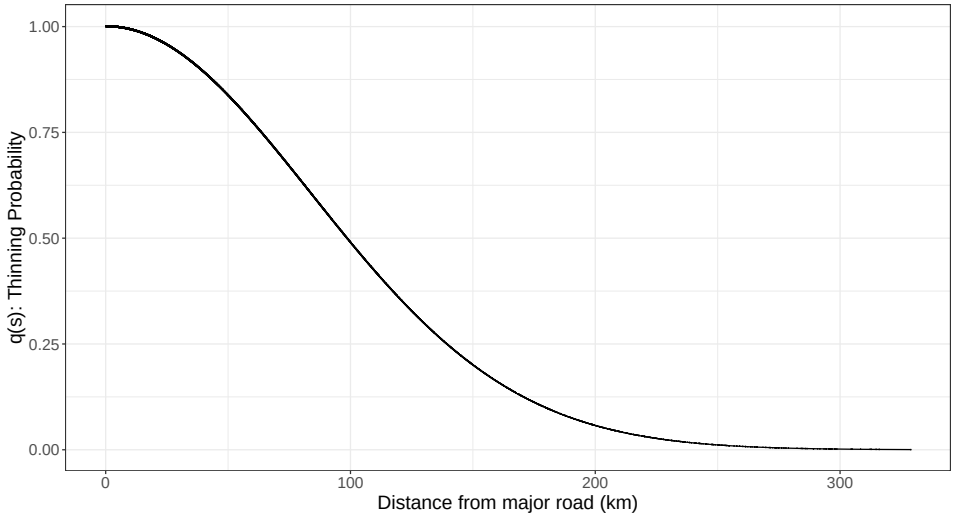

Figure 5.8: Estimated half-normal detection function with 95% credible intervals from the VSE model.

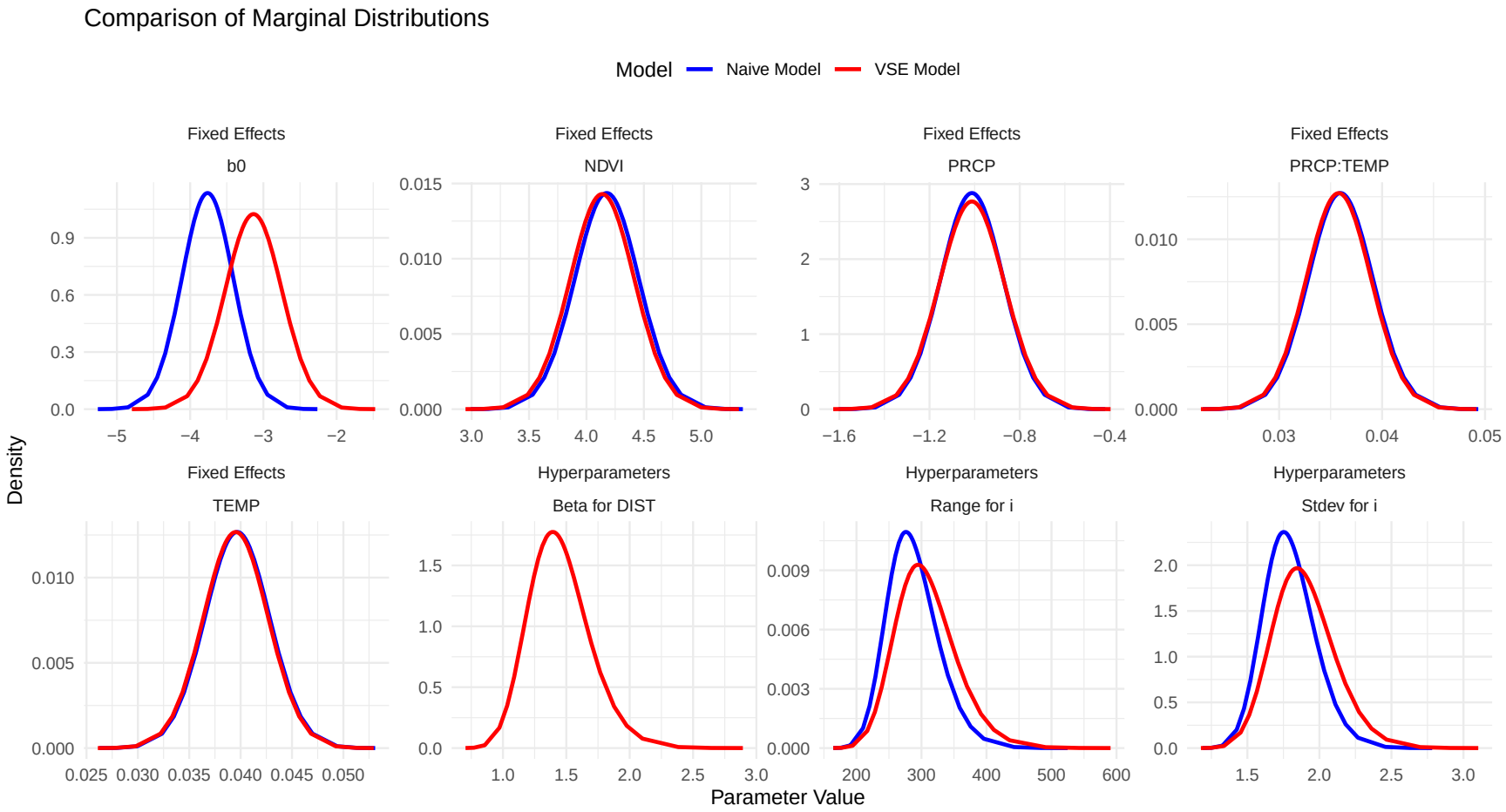

Figure 5.9: Estimated posterior marginal distributions of model parameters between the Naive and VSE models.

Table 5.2: Comparison of mean (SD) and quantiles of posterior parameter estimates between Naive and VSE models

	Naive Model				**VSE Model**			
Parameter	Mean (SD)	0.025	0.5	0.975	Mean (SD)	0.025	0.5	0.975
Range for i	287.83 (39.53)	220.22	284.39	375.54	307.92 (46.58)	228.36	303.82	411.32
Stdev for i	1.79 (0.18)	1.48	1.78	2.18	1.89 (0.22)	1.51	1.88	2.36
β_0	-3.76 (0.35)	-4.45	-3.76	-3.07	-3.13 (0.39)	-3.90	-3.13	-2.37
PRCP	-1.01 (0.14)	-1.28	-1.01	-0.74	-1.01 (0.14)	-1.29	-1.01	-0.73
TEMP	0.0397 (0.0031)	0.0335	0.0397	0.0458	0.0395 (0.0031)	0.0334	0.0395	0.0457
NDVI	4.18 (0.28)	3.63	4.18	4.72	4.14 (0.28)	3.59	4.14	4.68
PRCP:TEMP	0.03596 (0.0031)	0.02981	0.03596	0.04211	0.03579 (0.0031)	0.02963	0.03579	0.04194
DIST (γ)	–	–	–	–	1.45 (0.24)	1.03	1.43	1.98

Chapter 6

Future work

This chapter details some ideas for future work in preferential sampling.

6.1 Extension of the SLP to more flexible point processes

Current formulations of the SLP assume that the observations follow an inhomogeneous Poisson process (IPP), conditional on the underlying random field S. A key property of an IPP is the conditional independence of observation points, given S. However, this assumption has been widely criticized as unrealistic in many applications (Diggle et al., 2010; Gray and Evangelou, 2023).

Consider the air pollution monitoring example: Even with knowledge of high and low pollution areas, a network designer may deliberately space monitors apart to avoid redundancy. This violates the IPP assumption of conditional independence, as observation locations exhibit spatial inhibition rather than complete randomness. A promising direction for expanding the class of eligible observation processes is the (sparse) literature on point processes that exhibit large-scale clustering while maintaining local-scale inhibition (Lavancier and Møller, 2016). These processes exhibit

both aggregation at a broader scale and spatial regularity within clusters (or vice versa). Such models could be strong candidates for SLP-based inference in preferential sampling settings, particularly if they include the standard log Gaussian Cox process from the original SLP as a special case.

One potential extension is what we will call the determinantally thinned LGCP (DT-LGCP). This process begins as an LGCP but undergoes determinantal thinning, where the probability of retaining each point depends on the determinant of a covariance matrix. Determinantal thinning introduces repulsion between points, naturally discouraging excessive clustering while still preserving large-scale variation. This approach captures spatial inhibition while maintaining the flexibility of the LGCP. Sampling such a process is theoretically straightforward but presents computational challenges, particularly the potential singularity of the sampling matrix. This issue is exacerbated by the strong clustering typical of LGCPs. A detailed study of the properties of the DT-LGCP, along with an efficient software implementation, could significantly extend the utility of the traditional SLP framework, allowing it to model a wider range of preferential sampling designs beyond its current limitations.

6.2 Simultaneous weight estimation in ISIW

The ISIW method from Chapter 4 is a frequentist two-stage estimation approach. In the first stage, observation intensities are estimated. These estimates are then treated as fixed and used in the full likelihood estimation in the second stage. We demonstrated that existing nonparametric estimators for first-order intensity surfaces of point processes are often numerically unstable. In addition, without covariate information, no consistency result can be established (Guan, 2008), so that even with $n \to \infty$ there is no guarantee the estimated weights will improve. While parametric estimation theoretically offers better guarantees under appropriate assumptions

(Schoenberg, 2005; Simpson et al., 2016), our simulations showed that the estimates were still inaccurate and numerically unstable (prior to Winsorization) compared to using the true underlying weights.

An alternative is to integrate weight estimation directly into the likelihood, effectively merging the two stages. This approach parallels Bayesian methods for (non-spatial) models under informative sampling: Savitsky and Toth (2016) introduced a method treating weights as given, while León-Novelo and Savitsky (2019) extended it to jointly estimate the weights in a fully Bayesian framework. However, such models quickly become complex, potentially negating the computational advantages of ISIW and resembling a full marked point process model, akin to the SLP. Notably, several promising intensity estimators using sophisticated tree-based methods have emerged (Lamprinakou et al., 2023; Lu et al., 2024), which could be the solution to achieving the accuracy needed for proper PS adjustment.

6.3 Incorporation of positional error to preferential sampling models

Another common assumption in model-based geostatistics is that observation locations are known exactly. In reality, particularly in wildlife ecology, survey statistics, and other field-based applications, the exact location of observations is often uncertain and subject to measurement error. This issue, referred to as location uncertainty or positional error, is well-documented in the spatial epidemiology and model-based geostatistics literature (Gabrosek and Cressie, 2002; Zimmerman, 2007; Zandbergen, 2008). Several methods have been proposed to address positional error. These include kriging adjusted for location error (KALE) (Cressie and Kornak, 2003), a composite likelihood approach (Fronterrè et al., 2018), a regression calibration method (Warren et al., 2016), and Bayesian (Wilson and Wakefield, 2021; Miller et al., 2022) and hi-

erarchical model approaches (Altay et al., 2022). Wang et al. (2022) also conducted a theoretical study on the KALE method. For a comprehensive summary of geostatistical inference and prediction under location uncertainty, see Totland (2023).

Despite these advances, no methodological developments have addressed a geostatistical model that simultaneously accounts for both preferential sampling and location uncertainty. This gap is significant, as it is common in wildlife and ecological studies for sampling to be both preferential and spatially imprecise. For example, Dinsdale and Salibian-Barrera (2019b) highlight this issue in marine wildlife studies, while Conroy et al. (2023) describe a case of preferential sampling in coyote plague surveillance, where the actual location of infection is uncertain due to the large distances coyotes travel before being observed by ecologists. A hierarchical modeling approach could provide a natural framework for addressing these challenges, with an additional layer for latent locations alongside the existing layer for a doubly stochastic point process to account for preferential sampling. However, hierarchical models can quickly become computationally complex, and estimation may be difficult, even in advanced frameworks like `TMB` or Stan. Developing a reliable and numerically stable technique for geostatistical inference and prediction in the presence of both preferential sampling and location uncertainty remains an open problem. Given how frequently these issues arise in real-world spatial data, further research is needed to address this methodological gap.

Bibliography

Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A. (2022). Fast geostatistical inference under positional uncertainty: Analysing DHS household survey data. arXiv:2202.11035 [stat].

Amaral, A. V. R., Krainski, E. T., Zhong, R., and Moraga, P. (2023). Model-Based Geostatistics Under Spatially Varying Preferential Sampling. *Journal of Agricultural, Biological and Environmental Statistics*.

Bachoc, F. (2014). Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. *Journal of Multivariate Analysis*, 125:1–35.

Bachoc, F. (2020). Asymptotic analysis of maximum likelihood estimation of covariance parameters for Gaussian processes: an introduction with proofs. arXiv:2009.07002 [math].

Bachoc, F., Bevilacqua, M., and Velandia, D. (2019). Composite likelihood estimation for a Gaussian process under fixed domain asymptotics. *Journal of Multivariate Analysis*, 174:104534.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2008.00663.x.

Berman, M. and Diggle, P. (1989). Estimating Weighted Integrals of the Second-Order Intensity of a Spatial Point Process. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(1):81–92. Publisher: [Royal Statistical Society, Wiley].

Bevilacqua, M. and Gaetan, C. (2015). Comparing composite likelihood methods based on pairs for spatial Gaussian random fields. *Statistics and Computing*, 25(5):877–892.

Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012). Estimating Space and Space-Time Covariance Functions for Large Data Sets: A Weighted Composite Likelihood Approach. *Journal of the American Statistical Association*, 107(497):268–280. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2011.646928.

Bolin, D. and Kirchner, K. (2020). The Rational SPDE Approach for Gaussian Random Fields With General Smoothness. *Journal of Computational and Graphical Statistics*, 29(2):274–285. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2019.1665537.

Bolin, D. and Wallin, J. (2024). Spatial confounding under infill asymptotics. arXiv:2403.18961 [math].

Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference: Theory and Methods*. Springer Science & Business Media. Google-Books-ID: 8VNDAAAAQBAJ.

Cecconi, L., Biggeri, A., Grisotto, L., Berrocal, V., Rinaldi, L., Musella, V., Cringoli, G., and Catelan, D. (2016). Preferential sampling in veterinary parasitological surveillance. *Geospatial Health*, 11(1). Number: 1.

Chakraborty, A. and Gelfand, A. E. (2010). Analyzing spatial point patterns subject

to measurement error. *Bayesian Analysis*, 5(1):97–122. Publisher: International Society for Bayesian Analysis.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., and Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(5):757–776. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2011.00769.x.

Chen, H.-S., Simpson, D., and Ying, Z. (2000). Infill Asymptotics for a Stochastic Process Model with Measurement Error. *Statistica Sinica*.

Conn, P. B., Thorson, J. T., and Johnson, D. S. (2017). Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution*, 8(11):1535–1546. _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12803.

Conroy, B., Waller, L. A., Buller, I. D., Hacker, G. M., Tucker, J. R., and Novak, M. G. (2023). A Shared Latent Process Model to Correct for Preferential Sampling in Disease Surveillance Systems. *Journal of Agricultural, Biological and Environmental Statistics*, 28(3):483–501.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2007.00633.x.

Cressie, N. and Kornak, J. (2003). Spatial Statistics in the Presence of Location Error with an Application to Remote Sensing of the Environment. *Statistical Science*, 18(4).

Cressie, N. and Wikle, C. K. (2015). *Statistics for Spatio-Temporal Data*. John Wiley & Sons. Google-Books-ID: 4L_dCgAAQBAJ.

Cronie, O. and Van Lieshout, M. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, 105(2):455–462.

Curriero, F. C. and Lele, S. (1999). A Composite Likelihood Approach to Semivariogram Estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, 4(1):9–28. Publisher: [International Biometric Society, Springer].

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association*, 111(514):800–812. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2015.1044091.

Davis, A. K., Croy, J. R., and Snyder, W. E. (2024). Dramatic recent declines in the size of monarch butterfly (*Danaus plexippus*) roosts during fall migration. *Proceedings of the National Academy of Sciences*, 121(43):e2410410121.

Diggle, P. (1985). A Kernel Method for Smoothing Point Process Data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):138–147. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2307/2347366.

Diggle, P. J., Menezes, R., and Su, T.-l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2009.00701.x.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9876.00113.

Dinsdale, D. and Salibian-Barrera, M. (2019a). Methods for preferential sampling in geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(1):181–198. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssc.12286.

Dinsdale, D. and Salibian-Barrera, M. (2019b). Modelling ocean temperatures from bio-probes under preferential sampling. *Annals of Applied Statistics*, 13(2):713–745. Publisher: Institute of Mathematical Statistics.

Du, J., Zhang, H., and Mandrekar, V. S. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *The Annals of Statistics*, 37(6A):3330–3361. Publisher: Institute of Mathematical Statistics.

Erickson, E., Jason, C., Machiorlete, H., de la Espriella, L., Crone, E. E., and Schultz, C. B. (2023). Using community science to map western monarch butterflies (Danaus plexippus) in spring. *Ecology and Evolution*, 13(12):e10766. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.10766.

Fandos, G., Kéry, M., Cano-Alonso, L. S., Carbonell, I., and Luis Tellería, J. (2021). Dynamic multistate occupancy modeling to evaluate population dynamics under a scenario of preferential sampling. *Ecosphere*, 12(4):e03469. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ecs2.3469.

Ferreira, G. d. S. (2020). Geostatistics under preferential sampling in the presence of local repulsion effects. *Environmental and Ecological Statistics*, 27(3):549–570.

Ferreira, G. d. S. and Gamerman, D. (2015). Optimal Design in Geostatistics under Preferential Sampling. *Bayesian Analysis*, 10(3):711–735. Publisher: International Society for Bayesian Analysis.

Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for

multiple species. *Methods in Ecology and Evolution*, 6(4):424–438. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12242.

Flockhart, D. T. T., Larrivée, M., Prudic, K. L., and Ryan Norris, D. (2019). Estimating the annual distribution of monarch butterflies in Canada over 16 years using citizen science data. *FACETS*, 4(1):238–253. Publisher: Canadian Science Publishing.

Flockhart, D. T. T., Wassenaar, L. I., Martin, T. G., Hobson, K. A., Wunder, M. B., and Norris, D. R. (2013). Tracking multi-generational colonization of the breeding grounds by monarch butterflies in eastern North America. *Proceedings of the Royal Society B: Biological Sciences*, 280(1768):20131087. Publisher: Royal Society.

Fronterrè, C., Giorgi, E., and Diggle, P. (2018). Geostatistical inference in the presence of geomasking: A composite-likelihood approach. *Spatial Statistics*, 28:319–330.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing Priors that Penalize the Complexity of Gaussian Random Fields. *Journal of the American Statistical Association*, 114(525):445–452.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/106186006X132178.

Gabrosek, J. and Cressie, N. (2002). The Effect on Attribute Prediction of Location Uncertainty in Spatial Data. *Geographical Analysis*, 34(3):262–285. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.2002.tb01088.x.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. CRC Press. Google-Books-ID: Xf4leslPDzsC.

Gelfand, A. E., Sahu, S. K., and Holland, D. M. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics*, 23(7):565–578. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.2169.

Gelfand, A. E. and Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, 89(3):e01372. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ecm.1372.

Gihman, I. I. and Skorokhod, A. V. (2004). Absolute Continuity of Measures Associated with Random Processes. In Gihman, I. I. and Skorokhod, A. V., editors, *The Theory of Stochastic Processes I*, pages 440–524. Springer, Berlin, Heidelberg.

Gikhman, I. I. and Skorokhod, A. V. (2004). *The Theory of Stochastic Processes: I.* Springer Science & Business Media. Google-Books-ID: HBDiMpORJkoC.

Gilbert, B., Ogburn, E. L., and Datta, A. (2024). Consistency of common spatial estimators under spatial confounding. *Biometrika*, page asae070.

Gramacy, R. B. and Lee, H. K. H. (2012). Cases for the nugget in modeling computer experiments. *Statistics and Computing*, 22(3):713–722.

Gray, E. J. and Evangelou, E. (2023). A design utility approach for preferentially sampled spatial data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(4):1041–1063.

Guan, Y. (2008). On Consistent Nonparametric Intensity Estimation for Inhomogeneous Spatial Point Processes. *Journal of the American Statistical Association*, 103(483):1238–1247.

Guinness, J. (2018). Permutation and Grouping Methods for Sharpening Gaussian

Process Approximations. *Technometrics*, 60(4):415–429. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00401706.2018.1437476.

Guinness, J. (2021). Gaussian process learning via Fisher scoring of Vecchia's approximation. *Statistics and Computing*, 31(3):25.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F., and Zammit-Mangion, A. (2019). A Case Study Competition Among Methods for Analyzing Large Spatial Data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425.

Ho, L. P. and Stoyan, D. (2008). Modelling marked point patterns by intensity-marked Cox processes. *Statistics & Probability Letters*, 78(10):1194–1199.

Huser, R., Stein, M. L., and Zhong, P. (2023). Vecchia Likelihood Approximation for Accurate and Fast Inference with Intractable Spatial Max-Stable Models. *Journal of Computational and Graphical Statistics*, 0(0):1–22. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2023.2285332.

Ibragimov, I. A. and Rozanov, Y. A. (2012). *Gaussian Random Processes*. Springer Science & Business Media. Google-Books-ID: hB_SBwAAQBAJ.

Karcher, M. D., Palacios, J. A., Bedford, T., Suchard, M. A., and Minin, V. N. (2016). Quantifying and Mitigating the Effect of Preferential Sampling on Phylodynamic Inference. *PLOS Computational Biology*, 12(3):e1004789. Publisher: Public Library of Science.

Katzfuss, M. and Guinness, J. (2021). A General Framework for Vecchia Approximations of Gaussian Processes. *Statistical Science*, 36(1):124–141. Publisher: Institute of Mathematical Statistics.

Katzfuss, M., Guinness, J., Gong, W., and Zilber, D. (2020). Vecchia Approximations of Gaussian-Process Predictions. *Journal of Agricultural, Biological and Environmental Statistics*, 25(3):383–414.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets. *Journal of the American Statistical Association*, 103(484):1545–1555. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214508000000959.

Kaufman, C. G. and Shaby, B. A. (2013). The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, 100(2):473–484.

Kendrick, M. R. and McCord, J. W. (2023). Overwintering and breeding patterns of monarch butterflies (Danaus plexippus) in coastal plain habitats of the southeastern USA. *Scientific Reports*, 13(1):10438. Publisher: Nature Publishing Group.

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

Lamprinakou, S., Barahona, M., Flaxman, S., Filippi, S., Gandy, A., and McCoy, E. J. (2023). BART-based inference for Poisson processes. *Computational Statistics & Data Analysis*, 180:107658.

Lavancier, F. and Møller, J. (2016). Modelling Aggregation on the Large Scale and Regularity on the Small Scale in Spatial Point Pattern Datasets. *Scandinavian Journal of Statistics*, 43(2):587–609. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjos.12193.

Lee, A., Szpiro, A., Kim, S. Y., and Sheppard, L. (2015). Impact of preferential sampling on exposure prediction and health effect inference in the con-

text of air pollution epidemiology. *Environmetrics*, 26(4):255–267. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.2334.

Lee, D., Ferguson, C., and Scott, E. M. (2011). Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 174(1):109–126. Publisher: [Wiley, Royal Statistical Society].

León-Novelo, L. G. and Savitsky, T. D. (2019). Fully Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 13(1):1608–1645. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2011.00777.x.

Loader, C. (1999). Density Estimation. In Loader, C., editor, *Local Regression and Likelihood*, Statistics and Computing, pages 79–100. Springer, New York, NY.

Loh, W.-L. and Sun, S. (2023). Estimating the parameters of some common Gaussian random fields with nugget under fixed-domain asymptotics. *Bernoulli*, 29(3):2519–2543. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

Loh, W.-L., Sun, S., and Wen, J. (2021). On fixed-domain asymptotics, parameter estimation and isotropic Gaussian random fields with Matérn covariance functions. *The Annals of Statistics*, 49(6):3127–3152. Publisher: Institute of Mathematical Statistics.

Lu, C., Guan, Y., Lieshout, M. N. M. v., and Xu, G. (2024). XGBoostPP: Tree-based Estimation of Point Process Intensity Functions. arXiv:2401.17966 version: 1.

Manceur, A. M. and Kühn, I. (2014). Inferring model-based probability of occurrence from preferentially sampled data with uncertain absences using expert knowledge. *Methods in Ecology and Evolution*, 5(8):739–750. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12224.

Mardia, K. V. and Marshall, R. J. (1984). Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression. *Biometrika*, 71(1):135–146. Publisher: [Oxford University Press, Biometrika Trust].

Miller, M. J., Cabral, M. J., Dickey, E. C., LeBeau, J. M., and Reich, B. J. (2022). Accounting for Location Measurement Error in Imaging Data With Application to Atomic Resolution Images of Crystalline Materials. *Technometrics*, 64(1):103–113. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00401706.2021.1905070.

Moller, J. and Waagepetersen, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. CRC Press. Google-Books-ID: dBNOHvElXZ4C.

Momeni-Dehaghi, I., Bennett, J. R., Mitchell, G. W., Rytwinski, T., and Fahrig, L. (2021). Mapping the premigration distribution of eastern Monarch butterflies using community science data. *Ecology and Evolution*, 11(16):11275–11281. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.7912.

Moreira, G. A. and Gamerman, D. (2022). Analysis of presence-only data via exact Bayes, with model and effects identification. *The Annals of Applied Statistics*, 16(3):1848–1867. Publisher: Institute of Mathematical Statistics.

Moreira, G. A., Menezes, R., and Wise, L. (2023). Presence-Only for Marked Point Process Under Preferential Sampling. *Journal of Agricultural, Biological and Environmental Statistics*.

Paci, L., Gelfand, A. E., Beamonte, a. M. A., Gargallo, P., and Salvador, M. (2020). Spatial hedonic modelling adjusted for preferential sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1):169–192. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12489.

Pacifici, K., Reich, B. J., Dorazio, R. M., and Conroy, M. J. (2016). Occupancy estimation for rare species using a spatially-adaptive sampling design. *Methods in Ecology and Evolution*, 7(3):285–293. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12499.

Pati, D., Reich, B. J., and Dunson, D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, 98(1):35–48.

Pennino, M. G., Paradinas, I., Illian, J. B., Muñoz, F., Bellido, J. M., López-Quílez, A., and Conesa, D. (2019). Accounting for preferential sampling in species distribution models. *Ecology and Evolution*, 9(1):653–663. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.4789.

Putter, H. and Young, G. A. (2001). On the effect of covariance function estimation on the accuracy of kriging predictors. *Bernoulli*, 7(3):421–438. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

Reich, B. J. and Fuentes, M. (2012). Accounting for Design in the Analysis of Spatial Data. In Mateu, J. and Müller, W. G., editors, *Spatio-Temporal Design*, pages 131–141. John Wiley & Sons, Ltd, Chichester, UK.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12352.

Righetto, A. J., Faes, C., Vandendijck, Y., and Ribeiro, P. J. (2020). On the choice of the mesh for the analysis of geostatistical data using R-INLA. *Communications in Statistics - Theory and Methods*, 49(1):203–220. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610926.2018.1536209.

Rinaldi, L., Biggeri, A., Musella, V., Waal, T. d., Hertzberg, H., Mavrot, F., Torgerson, P. R., Selemetas, N., Coll, T., Bosco, A., Grisotto, L., Cringoli, G., and Catelan, D. (2015). Sheep and Fasciola hepatica in Europe: the GLOWORM experience. *Geospatial Health*, 9(2):309–317. Number: 2.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2008.00700.x.

Savitsky, T. D. and Toth, D. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10(1).

Schlather, M., Ribeiro Jr, P. J., and Diggle, P. J. (2004). Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):79–93. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1369-7412.2003.05343.x.

Schliep, E. M., Wikle, C. K., and Daw, R. (2023). Correcting for informative sampling in spatial covariance estimation and kriging predictions. *Journal of Geographical Systems*.

Schoenberg, F. P. (2005). Consistent parametric estimation of the intensity of a spatial–temporal point process. *Journal of Statistical Planning and Inference*, 128(1):79–93.

Scott, D. (1992). Kernel Density Estimators. In *Multivariate Density Estimation*, pages 125–193. John Wiley & Sons, Ltd. Section: 6 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470316849.ch6.

Shaddick, G. and Zidek, J. V. (2012). Preferential sampling in long term monitoring of air pollution: a case study. *TECHNICAL REPORT*, page 28.

Shirota, S. and Gelfand, A. E. (2022). Preferential sampling for bivariate spatial data. *Spatial Statistics*, 51:100674.

Sicacha-Parada, J., Steinsland, I., Cretois, B., and Borgelt, J. (2021). Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. *Spatial Statistics*, 42:100446.

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

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1):1–28. Publisher: Institute of Mathematical Statistics.

Stein, M. (1990a). Uniform Asymptotic Optimality of Linear Predictions of a Random Field Using an Incorrect Second-Order Structure. *The Annals of Statistics*, 18(2):850–872. Publisher: Institute of Mathematical Statistics.

Stein, M. L. (1988). Asymptotically Efficient Prediction of a Random Field with a Misspecified Covariance Function. *The Annals of Statistics*, 16(1):55–63. Publisher: Institute of Mathematical Statistics.

Stein, M. L. (1990b). Bounds on the Efficiency of Linear Predictions Using an Incorrect Covariance Function. *The Annals of Statistics*, 18(3):1116–1138. Publisher: Institute of Mathematical Statistics.

Stein, M. L. (1993). A simple condition for asymptotic optimality of linear predictions of random fields. *Statistics & Probability Letters*, 17(5):399–404.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer New York, New York, NY.

Tang, B., Clark, J. S., and Gelfand, A. E. (2021a). Modeling spatially biased citizen science effort through the eBird database. *Environmental and Ecological Statistics*, 28(3):609–630.

Tang, W., Zhang, L., and Banerjee, S. (2021b). On Identifiability and Consistency of The Nugget in Gaussian Spatial Process Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):1044–1070.

Totland, J. O. Z. (2023). *Geostatistical Modeling under Positional Uncertainty*. PhD thesis, Norwegian University of Science and Technology.

Vaart, A. W. v. d. (2000). *Asymptotic Statistics*. Cambridge University Press. Google-Books-ID: UEuQEM5RjWgC.

van Lieshout, M. (2021). Infill asymptotics for adaptive kernel estimators of spatial intensity. *Australian & New Zealand Journal of Statistics*, 63(1):159–181. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/anzs.12319.

Varin, C., Reid, N., and Firth, D. (2011). An Overview of Composite Likelihood Methods. *Statistica Sinica*, 21(1):5–42. Publisher: Institute of Statistical Science, Academia Sinica.

Vecchia, A. V. (1988). Estimation and Model Identification for Continuous Spatial Processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):297–312. Publisher: [Royal Statistical Society, Wiley].

Vedensky, D., Parker, P. A., and Holan, S. H. (2023). A Look into the Problem of Preferential Sampling through the Lens of Survey Statistics. *The American Statistician*, 77(3):313–322. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2022.2143898.

Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(4):595–601. Publisher: Institute of Mathematical Statistics.

Wang, D. and Loh, W.-L. (2011). On fixed-domain asymptotics and covariance tapering in Gaussian random field models. *Electronic Journal of Statistics*, 5(none):238–269. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Wang, W., Tuo, R., and Jeff Wu, C. F. (2020). On Prediction Properties of Kriging: Uniform Error Bounds and Robustness. *Journal of the American Statistical Association*, 115(530):920–930. Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2019.1598868.

Wang, W., Yue, X., Haaland, B., and Jeff Wu, C. F. (2022). Gaussian Processes with Input Location Error and Applications to the Composite Parts Assembly Process. *SIAM/ASA Journal on Uncertainty Quantification*, 10(2):619–650. Publisher: Society for Industrial and Applied Mathematics.

Warren, J. L., Perez-Heydrich, C., Burgert, C. R., and Emch, M. E. (2016). Influence of Demographic and Health Survey Point Displacements on Distance-Based Analyses. *Spatial demography*, 4(2):155–173.

Watson, J. (2021). A perceptron for detecting the preferential sampling of locations and times chosen to monitor a spatio-temporal process. *Spatial Statistics*, 43:100500.

Watson, J., Zidek, J. V., and Shaddick, G. (2019). A general theory for preferential sampling in environmental networks. *The Annals of Applied Statistics*, 13(4):2662–2700. Publisher: Institute of Mathematical Statistics.

Watson, S. I., Lilford, R. J., Sun, J., and Bion, J. (2021). Estimating the Effect of Health Service Delivery Interventions on Patient Length of Stay: A Bayesian Survival Analysis Approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(5):1164–1186.

Wilson, K. and Wakefield, J. (2021). Estimation of health and demographic indicators with incomplete geographic information. *Spatial and Spatio-temporal Epidemiology*, 37:100421.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, 36(2):280–296.

Yu, N. (2022). *Parametric Estimation in Spatial Regression Models*. PhD thesis, University of Maryland, College Park.

Yuan, Y., Bachl, F. E., Lindgren, F., Borchers, D. L., Illian, J. B., Buckland, S. T., Rue, H., and Gerrodette, T. (2017). Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *The Annals of Applied Statistics*, 11(4).

Zammit-Mangion, A., Ng, T. L. J., Vu, Q., and Filippone, M. (2022). Deep Compositional Spatial Models. *Journal of the American Statistical Association*, 117(540):1787–1808. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2021.1887741.

Zandbergen, P. A. (2008). Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy. *Transactions in GIS*, 12(1):103–130. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9671.2008.01088.x.

Zhang, H. (2004). Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

Zhang, H. and Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, 92(4):921–936.

Zhang, L., Tang, W., and Banerjee, S. (2024). Fixed-Domain Asymptotics Under Vecchia's Approximation of Spatial Process Likelihoods. *Statistica Sinica*, 34(4):1863–1881.

Zidek, J. V., Shaddick, G., and Taylor, C. G. (2014). Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. *The Annals of Applied Statistics*, 8(3):1640–1670. Publisher: Institute of Mathematical Statistics.

Zimmerman, D. (2007). Statistical Methods for Incompletely and Incorrectly Geocoded Cancer Data. In *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice*, pages 165–180. Journal Abbreviation: Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice.