SAND2023-06233T

ADVANCEMENTS IN ENVIRONMENTAL STATISTICS CONCERNING MULTIPLE DATA SOURCES

BY

MAURICIO CAMPOS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Doctoral Committee:

Professor Bo Li, Chair
Professor Douglas G. Simpson
Clinical Associate Professor Trevor H. Park
Clinical Associate Professor Lelys Bravo de Guenni

Abstract

This work considers three different applications of environmental spatial statistics. Two of them relate to integrating data coming from different sources, while the other does the opposite and instead breaks down an aggregated response into several components of interest. The contributions of this work are separated by application into three parts.

The first part is motivated by the interest in studying the biotic responses of species during the Last Glacial Maximum (LGM) due to rapid anthropogenic climate change. During this period, species retreated to highly spatially restricted geographic regions where survival was possible, known as glacial micro-refugia, from which they migrated and expanded when conditions became more suitable. Several distinct sources of evidence have contributed to developing a new understanding of how these regions might have impacted the sustainability of the natural populations of many species. Pollen records in Eastern Beringia (EB) have been used to explore the possibility that the region harbored glacial refugia for several plants from the arctic tundra and/or the boreal forest biomes common to the region. Our study focuses on *Alnus viridis* and *Picea glauca*, two predominant species of arcto-boreal vegetation. We propose to integrate genomic, SDM, and existing fossil data in a hierarchical Bayesian modeling (HBM) framework to determine whether multiple refugia existed in isolated geographic areas. This study demonstrates how the flexibility of HBMs makes the formal synthesis of such disparate data sources feasible. Our results highlight the regions of plausible refugia that can guide future investigations into studying the role of glacial refugia during climate change.

The second part reverses the data integration of the first in hopes of utilizing present technologies better for the purposes of crop monitoring. The amount of carbon assimilated by plants through photosynthesis, called Gross Primary Productivity (GPP), is the largest carbon flux between the

terrestrial biosphere and the atmosphere and, if quantified accurately, can grant insights into understanding several ecosystem functions as well as the impact of climate change to crop yields, in particular corn and soy. Recently, satellite-based measurements of solar-induced chlorophyll fluorescence (SIF) have been used as a strong proxy to measure GPP. SIF values will depend on the type of vegetation land cover; thus, the observed values can be decomposed into the specific vegetation type components to obtain its particular SIF yield information. We propose to implement a spatially varying coefficient regression model where the coefficients represent the specific SIF yields. For each land type coefficient, we induce spatial smoothness by penalizing the square deviations among adjacent sites according to some data-driven threshold value. The adjacent sites are chosen according to a minimum spanning tree (MST) in order to reduce redundancy in site pairing. Special characteristics of the data impose additional challenges, such as a non-negativity constraint on the estimations, as well as the presence of deterministic information that changes the structure of the MST. This study is able to retrieve accurate and fast results for the two main crops of interest when compared to other similar methods.

Finally, the third part returns to data integration in handling the change of support problem in spatial statistics. Environmental applications are highly dependent on accurate and complete datasets of world temperature. However, the recollection of these datasets is limited by technology and poses additional challenges that must be handled before doing any environmental analysis. In particular, the change of support problem occurs when trying to combine data that has been collected at different resolutions. Some data is collected in situ, whereas other can be collected by satellites that cover wider areas. This study uses INLA to accurately handle data coming from the Integrated Global Radiosonde Archive (IGRA) and from the TIROS Operational Vertical Sounder (TOVS), both from the National Oceanic and Atmospheric Administration (NOAA), from 1990 to 1993. Both datasets are considered to measure the same latent process but in different ways that must be integrated to produce a complete picture of global temperature. This must also be done close to real-time, so an alternative model is also proposed to speed computations at the cost of accuracy. We are able to obtain similar results from both approaches as well as provide accurate uncertainty measurements for both. These results can then be used for future applications of environmental studies.

To my parents. Thank you for everything.

Acknowledgments

I want to first thank the teachers who made this possible. In particular, I want to thank my advisor, Bo Li, who contributed tremendous time and effort to guide me throughout this whole process. I am fortunate to have had Bo push me to become a better statistician in this time.

I would also like to thank all the collaborators with whom I've worked in all my different project. I thank Guillaume de Lafontaine, Joe Napier and Feng Sheng Hu in helping me understand the details about different paleoecological sources of evidence. I thank Kaiyu Guan and his team in providing several references that helped understand the relation between SIF and different vegetation types. I also want to thank Audrey McCombs, Justin Li, Gabriel Huerta and Lyndsay Shand from Sandia National Laboratories for welcoming me into their time and give me a chance to work alongside them.

I also thank my parents, Raul and Liliana, for always encouraging and supporting me, and for being great parents overall. I also thank the friends that have stuck with me since the beginning and those that I have made along the way, both in Urbana-Champaign and Costa Rica, for their support. I wouldn't have made it this far without the good times we had.

Table of contents

Chapter 1

Introduction

In an era of unprecedented environmental challenges and growing concerns about the sustainability of our planet, the need for reliable data and rigorous analysis has become paramount. Environmental statistics, as a specialized field within the broader realm of statistics, plays a pivotal role in providing a systematic framework for understanding and addressing complex environmental issues. This dissertation aims to delve into the realm of environmental statistics, exploring its inherent importance, methodologies, and practical applications in addressing contemporary environmental challenges.

The field of environmental statistics encompasses a diverse range of quantitative techniques and methodologies that facilitate the collection, analysis, and interpretation of data related to various environmental phenomena. Its primary objective is to generate meaningful insights and inform decision-making processes pertaining to environmental management, policy formulation, and resource allocation.

Furthermore, environmental statistics serves as a crucial tool for decision-making under uncertainty. Environmental systems are inherently complex and subject to numerous interdependencies, making accurate predictions and risk assessments challenging. By utilizing statistical models and techniques, practitioners can evaluate the potential impacts of various environmental factors, predict future trends, and assess the associated uncertainties. Such analyses provide decision-makers with a clearer understanding of the potential consequences of alternative courses of action and enable them to make informed choices that minimize negative environmental outcomes.

The first application investigated in this work involves the integration of three distinct sources

of evidence to identify glacial refugia for two species of arctic shrubs during the Last Glacial Maximum. Glacial refugia are small pockets of land with sustainable conditions that allowed certain species to survive in northern regions during the harsh winters and ice sheet expansion. To identify potential refugia locations, this research combines evidence from pollen fossil records, genetic history of current vegetation, and mathematical species distribution models. A Bayesian hierarchical model is employed to relate all data sources to a common latent process. Estimation is conducted using integrated nested Laplace approximations (INLA), a computationally efficient alternative to traditional Markov chain Monte Carlo methods. The results provide researchers with new locations to direct their future studies in the search for additional evidence of glacial refugia.

The second application focuses on improving crop-monitoring methodologies using satellite spectroscopy and solar induced chlorophyll fluorescence (SIF). SIF, a byproduct of photosynthesis, can enhance crop yield estimations for a given season. However, the computational time required for SIF decomposition, which relates total SIF measured by satellites to specific SIF from different vegetation types, currently poses a challenge. This research builds upon existing methods to optimize the decomposition process, reducing computation time without sacrificing accuracy. By enhancing the efficiency of SIF decomposition, this approach contributes to improving food security and addressing concerns regarding the sustainability of our planet.

The third application addresses the change of support problem, a classical issue in spatial statistics, which arises when evidence is measured at different resolutions and needs to be combined while considering varying uncertainties. Specifically, this research aims to interpolate global temperature in the early 90s using satellite measurements conducted in areal blocks and local observations from weather balloons. The flexibility offered by INLA is explored to account for the distinct nature of these data sets. However, due to the computational demands associated with the resolution of satellite images, an alternative Bayesian hierarchical model is proposed to expedite estimations while maintaining consideration for the different uncertainties from each data source. Although the resulting estimation of the field is slightly less accurate, it can be obtained within a more realistic timeframe, facilitating comprehensive analysis of real-world data and the examination of global temperature trends over the years.

In conclusion, this dissertation's exploration of environmental statistics applications involving multisource data analysis showcases the discipline's significance in addressing complex environ-

mental challenges. By integrating diverse sources of evidence and employing advanced statistical methodologies, researchers can gain valuable insights into glacial refugia, improve crop-monitoring techniques, and tackle the change of support problem in spatial statistics.

Ultimately, the findings of this research will contribute to a better understanding of environmental dynamics, inform evidence-based decision-making, and foster sustainable development practices. Environmental statistics is a discipline that bridges the gap between statistical methodologies and environmental sciences. By acknowledging the importance of this field and further advancing its methodologies, we can equip ourselves with the necessary tools to address pressing environmental challenges and safeguard the future of our planet.

Chapter 2

Integrating Different Data Sources Using a Bayesian Hierarchical Model to Unveil Glacial Refugia

2.1 Introduction

Anthropogenic climate change has become a major concern for the sustainability of the natural populations of many species. This has renewed interest in understanding the biotic responses to climate variations in the paleorecord, because such understanding will be essential in anticipating future changes in biodiversity and informing ecosystem management (e.g., Dawson et al. 2011). To shed light on this issue, we particularly study the species range shifts within the Quaternary from the Pleistocene during the Last Glacial Maximum (LGM) (colloquially referred to as the *Ice Age*) to the current-day Holocene. During this period, the varying climates had a major impact on altering the biodiversity patterns of the region, and thus understanding shifts in species distribution during this period offers much evidence of the species response to climate change (Davis and Shaw 2001; Lafontaine et al. 2018; Napier, Lafontaine, and Chipman 2020).

During periods of atypical regional climate, species retreated to geographic regions where survival was possible, known as glacial refugia, from which they migrated and expanded when conditions became more suitable (Hampe and Jump 2011; Keppel et al. 2012; Gavin et al. 2014). Recent genetic studies (e.g. Anderson et al. 2006; Parducci et al. 2012; De Lafontaine et al. 2013; Hao

et al. 2018; Napier et al. 2019; Napier et al. 2020) have demonstrated the possibility that many arcto-boreal plants survived the LGM in small disjunct populations that later expanded in the post-glacial period. These "cryptic refugia", usually undetected using the fossil record (Provan and Bennett 2008), challenge the traditional understanding regarding the role of low-latitude refugia in the post-glacial vegetation development (e.g. Petit et al. 2003; McLachlan, Clark, and Manos 2005; Magri et al. 2006; Stewart et al. 2010; Mosblech, Bush, and Woesik 2011). It was previously believed that most of the post-glacial colonization came from refugia located in warmer lower latitudes, but now high-latitude refugia offer another insight into the process of how species flourished into the Holocene (Feurdean et al. 2013). This is of particular importance since the existence of small refugial populations might contribute to explaining the "Quartenary conundrum" - there being little evidence of species extinction during the dramatic climate shifts of the Quaternary, as opposed to the massive extinctions predicted by our current climate change (Botkin et al. 2007) - thus creating forecasts that lessen the overestimation of extinction likelihood (Luoto and Heikkinen 2008; Randin et al. 2009; Mosblech, Bush, and Woesik 2011).

Evidence of the existence of cryptic refugia has risen from many regions in the northern hemisphere but our study will focus mostly on Eastern Beringia (Alaska and adjacent Canada), which has been featured extensively in the literature and recognized as a site of possible refugia (e.g. Shafer et al. 2010). The dense network of fossil pollen records recovered from lake sediments captured over several decades in the region has been used to examine the possibility that it harbored glacial refugia for arcto-boreal taxa (Hopkins, Smith, and Matthews 1981; Bigelow et al. 2003; Brubaker et al. 2005). Additionally, phylogeographic surveys have also contributed to these studies by analyzing DNA markers of extant populations (e.g., Abbott and Brochmann 2003; Anderson et al. 2006; Anderson, Hu, and Paige 2011; Lafontaine, Turgeon, and Payette 2010; Napier et al. 2019; Napier et al. 2020). Altogether, the evidence seems to indicate that several arcto-boreal species managed to persist through the LGM in Eastern Beringia. However, details about the whereabouts of such refugial populations remain unknown.

Much of the evidence used in uncovering the refugia comes from three data sources: pollen fossil records, phylogeographic surveys, and species distribution models (SDMs). Fossil pollen is recovered from lake-sediment cores. If enough pollen that dates back to the LGM is found in the cores, it would be direct proof of past presence in the vicinity of the coring site. As such it is likely the

most robust line of evidence, but recovering this information is a resource-intensive procedure thus we only have limited data collected over various decades. Phylogeography relies on analyzing the geographical pattern of DNA diversity from modern-day samples to infer the past evolutionary scenarios that generated the observed modern-day genetic lineages. Since it relies on sampling present-day individuals, genetic information is easier to obtain than pollen fossil records, at the expense that the inferences about past refugia are less direct. Finally, SDM is the association between known modern-day occurrence and climate variables that is projected on past climate reconstructions to obtain probabilities of suitable climate for a given species over the landscape. This provides insight into regions where climate conditions might have been suitable for the species to be present but provides no direct evidence of past presence.

A review of the literature in paleoecology has revealed that many different statistical techniques have been employed to recover refugia from each data source (Gavin et al. 2014). For example, analysis of fossil data typically consists of comparing, modern-day pollen assemblages with those observed in the past to infer the composition and location of ancient forests.

Phylogeography relies on analyzing geographical patterns of genetic diversity and structure from natural populations to infer the historical evolutionary processes that lead the distribution of past geographical genealogical lineages to the present distributions. Refugia locations can usually be identified due to their lower intrapopulation genetic diversity and higher interpopulation diversity, resulting in stronger genetic differentiation but lower spatial genetic structure, than biological populations located in recolonized areas (Hewitt 2000; De Lafontaine et al. 2013). Population genetics have employed different approaches to studying genetic variation used for phylogeographic inferences. For instance, techniques such as AMOVA (Analysis of Molecular Variance; Excoffier and Smouse 1994) have borrowed statistical methods to provide objective historical inferences. AMOVA aims to estimate population differences similar to the statistical analysis of variance (ANOVA) (Meirmans and Liu 2018). The total genetic variance is decomposed into three covariance components: between-population, between-individuals within a population, and within-individuals, which are then used to construct the test statistics similar to F-statistics (Meirmans and Liu 2018).

Lemmon and Lemmon 2008 used likelihood methods to both test a prior conjecture regarding refugia as well as estimate the phylogeographic history of a gene in the absence of such conjecture. A Bayesian alternative to the estimation methods has also been employed to model the locations of

taxa along each branch in the phylogeny (e.g. Lemey et al. 2009; Lemey et al. 2010; Manolopoulou and Emerson 2012; Marske, Leschen, and Buckley 2012). Due to the increasing complexity of the likelihood models for estimating ancestral refugia, it is common in the field to fit Bayesian models using Approximate Bayesian Computation methods (Gao et al. 2012; Li et al. 2013; Budde et al. 2013; Tsuda et al. 2016; Wang et al. 2016; Cornejo-Romero et al. 2017; Ren et al. 2017; Aoki et al. 2019).

Each line of evidence provides its own set of strengths and weaknesses. Different data sources also seem to capture different information regarding refugia and postglacial expansion. For example, evidence from fossil pollen records (e.g., Anderson and Brubaker 1994) implies that taxa, such as spruce, resided in one general area and expanded in a single direction during the postglacial. However, genetic analyses suggested the existence of multiple microrefugia in Eastern Beringia (Napier et al. 2019), consistent with the prevailing pattern that has also emerged in other regions around the globe (Hao et al. 2018). It is imperative to develop an integrative method that can jointly glean information from all lines of evidence. Several attempts have been made and most of them emphasized the integration of genetic data and SDM information to obtain better estimates of refugia location (see Section IV of Gavin et al. 2014). These methods typically use SDM as a filter for identifying plausible refugia locations over which multiple genetic scenarios are then simulated and compared to the observed genetic data with statistical tests (e.g. MANOVA/ANOVA) to determine which ones provide the most likely locations (e.g. Knowles and Alvarado-Serrano 2010; Brown and Knowles 2012; Espíndola et al. 2012; Aoki et al. 2019; Napier et al. 2019). Bayesian hierarchical models (BHMs) have also been used for this purpose as a foundation of dynamic geographical range models. These models combine abundance information (usually obtained from SDMs) with environmental data and demographic rates to estimate niches and range dynamics, which in turn inform the presence of refugia (Marion et al. 2012; Pagel and Schurr 2012; Schurr et al. 2012).

BHMs have been a popular approach for data fusion due to their advantage of enabling joint modeling while being flexible to take into account the unique characteristics of each data type (Clark 2005). In addition, the posteriors of BHMs naturally provide uncertainty quantification for the estimates of unknown variables. BHMs have shown great promise in paleoclimate and paleoecological studies (e.g. Li, Nychka, and Ammann 2010; Urban et al. 2013). Advances in computation power as well as alternatives to MCMC, such as INLA (Rue, Martino, and Chopin

2009), have made it possible for the estimation to be timely and efficient. To our knowledge, there is no systematic and rigorous method to combine all three major data sources to infer refugia locations. We propose to integrate species-distribution, genomic, and existing fossil data in a BHM framework to elucidate glacial refugia of green alder (*Alnus viridis*) and white spruce (*Picea glauca*) in Eastern Beringia. Our method allows for the strengths of one data source to compensate for the weaknesses of others. We hope that the uniqueness and strength of this proposed method make it a useful tool for paleoecology and enlighten new follow-up studies.

The rest of the chapter is organized as follows: Section 2.2 reviews the three distinct lines of evidence that are commonly used to locate the most possible arcto-boreal refugia in Eastern Beringia. Section 2.3 introduces the BHM that integrates all three lines. Section 2.3.2 contains a brief explanation of how the estimation procedure is implemented using Integrated Nested Laplace Approximation (INLA). A small simulation study verifying our method is shown in Section 2.4. Finally, Section 2.5 presents the results for both arcto-boreal species under study.

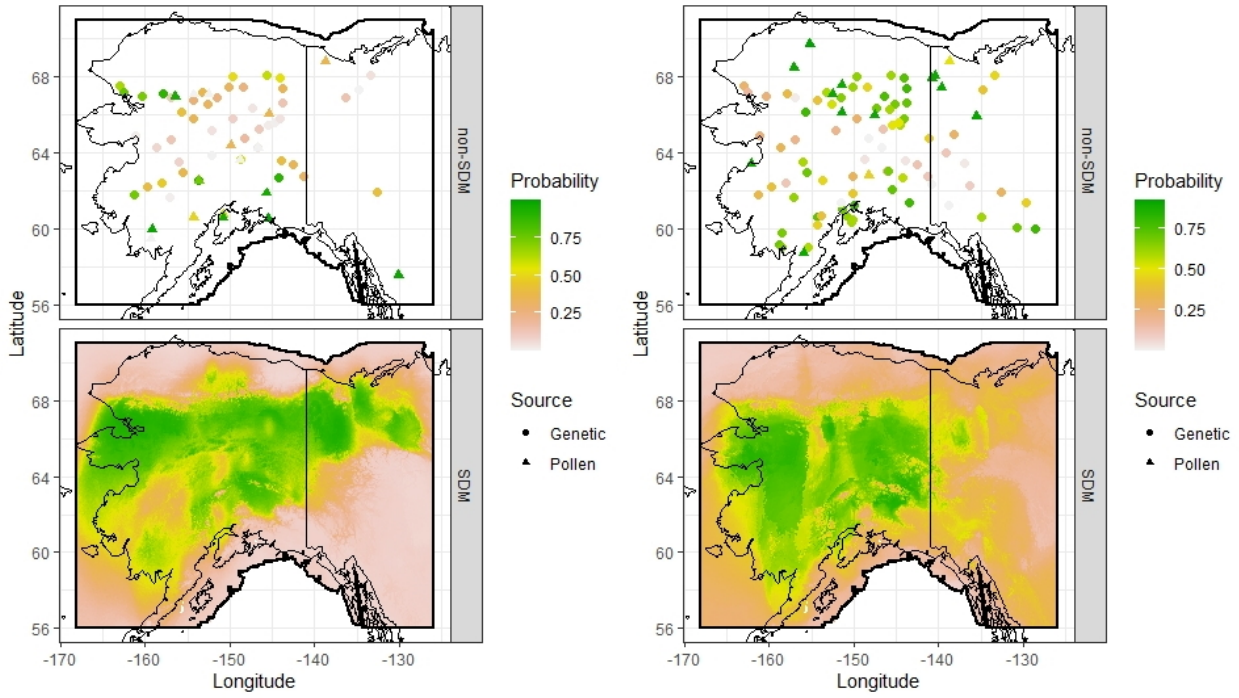
2.2 Data

Our data comes from three different sources: niche models, genetic lineages, and pollen fossil records. All three types of data are acquired for green alder (*Alnus viridis*) and white spruce (*Picea glauca*), and are shown in Figure 2.1. The particular type of niche modeling used in this chapter is Species Distribution Models (SDMs). The pollen and genetic data are much more sparse than the SDM.

2.2.1 Species distribution model

SDMs determine the probability of suitable climates using environmental variables (Franklin 2010) and have been widely applied due to their simplicity and growing accessibility (e.g., Thuiller et al. 2009). SDMs were developed based on available modern species occurrence and climate data and then applied to climate simulations to hindcast probabilities of species past occurrence. SDM for green alder is taken from Napier et al. 2019 whereas for white spruce it was generated using the same approach. The improved availability of paleoclimate simulations has led to the increasing application of SDMs to paleoecology (Nogués-Bravo 2009; Svenning et al. 2011), including the study of historical refugia. As the output of numerical models, SDM can be obtained at a very fine resolution, with around 334,000 sites in our study region.

However, several assumptions and uncertainties, such as the assumed static species-climate



(a) (b)

Figure 2.1: Observed data from all sources for (a) green alder and (b) white spruce. The upper panel in (a) and (b) shows the genetic and pollen data while the lower panel shows the species distribution model (SDM) data. The pollen and genetic data shown here are already processed with the interpolations discussed in Sections 2.2 and 2.3. The shaded region corresponds to Eastern Beringia during the Last Glacial Maximum, with modern-day Alaska superimposed for reference.

relationships despite changes in the environment, unaccounted putative dispersal limitations, and biotic interactions, limit the utility of SDMs and complicate their interpretation (Guisan and Thuiller 2005).

Thus SDMs are best viewed as a tool for preliminary analysis regarding the past locations (e.g., Porto, Carnaval, and Rocha 2013) and dynamics (e.g., Graham et al. 2010) of species that needs cross-validation with independent evidence such as genomic and other paleoecological data. Following Allouche, Tsoar, and Kadmon 2006, we consider SDM probabilities greater than a model-specified threshold τ_m (maximizing the accuracy of the model based on the True Skill Statistics) as an indicator for favorable conditions for refugia and otherwise unfavorable. For regions where SDM probabilities are below τ_m, it is safe to assume that those regions are not refugia. However, where SDM probability is above τ_m, we need complementary lines of evidence to better locate the exact refugia area. This characteristic of SDM makes SDM more appropriate for playing the role of classifying whether refugia are present or absent. Therefore, we transform SDM into binary data.

Let $P_m(\boldsymbol{s})$ represent the SDM probability that site $\boldsymbol{s}$ is a refugia location. We define $Y_m(\boldsymbol{s}) = I\{P_m(\boldsymbol{s}) \geq \tau_m\}$, where $I(A)$ is an indication function with $I(A) = 1$ if A is true and 0 otherwise. The threshold τ_m was chosen to be the True Skill Statistic (TSS) defined in Allouche, Tsoar, and Kadmon 2006. TSS is defined as TSS = sensitivity + specificity - 1, where the sensitivity and specificity are obtained by comparing the SDM predictions with a set of validation sites. Napier et al. 2019 suggest using 0.54 and 0.506 thresholds for green alder and white spruce, respectively.

2.2.2 Pollen data

The fossil data comes from pollen records that have been collected in Beringia since the early 1980s. The information used represents the effort over multiple decades of several research teams to uncover evidence of refugia and yet only a few of them can be used to posit our species of interest during the LGM. Coring sites that actually date back to the LGM are scarce and thus the spatial resolution of this database is coarse. The observed pollen data measures the proportion of pollen fossil records belonging to a specific species at a given depth of a sediment core. The greater this proportion, the stronger the evidence of the site being refugia. Despite this continuous association, the pollen data was mainly used as a binary indicator by thresholding the records.

We wish to utilize pollen data to a greater extent than as a binary variable indicating pres-

ence/absence. We will still respect that usually a site $\boldsymbol{s}$ is considered to be refugia of a species with probability $P_p(\boldsymbol{s}) \geq \gamma_p$ for a large γ_p, if its composition proportion $c(\boldsymbol{s}) \geq \tau_p$ for a species-specific threshold τ_p. Also, since there are many pollen types in the sediment samples, the composition percentages are usually small. Due to the small nature of these percentages, we consider $P_p(\boldsymbol{s}) = 1$ if $c(\boldsymbol{s}) = 0.5$. In the observed data, no composition percentage reaches this threshold. To use the pollen data properly, we propose to transform the composition proportions $c(\boldsymbol{s})$ into probabilities $P_p(\boldsymbol{s})$ subject to the above considerations:

$$p_p(\boldsymbol{s}) = \begin{cases} \frac{c(\boldsymbol{s})}{\tau_p}\gamma_p & \text{if } c(\boldsymbol{s}) \leq \tau_p \\ \{2c(\boldsymbol{s})\}^{\log(\gamma_p)/\log(2\tau_p)} & \text{if } c(\boldsymbol{s}) > \tau_p. \end{cases} \tag{2.1}$$

The transformation (2.1) features a linear interpolation of probabilities when $c(\boldsymbol{s}) \leq \tau_p$, and approaching probability 1 in polynomial when $c(\boldsymbol{s}) > \tau_p$, as shown in Figure 2.2. The coefficient 2 and the polynomial power are determined by the conditions $P_p(\boldsymbol{s}) = \gamma_p$ when $c(\boldsymbol{s}) = \tau_p$ and $P_p(\boldsymbol{s}) = 1$ when $c(\boldsymbol{s}) = 0.5$. There is no established literature with regard to the interpolations formula, and our proposed transformations were simply constructed for being sound choices and conforming to our prior knowledge. There could be other choices. For example, the linear interpolation for the range $c(\boldsymbol{s}) \leq \tau_p$ can be generalized to a polynomial interpolation with power r, i.e. $p_p(\boldsymbol{s}) = \left\{\frac{c(\boldsymbol{s})}{\tau_p}\right\}^r \gamma_p$ for $c(\boldsymbol{s}) \leq \tau_p$. Curves for different r are shown in Figure 2.2. We choose $r = 1$ for its simplicity and for representing more reasonable probabilities for small $c(\boldsymbol{s})$ than $r > 1$. We found our results are insensitive to other reasonable transformations.

The threshold τ_p depends on species: green alder uses $\tau_p = 2.5\%$ whereas white spruce uses 1% (Napier et al. 2019; Warren et al. 2016). Likewise, γ_p differs for both species. These were chosen as $\gamma_p = 0.95$ for alder and $\gamma_p = 0.90$ for spruce as sensible choices that represent 'high' probabilities.

2.2.3 Genetic evidence

We obtain genetic data from genetic surveys that report separate lineages (see Napier et al. 2019 for more detail). Genetic data for green alder consists of evidence from only two lineages, while white spruce has five different lineages. For a particular site $\boldsymbol{s}_i$, the ancestry coefficient $a_k(\boldsymbol{s}_i)$ is defined as the proportion of site i's genome that originated from lineage k (Pritchard, Stephens, and Donnelly 2000). This implies that $\sum_k a_k(\boldsymbol{s}_i) = 1$, where the summation is taken over all lineages

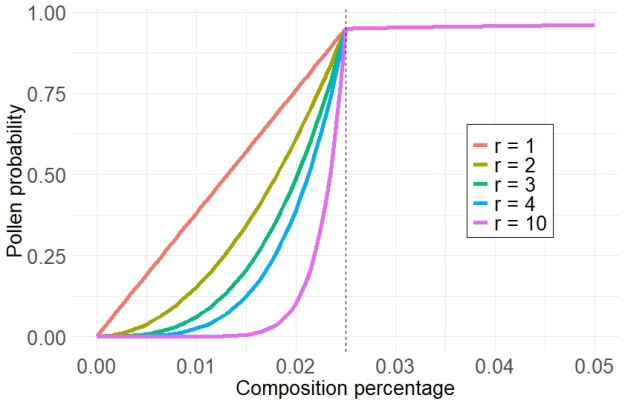


Figure 2.2: Different polynomial interpolations for pollen data. The $r = 1$ curve corresponds to the choice used in Equation (2.1). The dashed line represents τ_p and how all interpolations have the same tail afterward.

represented in the study for a particular species. Similar to pollen data, genetic information was traditionally used as a binary source of evidence. To use genetic data more efficiently, we likewise transform each genetic assemblage to a probability of being refugia.

The transformation is based on the relative percentages of different lineages for each species. Only the dominance of one single lineage indicates the higher chance of this location being a refugium. The transformation differs depending on species as each one has a different number of lineages, nevertheless, the underlying principle remains the same.

For both species, let $\tilde{A}(\boldsymbol{s}) = \max_k a_k(s)$ for site $\boldsymbol{s}$ and let $P_g(\boldsymbol{s})$ represent the probability of site $\boldsymbol{s}$ being refugia according to genetic data. Since green alder only has two lineages, the transformation should interpret a site with $\tilde{A}(\boldsymbol{s})$ closer to 0.5 corresponding to a smaller $P_g(\boldsymbol{s})$, and the probability should grow larger as $\tilde{A}(\boldsymbol{s})$ increases. To meet those requirements, we propose a transformation as a polynomial of power r:

$$P_g(\boldsymbol{s}) = \left\{2\left(\tilde{A}(\boldsymbol{s}) - 0.5\right)\right\}^r. \tag{2.2}$$

Using this method we obtain a "soft" threshold for the data, where we retain all information in the data but only a few sites receive high probabilities while the rest are much lower, reflecting the higher uncertainty of being refugia if the sample at that site is more mixed (See Figure 2.3a). The value of r can be chosen according to how well the transformed data conforms with expert knowledge or prior information. A sensitivity analysis for different r values shows that overall as r increases the higher $P_g(\boldsymbol{s})$ remain similar, although the bulk of $P_g(\boldsymbol{s})$ decreases. We choose $r = 3$

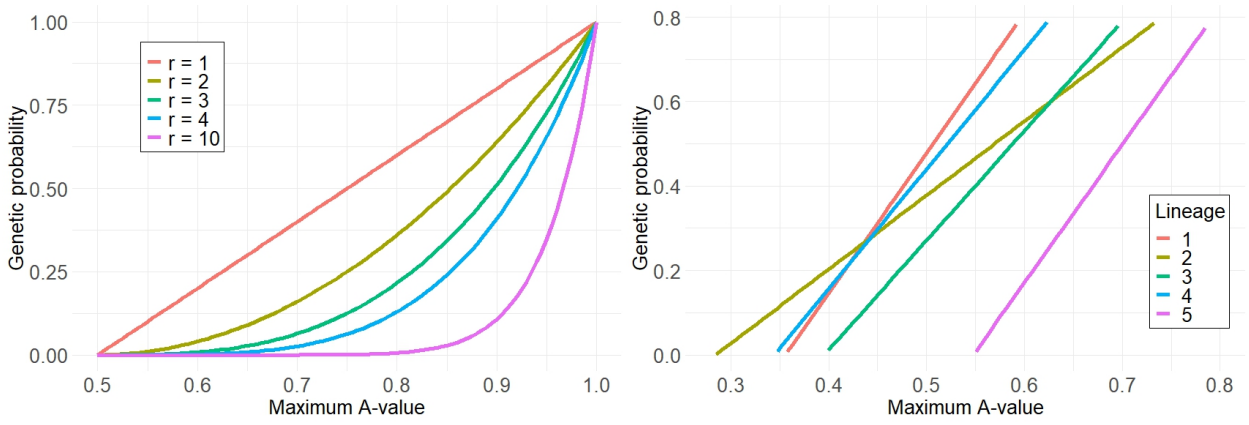
because this value seems to reach a better balance between keeping the sites with high $\tilde{A}(\boldsymbol{s})$ as high $P_g(\boldsymbol{s})$ and decreasing the rest to more conservative levels, according to expert knowledge.

White spruce has five different lineages. We first identify the dominant lineage for each site by finding which lineage corresponds to $\tilde{A}(\boldsymbol{s})$. Denote m_j as the total number of sites that have the j-th lineage as the dominant. Let $\tilde{A}_j(\boldsymbol{s}_i)$, $i \in \{1, 2, 3, ..., m_j\}$ represent the ancestry coefficient at the i-th location that is dominated by the j-th lineage. Let $\xi_j = \max_i \tilde{A}_j(\boldsymbol{s}_i)$ and $\delta_j = \min_i \tilde{A}_j(\boldsymbol{s}_i)$ be the maximum and minimum ancestry coefficient for lineage j, respectively. Also, let A_{max} and A_{min} represent the maximum and minimum ancestry coefficients observed among all sites and all lineages, which for our white spruce data are 0.789 and 0.0001 respectively. We define the probability of refugia for the genetic data:

$$P_g(\boldsymbol{s}) = \frac{A_{min}[\xi_j - \tilde{A}_j(\boldsymbol{s})] + A_{max}[\tilde{A}_j(\boldsymbol{s}) - \delta_j]}{\xi_j - \delta_j}. \tag{2.3}$$

With this definition, the lowest ancestry coefficient for each lineage will be assigned A_{min} as its probability of being refugia, whereas the largest lineage-specific coefficient will be assigned A_{max} as its corresponding probability (see Figure 2.3b), meaning that all lineages will have the same interpolated probability range. Note that the probability $P_g(\boldsymbol{s})$ can be interpreted as the weighted average of A_{min} and A_{max}, weighing by the distances from $\tilde{A}_j(\boldsymbol{s})$ to the extremes ξ_j and δ_j. In other words, the closer $\tilde{A}_j(\boldsymbol{s})$ is from its lineage's highest possible ancestry coefficient (i.e., ξ_j), the more the probability will approach A_{max}. Likewise, the closer $\tilde{A}_j(\boldsymbol{s})$ is from its lineage's lowest possible ancestry coefficient (i.e., δ_j), the more the probability will approach A_{min}. Since the values used in this interpolation method are the lineage-specific ancestry coefficients, they are naturally bounded by A_{min} and A_{max}, which represent the highest and lowest possible values of $P_g(\boldsymbol{s})$, respectively. This guarantees that none of the interpolated probabilities will be less than 0 or greater than 1.

Like pollen fossil data, the interpolations presented here are not unique and there is no established formula to follow. Other transformations could be considered, but we merely wish to translate the raw lineage information into more interpretable probabilities that can then be used in our model. The linear interpolation was chosen for its simplicity in interpretation and reasonable performance.



(a) (b)

Figure 2.3: Genetic data interpolations: (a) different polynomial curves are shown for green alder whereas (b) linear interpolations are shown for each genetic lineage of white spruce.

2.3 Bayesian Hierarchical Model

All three lines of evidence contain useful information in unveiling the refugia, though they each have their strength and weakness. We aim to integrate these complementary data sources to identify the possible locations of refugia, which is expected to be more efficient and powerful than using a single line of evidence. As discussed earlier, we have transformed the three data sources into $Y_m(\boldsymbol{s})$, $P_p(\boldsymbol{s})$, and $P_g(\boldsymbol{s})$, according to their characteristics. Our model is constructed based on the transformed data.

2.3.1 Model specification

Let $P(\boldsymbol{s})$ denote the probability of location $\boldsymbol{s}$ being a refugium. We attempt to obtain coherent estimates of $P(\boldsymbol{s})$, given the binary $Y_m(\boldsymbol{s})$ derived from the SDM and the probabilities $P_p(\boldsymbol{s})$ and $P_g(\boldsymbol{s})$ derived from pollen and genetic respectively. To accomplish this, we need to carefully model the relationship between $P(\boldsymbol{s})$ and the three data sources based on the characteristics of each data. Since all three data are observations from different perspectives given the true refugia, it is natural to establish the forward model of each data given a common underlying $P(\boldsymbol{s})$. The SDM has been mainly used as a preliminary screening tool through the binary $Y_m(\boldsymbol{s})$, hence we will employ a logistic model for $Y_m(\boldsymbol{s})$. Since the information from pollen and genetics is more quantitatively related to the probability of refugia, we build a model to reflect this feature. The forward models should also recognize that, when compared with temporally variable and spatially inconsistent pollen data, genetic data is often easier to obtain from comprehensive spatial grids.

Let S_m, S_g, S_p denote the collection of sites for SDM, genetic and pollen data, with sizes n_m, n_g, and n_p respectively. The total sample size is given by $n = |S| = |S_m \cup S_g \cup S_p|$ where $|S|$ denotes the cardinality of S. In our application, all data subsets are disjoint so that $n = n_m + n_g + n_p$. There is a strong unbalance in the sample sizes, such that $n_m \gg n_g + n_p$. Furthermore, we define two subregions for the SDM data: S_{m0} and S_{m1}, where S_{m1} is defined as the collection of SDM sites where SDM is greater than the threshold, i.e. $Y_m(\boldsymbol{s}) = 1$, and S_{m0} the rest. To lift the constraint of modeling probabilities, we first perform an inverse probability integral transform on $P(\boldsymbol{s})$, $P_p(\boldsymbol{s})$, and $P_g(\boldsymbol{s})$ using a standard normal cumulative distribution function $\Phi(\cdot)$ to turn probabilities into Gaussian random variables. Specifically, we have $\mu + X(\boldsymbol{s}) = \Phi^{-1}(P(\boldsymbol{s}))$, where $X(\boldsymbol{s})$ is assumed to be a mean zero Gaussian random variable, $Y_p(\boldsymbol{s}) = \Phi^{-1}(P_p(\boldsymbol{s}))$ and $Y_g(\boldsymbol{s}) = \Phi^{-1}(P_g(\boldsymbol{s}))$. Then we propose the following forward models as the first level of our BHM:

First level: Data models

$$\begin{aligned} \text{logit}\left(P\left(Y_m(s)=1\right)\right) &= \alpha_m + \beta_m\left\{\mu + X(\boldsymbol{s})\right\} + Z(\boldsymbol{s}), \quad \mathbf{Z} \sim GP\left(\mathbf{0}, \boldsymbol{\Sigma}(\sigma_m^2(\boldsymbol{s}), \rho_m)\right), \\ Y_p(\boldsymbol{s}) &= \alpha_p + \beta_p\left\{\mu + X(\boldsymbol{s})\right\} + \epsilon_p, \qquad \epsilon_p \sim N(0, \sigma_p^2), \\ Y_g(\boldsymbol{s}) &= \mu + X(\boldsymbol{s}) + \epsilon_g, \qquad \epsilon_g \sim N(0, \sigma_g^2), \end{aligned} \tag{2.4}$$

where $\mathbf{Z}$ is the vector consisting of all $Z(\boldsymbol{s})$ for $\boldsymbol{s} \in S_m$.

Our model respects the fact that all three data are trying to capture the true probability of refugia in different manners, albeit with uncertainties. Additionally, the model assumes that $Y_m(s), Y_p(\boldsymbol{s})$, and $Y_g(\boldsymbol{s})$ are conditionally independent given the latent process $X(\boldsymbol{s})$. Since genetics is considered the more spatially comprehensive quantitative data source, we model the genetic data as an unbiased source of the true refugia probability. SDM and pollen data are taken as deviations with both additive and multiplicative biases, in addition to Gaussian errors on the models. This model specification also ensures the identifiability of unknown parameters.

The Gaussian error process $Z(\boldsymbol{s})$ models the extra uncertainty in SDM data beyond what a Bernoulli distribution can capture. Considering different levels of credibility of SDM in showing whether there are refugia, we employ a non-stationary covariance function with unknown spatially varying variance, $\sigma_m^2(\boldsymbol{s})$, and an invariant range parameter, ρ_m for $Z(\boldsymbol{s})$. The covariance between

two sites can be expressed as

$$\text{cov}(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)) = \sigma_m(\boldsymbol{s}_i)\sigma_m(\boldsymbol{s}_j)C(||\boldsymbol{s}_i - \boldsymbol{s}_j||),$$

where $C(||\boldsymbol{s}_i - \boldsymbol{s}_j||)$ can be any valid correlation function. We choose the Matérn correlation function for $Z(\boldsymbol{s})$. Let $d = ||\boldsymbol{s}_i - \boldsymbol{s}_j||$, a Matérn correlation function is defined as

$$C(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{8\nu}}{\rho} d \right)^{\nu} K_\nu \left(\frac{\sqrt{8\nu}}{\rho} d \right), \tag{2.5}$$

where ν is the smoothness parameter, ρ represents the range and K_ν is the modified Bessel function of the second kind. The range parameter measures how quickly the spatial correlation decays with spatial distance and the smoothness parameter determines how smooth the random process is in terms of mean square differentiability (Stein 1999). We fix the smoothness parameter to $\nu = 1$ in our model due to the limitation of INLA computing algorithm (Bakka et al. 2018). Currently, R-INLA only allows values of $\nu \in (-1, 1]$ for spatial applications in two dimensions, where fields with negative values lack a point-wise interpretation (Lindgren and Rue 2015). Nevertheless, $\nu = 1$ is a reasonable choice for environmental processes and Whittle 1954 has argued that $\nu = 1$ is a more natural choice for two-dimensional processes than the exponential $\nu = 1/2$ alternative.

The spatially varying variance parameter, $\sigma_m(\boldsymbol{s})$, for $Z(\boldsymbol{s})$ switches between two regions: S_{m0} and S_{m1}. It is believed that evidence in S_{m0} is often more certain to show that this land was not suitable as refugia (e.g. evidence of ice sheets), compared to evidence in favor of S_{m1}. This is also why SDM data is often used as a classifier for refugia. Thus, we model SDM in the region S_{m0} with relatively smaller variance compared to S_{m1} as follows:

$$\log(\sigma_m(\boldsymbol{s})) = \theta_1 + \theta_2 \cdot I\{\boldsymbol{s} \in S_{m1}\},$$

where

$$I\{\boldsymbol{s} \in S_{m1}\} = \begin{cases} 1 & \text{if } \boldsymbol{s} \in S_{m1}, \\ 0 & \text{otherwise.} \end{cases}$$

In the second level of BHM, we model the unknown latent process $X(\boldsymbol{s})$ as a spatially correlated Gaussian process, and then we specify the priors in the third level to close the hierarchy.

Second level: Latent spatial process

$$X(\boldsymbol{s}) \sim GP\left(\mathbf{0}, \boldsymbol{\Sigma}(\sigma_x^2, \rho_x)\right). \tag{2.6}$$

To model the spatial correlation in $X(\boldsymbol{s})$, we again use a Matérn correlation function as defined in (2.5), with a range parameter ρ_x. Since there is no evidence to support that the variance of $X(\boldsymbol{s})$ should be spatially varying, we assume $X(\boldsymbol{s})$ is a stationary random process with a constant variance σ_x^2. We still fix the smoothness parameter to $\nu = 1$ for the reasons elaborated earlier, and treat the variance and range parameters, σ_x^2 and ρ_x, as unknown.

Third level: Priors

$$\begin{aligned}
\mu, \alpha_m, \alpha_p &\sim N(0, 1000), \\
\beta_m, \beta_p &\sim N(1, 1000), \\
\log(1/\sigma_p^2) &\sim \text{LogGamma}(1, 0.00005), \\
\sigma_g &\sim \text{PC Prior}, \\
(\sigma_x^2, \rho_x)^T &\sim \text{PC Prior}, \\
(\theta_1, \theta_2, \log \rho_m)^T &\sim N\left((0,1,0)^T, \boldsymbol{I}_3\right).
\end{aligned}$$

The *penalized complexity* (PC) prior was introduced in Simpson et al. 2017 to create a weakly informative prior that penalizes the complexity of a hierarchical model structure. These priors assign a non-zero mass to the simplest possible model, thus allowing the data to manifest itself freely when considering the necessity of including more parameters. Additionally, the PC priors have the following nice properties: invariant to reparameterizations, having a natural connection to Jeffreys' priors, supporting Occam's Razor, and are robust. They are also easily defined by the user, who only has to specify the tail probability of the prior, giving them more straightforward interpretability. The PC prior for σ_g is an exponential with rate determined by specifying the tail probability $P(\sigma_g > 1) = 0.01$. This prior form is determined by penalizing the distance (in terms of Kullback-Leibler divergence) from a simple model with no nugget to that of a more complex model that includes one. This prior further assists in respecting genetic data as a more spatially consistent data source than pollen by giving σ_g^2 a smaller value a priori.

For the Matérn covariance parameters, the joint PC prior is specified by the following marginal tail probabilities: $P(\sigma_x > 3) = 0.01$ and $P(\rho_x < 1) = 0.01$. These tail probabilities are chosen

to reflect unlikely events so that the prior can be considered weakly informative. Additionally, it penalizes complexity by shrinking the range toward infinity and the marginal variance toward zero (Fuglstad et al. 2019). In our particular scenario, the joint prior can be expressed as the product of a marginal Inverse Weibull density for the range and another Exponential for its standard deviation.

After we obtain posterior samples of $X(\boldsymbol{s})$ and μ, we apply the probability integral transform to derive the probability of refugia, $P(\boldsymbol{s}) = \Phi(\mu + X(\boldsymbol{s}))$. Then posterior inference is made on the samples of $P(\boldsymbol{s})$.

2.3.2 Estimation using INLA

A well-known bottleneck for large spatial data analysis is the computation of its likelihood. The number of sites in our data and of our interest makes computation a serious issue. This restricts the usage of traditional Markov chain Monte Carlo (MCMC) sampling methods for our BHM.

To bypass the computational challenge, we resort to the Integrated Nested Laplace Approximation (INLA) (Rue, Martino, and Chopin 2009) to derive the posterior densities. INLA has been a popular strategy for Bayesian estimation for large spatial random fields, by employing approximate Bayesian inference for latent Gaussian models controlled by a small number of hyperparameters. Using integrated nested Laplace approximations, INLA can obtain fast and accurate posterior estimates compared to MCMC (Rue, Martino, and Chopin 2009; Lindgren and Rue 2015).

INLA assumes that the latent field follows a Gaussian Markov random field (GMRF) with a sparse precision matrix, which allows for faster computations of the approximations and integrals. However, this becomes a limitation when modeling continuously indexed spatial fields.

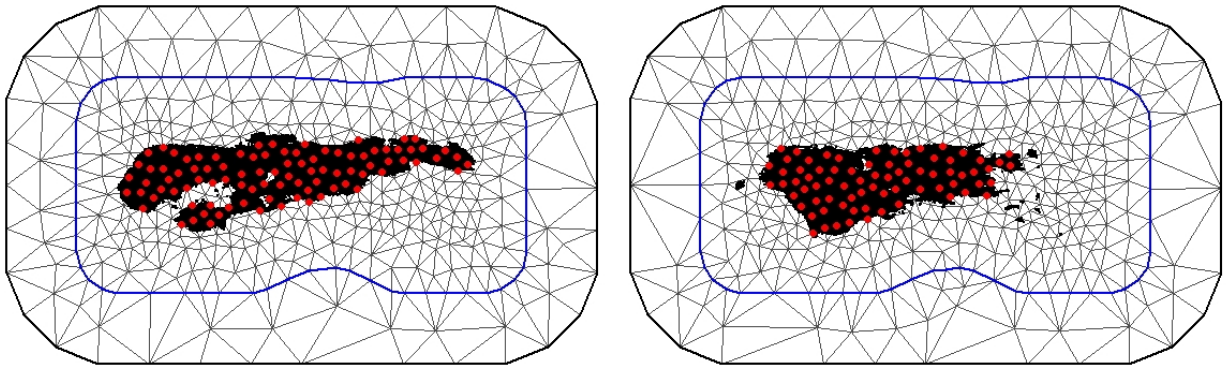
Nevertheless, Lindgren, Rue, and Lindström 2011 showed that an approximate stochastic weak solution to a linear stochastic partial differential equation (SPDE) will provide a Gaussian random field (GRF) with a Matérn covariance function, defined by the parameters of the SPDE. This means that modeling can be done in continuous space using GRFs, but the inference will gain the computational speed obtainable from working with sparse precision matrices on GMRFs formulated on a triangulation of the spatial domain.

R-INLA employs a Delaunay triangulation mesh (see Figure 2.4) where each vertice corresponds to a point where the GMRF is fitted. Following the approach from Lindgren, Rue, and Lindström 2011, the solution to the SPDE is approximated by a finite sum of basis functions, giving a

continuously indexed approximation of the Gaussian random field. Any point in a triangle is approximated by a linear interpolation of the basis functions used at each node. To improve the mesh construction, the border of the study region is used to delineate small and big triangles, with the smaller, more regular ones inside. This increases variability near the boundaries which helps mitigate the boundary effect of the estimations (Lindgren and Rue 2015; Bakka et al. 2018).

Having to define the spatial model in the discrete field means that the special regions for SDM, S_{m0} and S_{m1}, must also be defined in the triangular mesh. For that purpose, we must identify the nodes of the mesh that are contained in said regions. This is done by defining a radius around each node and counting the proportion of SDM sites that belong to S_{m1}. If the proportion is above a certain threshold value, then we count the node as being part of S_{m1}. Both the radius and threshold are tailored by the user to achieve reasonable results. For both species, a radius of 0.7 and a threshold proportion of 0.25 were used to define the S_{m1} mesh nodes (see the red dots in Figure 2.4).

The R-INLA package, obtained from www.r-inla.org, was used to run the INLA method for the Bayesian inference (Lindgren and Rue 2015).



(a) (b)

Figure 2.4: Constrained refined Delaunay triangulation mesh for (a) green alder and (b) white spruce. The black regions represent S_{m0} and the red dots represent the node points that correspond to this region.

2.4 Simulation Study

We conduct a small simulation study to verify that our method recovers the underlying refugia probability $P(\boldsymbol{s})$ if the data follow the models outlined in (2.4). We also evaluate the sensitivity of

our estimates to the number of SDM sites as we will use only a subset of the very dense SDM data in our real data analysis.

Sampling Scenario	n_m	n_p	n_g
S1	500	15	60
S2	1000	15	60
S3	1000	1000	1000

Table 2.1: Different sampling scenarios for SDM (n_m), pollen (n_p) and genetic (n_g) data

2.4.1 Setup

To mimic the real data, we adopt Eastern Beringia as our spatial domain and randomly select 60 locations for genetic data and 15 locations for pollen data. We randomly choose 500 SDM locations which are much sparser than the SDM data we have, for ease of computation. We also consider the scenario of 1000 SDM locations to evaluate the sensitivity of our estimation to the amount of SDM data used in the model. Additionally, we consider a third scenario with 1000 locations for each of the genetic, pollen, and SDM data, to study the effect of the sparsity of genetic and pollen data on the refugia probability estimation. These three scenarios are summarized in Table 2.1. In addition to the sites with observations, we also randomly sample 200 locations that do not overlap with the data sites and will be used for evaluating the probability estimation. For each scenario, we run the simulation 50 times.

We first simulate the $X(\boldsymbol{s})$ process by following the latent process model (2.6), and then generate $Y_p(\boldsymbol{s})$ and $Y_g(\boldsymbol{s})$ following their corresponding models in (2.4). The parameter values in the simulation were chosen to resemble those obtained in the real data application. The mean of the latent fields was chosen as $\mu = -0.5$ since overall most of the data indicates low probabilities. However, since there are still regions of high probability, this had to be reflected in the variance and range parameter of the Matérn covariance function of $X(\boldsymbol{s})$, which were chosen as 0.64 and 3, respectively. Additionally, to respect the signal-to-noise ratio of our data, the variance terms σ_p^2 and σ_g^2 were both set at 0.2. Additionally, pollen's additive and multiplicative bias were set at $\alpha_p = -0.8$ and $\beta_p = 1.4$, respectively. This indicates that overall, pollen sites have a lower probability than genetic but are more linearly related to $X(\boldsymbol{s})$.

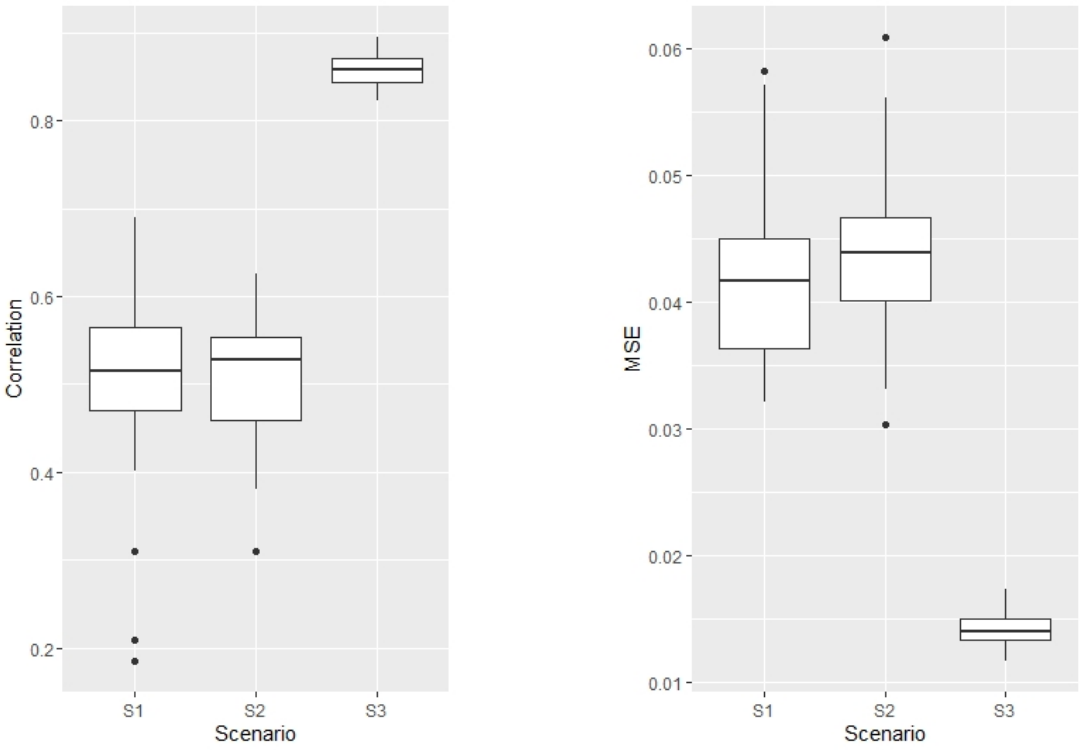
To simulate the SDM sites, we started by setting the additive bias at $\alpha_m = -10$, meaning that most of our SDM sites were going to provide evidence of no refugia (i.e. $Y_m(\boldsymbol{s}) = 0$). The multiplicative bias was left at $\beta_m = 1$. The range parameter of the covariance function was set at 10, whereas the variance at S_{m0} was 100, whereas at S_{m1} it was 144. The high variances allowed for the presence of only a few close areas of high probability to be present in what was mostly a low probability field, thus resembling the SDM data present in Figure 2.1.

The SDM data generation requires simulating the spatial error $Z(\boldsymbol{s})$ that has different variances depending on the region. To accomplish that, we first simulate a $Z(\boldsymbol{s})$ process with fixed variance $\sigma_m^2(S_{m0})$ for all n_m sites, based on which we calculate a preliminary $P\left(Y_m(s) = 1\right)$ using the logit model in (2.4). For a given τ_m, all sites such that $P\left(Y_m(s) = 1\right) > \tau_m$ have their $Z(\boldsymbol{s})$ multiplied by a correction factor to switch their variance to $\sigma_m^2(S_{m1})$. We recalculate $P\left(Y_m(s) = 1\right)$ and finally, the $Y_m(\boldsymbol{s})$ process is generated based on these updated probabilities and the threshold τ_m.

2.4.2 Comparing estimated and true probabilities

To obtain estimates of $P(s)$, for a given site $\boldsymbol{s}$, we generate posterior samples of $P(\boldsymbol{s}) = \Phi(\mu + X(\boldsymbol{s}))$ and take the posterior mean as our estimate $\hat{P}(\boldsymbol{s})$. We estimate $P(\boldsymbol{s})$ at all sites with observations and the extra 200 data-absent locations. We calculate the typical mean squared error (MSE) of $\hat{P}(\boldsymbol{s})$ as one measure of the estimation performance. Since in our data application, identification of refugia is based on the relative size of the $\hat{P}(\boldsymbol{s})$, capturing the spatially varying pattern is the key to correctly differentiating refugia and non-refugia areas. For this reason, we also use the Pearson correlation between $\hat{P}(\boldsymbol{s})$ and $P(\boldsymbol{s})$ to measure the performance of $\hat{P}(\boldsymbol{s})$.

Figure 2.5 presents the simulation results summarized over all locations. Scenario 1 (S1) and 2 (S2) are comparable in their correlation and MSE, with S2 performing slightly worse in terms of MSE but slightly better in correlation. This indicates that the probability estimation is insensitive to the number of SDM sites, due to the binary feature of SDM data. This supports our choice of using a random subset of all available SDM sites in the real data analysis. However, since genetic and pollen data are more informative, the abundance of these data can have a significant impact on the estimation. We observe low correlations between the estimated and true probabilities, around 0.5, for Scenarios S1 and S2, indicating the struggle of estimates with scarce genetic and pollen data. Results for S3 show that with dense genetic and pollen data, the capacity of our model for



(a) (b)

Figure 2.5: Simulation results showing (a) correlation and (b) MSE between the estimated probabilities, $\hat{P}(\boldsymbol{s})$, and true probabilities, $P(\boldsymbol{s})$, for the three distinct sampling scenarios, S1, S2, and S3.

estimating the refugia probability is much improved.

Figure 2.7 compares the estimated and the true refugia probability of one particular simulation run for both S2 and S3. The particular run was chosen to correspond to the median correlation between $\hat{P}(\boldsymbol{s})$ and $P(\boldsymbol{s})$ for their respective scenario. With fewer genetic and pollen data in S2, the estimation show a blurry version of the true probability. Nevertheless, the estimation still captures a rough pattern of high and low probabilities. With the ideal situation of S3 that has dense genetic and pollen data, the estimated refugia probabilities recover considerable amount of details of the true probabilities. Although it is unclear how to exactly interpret the empirical coverage of credible intervals, we report the coverage for $P(\boldsymbol{s})$ in the Figure 2.6. The empirical coverage for S2, which resembles the real data, is lower than 95% by 10%. This might imply our uncertainty estimate for the real data could be somewhat lower than it should be. Note that even if our primary interest, $P(\boldsymbol{s})$, can be estimated well, we find that the estimation of nuisance parameters of the complex system, such as some variance parameters, could still carry bias.

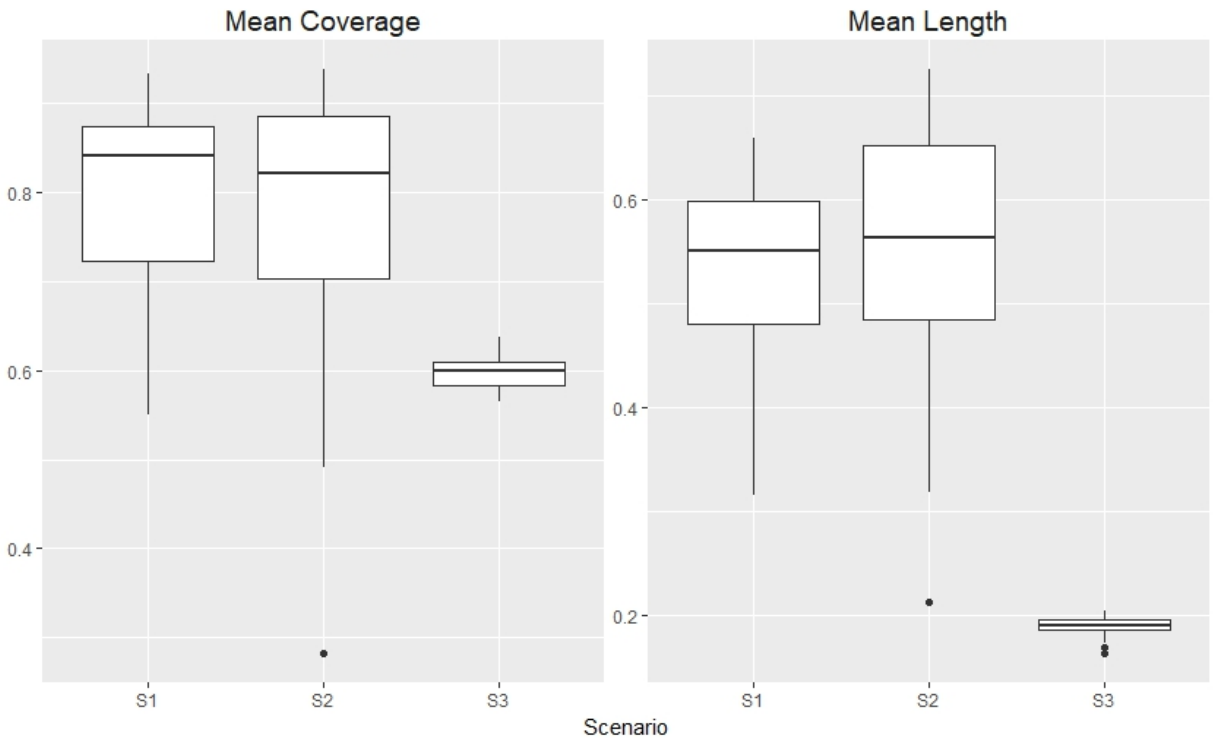
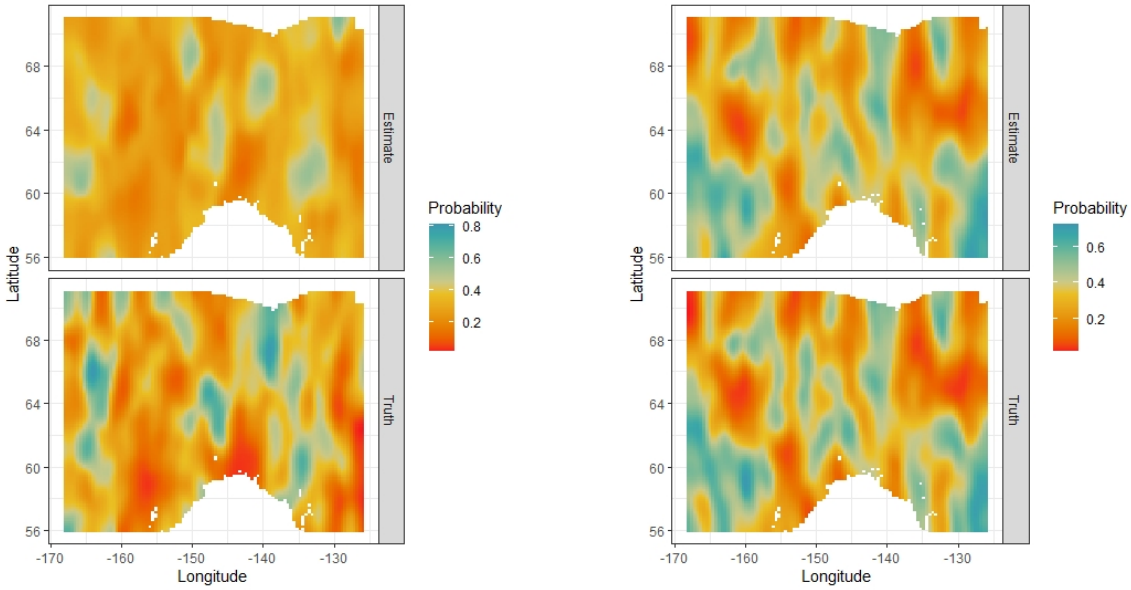


Figure 2.6: Mean coverage (left) and mean interval length (right) of 95% credible intervals for estimating $P(\boldsymbol{s})$ for each different sampling scenario.



(a) (b)

Figure 2.7: $\hat{P}(\boldsymbol{s})$ and $P(\boldsymbol{s})$ of the simulation run for which the correlation between $\hat{P}(\boldsymbol{s})$ and $P(\boldsymbol{s})$ is the median for (a) S2 and (b) S3.

2.4.3 Simulation results by data source

We also evaluate the estimation performance at genetic, pollen, and SDM sites as well as data-absent locations separately. This allows us to have a better understanding of how the different sources of evidence interact with each other and their roles in estimating $P(\boldsymbol{s})$.

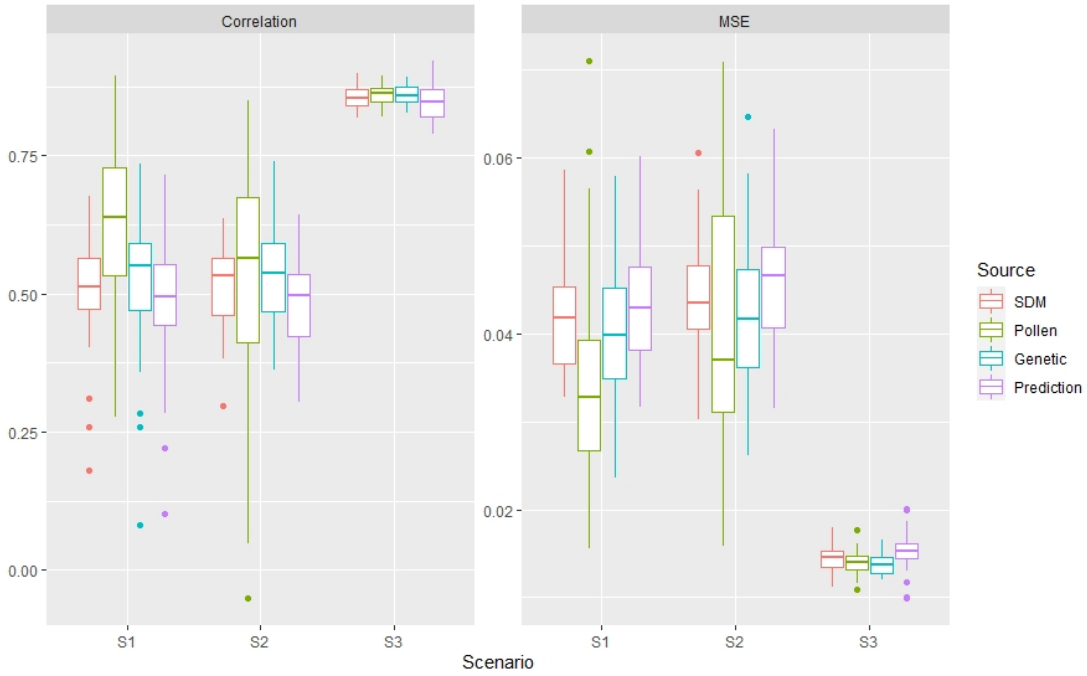


Figure 2.8: Correlation and MSE between $\hat{P}(\boldsymbol{s})$ and $P(\boldsymbol{s})$ for the data-source-specific sites as well as prediction sites.

As seen in Figure 2.8, pollen data does the best both in terms of correlation and MSE in S1 and S2. As stated in the previous section, pollen has a much stronger linear relationship to $X(\boldsymbol{s})$ than genetic (and SDM) and this is reflected in the simulation results for both sampling scenarios. When increasing SDM sample size it appears that pollen sites perform better, whereas SDM sites perform slightly worse. Conversely, the performance of genetic sites does not vary much from S1 to S2. This could be due to having a stronger prior than the other two data sources, which impacts the estimation of $P(\boldsymbol{s})$ much more in the presence of a small sample size. However, when having much more (and equal) sample size, genetic data performs the best out of all three data sources, as a result of being treated as the unbiased data source. The other two data types behave in expectancy to their modeling choices with pollen being linearly biased from $X(\boldsymbol{s})$ and SDM presenting further departures.

In all sampling scenarios, the prediction sites have the worst performance. However, the difference in estimation accuracy between prediction sites and the rest is not too significant indicating that the model does not suffer much from having too many prediction sites. Additionally, SDM seems to perform very similarly to genetic sites in S2 as compared to S1, meaning that the inclusion of more SDM sites does not seem to be biasing the results.

2.5 Refugia for Green Alder and White Spruce

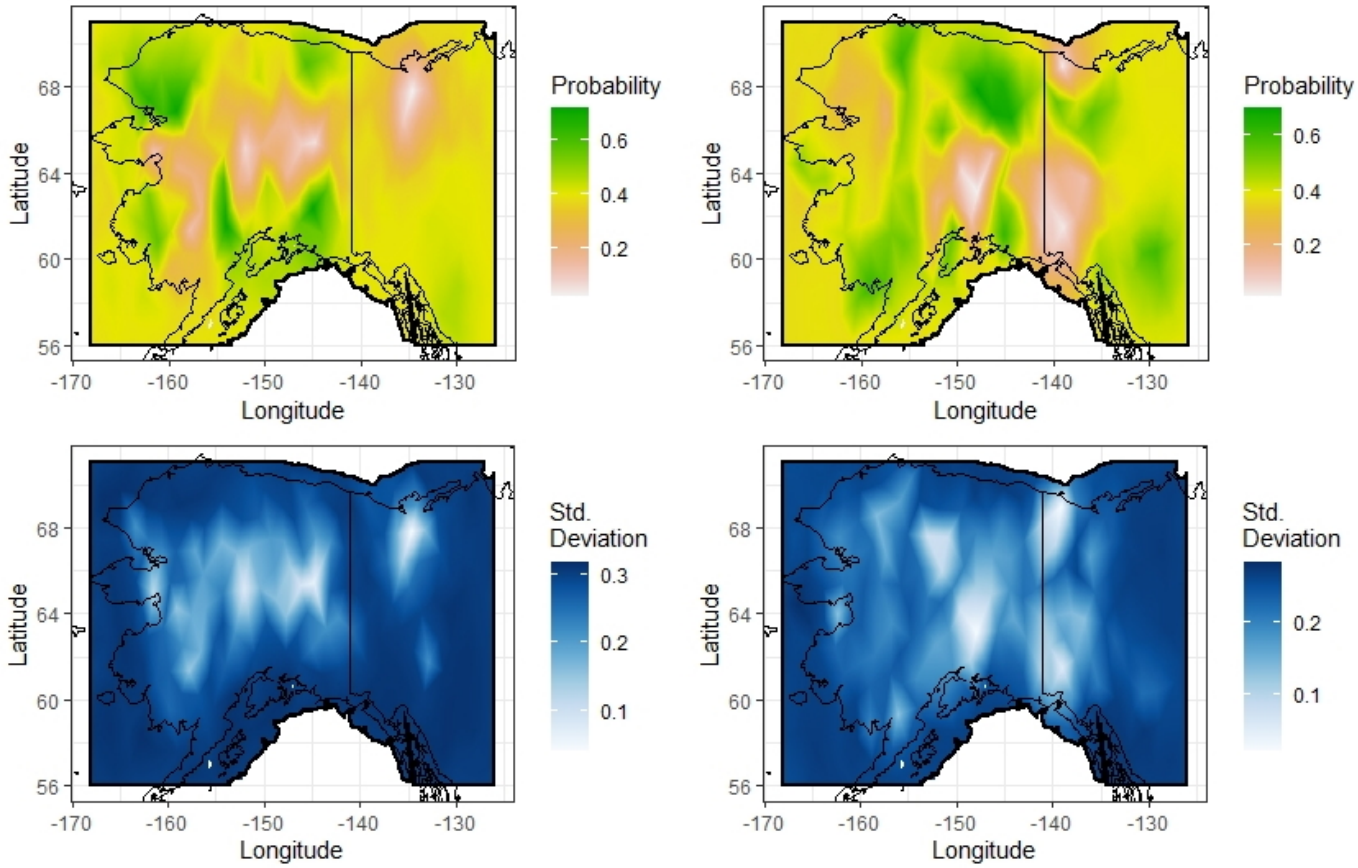
We apply our BHM to the Green Alder and White Spruce data to unveil the refugia of these two species in Eastern Beringia during the Last Glacial Maximum (LGM, 23-19 thousand years ago) period.

2.5.1 Green Alder

SDM output for green alder is composed of 334,000 sites that cover the Alaskan peninsula and parts of adjacent Canada. To ease the computation for this very large and dense data set, a random sample of 50,000 sites was drawn and used in the analysis. For the other two data sources, there were 47 genetic and 18 pollen observations, as shown in Figure 2.1a.

The posterior estimates of model parameters and hyperparameters are reported in Table 2.2. The negative estimate for μ suggests that on average the true probabilities are smaller than 0.5. This is mostly due to the overall SDM values being zero, but also to a lesser extent to using $r = 3$ for genetic interpolation which ensures most genetic probabilities are low. Additionally, all sources of evidence seem to agree amongst themselves since they all have positive β, i.e., if an area has high observed probabilities among data sources then the common probability of that area being refugia is also expected to be high. The posterior probability for $\theta_2 > 0$ was calculated and was found to be almost 1. This seems to signal that there is some tangible difference in uncertainty between S_{m0} and S_{m1}, with S_{m1} having a greater variability.

Figure 2.9a shows the posterior mean and standard deviation of $P(\boldsymbol{s})$ on all sites within the study region. We can see that the areas of low variability coincide with the regions where we have pollen or genetic data, whereas high uncertainty occurs in places where neither of these two data sources is present. Reduced variability at the co-occurrence of disparate data suggests that the different lines of evidence provide unique, complementary information about the true source locations of past populations (Gavin et al. 2014). Our results broadly match the findings of Napier et al. 2019 who



(a) (b)

Figure 2.9: Posterior mean and standard deviation of the true probability of being refugia, $P(\boldsymbol{s})$, of (a) green alder and (b) white spruce. The shaded region corresponds to Eastern Beringia during the Last Glacial Maximum, with modern-day Alaska superimposed for reference.

analyzed the same SDM and genetic dataset but in a framework that did not integrate those two data sources.

Parameter	Mean	St. Dev	2.5% Quantile	97.5% Quantile
μ	-0.5599	0.307	-1.1433	0.0805
α_m	-29.5723	8.0242	-48.8235	-17.0337
α_p	-0.2138	1.0189	-1.9813	2.0104
β_m	0.661	1.2104	-1.8335	2.912
β_p	1.2099	1.1049	-0.8111	3.5188
$\sigma_m(S_{m0})$	19.6507	4.5121	14.5895	27.0541
$\sigma_m(S_{m1})$	23.6677	4.7163	17.8428	31.3969
σ_p^2	8.1744	3.0616	3.8316	15.7032
σ_g^2	0.4164	0.3083	0.0897	1.2436
σ_x	1.4068	0.3866	0.7915	2.2974
ρ_m	11.5636	2.9265	8.4688	16.2158
ρ_x	2.564	0.9674	1.1734	4.9291

Table 2.2: Posterior estimates of *Alnus viridis* model.

2.5.2 White Spruce

For white spruce, there are also 334,000 sites of SDM information covering the same study region as green alder. Likewise, we randomly sampled 50,000 sites for our analysis. Additionally, there were 79 genetic and 14 pollen observations, as shown in Figure 2.1b.

The posterior estimates for all model parameters and hyperparameters are shown in Table 2.3. Notice that the estimate for σ_p^2 is very small. This is most likely due to the pollen data for white spruce being almost invariable. Even though this hyperparameter represents the uncertainty of pollen data, which is expected to be larger than that of genetic information, the small variation of the pollen data caps the magnitude of the values σ_p^2 can take. Similar to green alder, the overall mean is also negative for the white spruce estimation, thus suggesting that indeed the locations of refugia are sparse. Furthermore, the variance for the $\boldsymbol{z}$ process has a posterior probability of 0.896 of being larger in S_{m1} than in S_{m0}.

Parameter	Mean	St. Dev	2.5% Quantile	97.5% Quantile
μ	-0.4612	0.247	-0.9454	0.0263
α_m	-25.6744	6.7762	-39.3970	-12.9561
α_p	1.5932	0.1800	1.2677	1.9845
β_m	1.2203	0.2027	0.7864	1.6075
β_p	1.021	0.192	0.6602	1.4474
$\sigma_m(S_{m0})$	15.0940	1.6044	11.4026	17.5576
$\sigma_m(S_{m1})$	16.4105	1.7294	13.0457	18.7793
σ_p^2	0.0002	0.0003	0	0.0009
σ_g^2	0.6045	0.1185	0.4148	0.8778
σ_x	1.0954	0.0888	0.9424	1.3586
ρ_m	12.2107	1.4162	9.0695	14.8249
ρ_x	3.6014	0.5473	2.5171	4.967

Table 2.3: Posterior estimates of *Picea glauca* model

Figure 2.9b shows the mean and standard deviation of the posterior density of $P(\boldsymbol{s})$ on all sites within the study region. Same as with green alder, the regions where there are genetic and pollen data have lower variability, while the regions lacking those two types of data have increased uncertainty. This is again evidence of how sites with multiple sources of evidence coincide in the estimation of $P(\boldsymbol{s})$. Since our results for white spruce are the first attempt to find the exact locations of refugia, there is no existing literature to verify our results yet. However, the resulting highlighted regions, such as a possible refugium in Alaska, seem reasonable from a scientific point of view.

2.6 Conclusions

We propose an innovative Bayesian hierarchical model to utilize the diverse pieces of evidence collected in paleoecology to uncover cryptic refugia. Specifically, we integrate SDMs, pollen fossil records, and genetic surveys as three complementary sources of evidence to produce a single unified map showing the possible refugia locations in Eastern Beringia for the *Alnus viridis* and *Picea glauca*, respectively.

The simplicity of the model plus the computational convenience offered by INLA allow for

researchers to quickly and efficiently implement the model with their data sources. The flexibility given by the Bayesian hierarchical modeling also allows researchers to model their specific data sources in any way they consider best.

Due to the intricate relationship between the raw data and the true probability of being refugia, we transform the data beforehand to enable a more straightforward relationship to the true probabilities based on empirical experience and prior knowledge. Alternatively, we could consider incorporating the data pre-processing process into the hierarchical model and learning all parameters from the data. However, there are no widely accepted pre-processing models for us to borrow at this point. We thus only focus on integrating different lines of evidence rather than extensively exploring pre-processing approaches. Although we consider those transformations sound choices, we acknowledge that further investigation is needed.

Our models are constructed based on the perceived data quality and where each data source exhibits credibility. According to their reliability, pollen and genetic data were modeled using linear biases, whereas SDM used a non-linear bias with a logit transformation. This allows the results to resemble more accurate data sources while receiving due help from other available evidence. The high uncertainty of the probability estimates in regions far from pollen and genetic data suggests that SDM provides only a little information to identify high-plausible refugia. However, SDM helps identify the areas where refugia can be safely discarded, as evidenced by the reduced variance of S_{m0} in both species. Pollen and genetic data are then responsible for bringing to light some specific regions of high probability.

It would be overly optimistic to conclude that these highlighted regions are the sites of glacial refugia, however, we feel that they can be safely used to inform future field expeditions for the further study of cryptic refugia in Eastern Beringia. Furthermore, we hope that the method disclosed in this chapter can turn into a powerful and useful tool for further investigations in paleoecology for many other species. With the insight gained from such studies, we can better prepare for the coming challenges brought by rapid climate change.

Even though we try to rigorously integrate three different data sources to provide the refugia estimates, our method is still an indirect approach to detecting refugia. The results obtained through our analysis should be treated with caution. To formally confirm some locations to be refugia of a species, we need to find in situ macrofossils that can prove this species' past presence.

Searching for such macrofossils is an extremely difficult task without prior knowledge of where to target (Lafontaine et al. 2014). The plausible refugia we identified here can effectively guide the search for actual proof of the species' past presence.

Chapter 3

Estimation of Solar-Induced Chlorophyll Fluorescence Yield of Various Vegetation Land Types Using a Spatially Varying Coefficient Model

3.1 Introduction

Large-scale crop monitoring provides numerous benefits for agriculture and related sectors. It enables farmers and agricultural managers to make informed decisions about crop health, optimizing yields and reducing resource waste. Early detection of crop stress factors and diseases helps prevent damage and economic losses. Precision agriculture techniques can be implemented, by utilizing remote sensing technologies to gather high-resolution data for site-specific interventions.

Historically, these approaches relied on correlations between the reflectance of biomass and its relation with photosynthesis. Recent advances in satellite spectroscopy have allowed a new approach to estimating photosynthesis based on the flux of chlorophyll fluorescence emitted by the canopy (Guanter et al. 2007; Guanter et al. 2014; Guan et al. 2016). One such flux, Solar-induced fluorescence (SIF), refers to the emission of light by plants and vegetation in response to sunlight. When plants undergo photosynthesis, they absorb sunlight and convert it into chemical energy. Part

of the absorbed light is re-emitted as fluorescence, which is emitted at longer wavelengths than the absorbed light. This means that SIF is a direct by-product of the photosynthesis process. SIF is a relatively weak signal and occurs in the far-red and near-infrared regions of the electromagnetic spectrum. The intensity of SIF is influenced by various factors such as vegetation type, leaf chlorophyll content, plant stress, and environmental conditions.

Guan et al. 2016 have provided a framework that links SIF retrieval and crop yield. It utilizes SIF retrievals from the Global Ozone Monitoring Experiment-2 satellite to provide better crop productivity measures than other traditional crop monitoring approaches. Nevertheless, the information retrieved by the satellites is an aggregate of different vegetation types and must be decomposed to be useful for specific crop monitoring.

Satellite images contain a multitude of footprints that contains undifferentiated SIF information for different vegetation types. Let the footprint location be denoted by $\boldsymbol{s}$, $\boldsymbol{s} \in D \subset \mathbb{R}^2$. Wang et al. 2020 break down the total SIF value for footprint $\boldsymbol{s}$ as:

$$\text{SIF}(\boldsymbol{s}) \approx \sum_{j=1}^{p} \pi_j(\boldsymbol{s}) \cdot \text{PAR}(\boldsymbol{s}) \cdot \text{fPAR}_j(\boldsymbol{s}) \cdot \text{SIF}_{yieldj}(\boldsymbol{s}), \tag{3.1}$$

where PAR stands for photosynthetically active radiation (i.e. light that can be used for photosynthesis), fPAR_j is the fraction of PAR absorbed by the canopy of vegetation j, π_j is the proportion of the footprint that contains vegetation j and SIF_{yieldj} is the specific SIF value coming from vegetation j. The approximation in the equation is due to only using the p most significant vegetation types in the footprint.

If we define our known quantities per footprint and land type as $x_j(\boldsymbol{s}) = \pi_j(\boldsymbol{s}) \cdot \text{PAR}(\boldsymbol{s}) \cdot \text{fPAR}_j(\boldsymbol{s})$, rename the SIF yield as $\beta_j(\boldsymbol{s}) = \text{SIF}_{yieldj}(\boldsymbol{s})$ and the total SIF as $y(\boldsymbol{s})$, then the form of linear regression with spatially varying coefficients (SVC) becomes much clearer:

$$y(\boldsymbol{s}) = \sum_{j=1}^{p} x_j(\boldsymbol{s})\beta_j(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}) \tag{3.2}$$

where $\varepsilon(\boldsymbol{s})$ is an error term that takes into account all the non-measured land types. We would expect this error to be small in size, as well as strictly non-negative given that all measurements are non-negative. Nevertheless, fitting this model requires some sort of penalty or modeling choice to

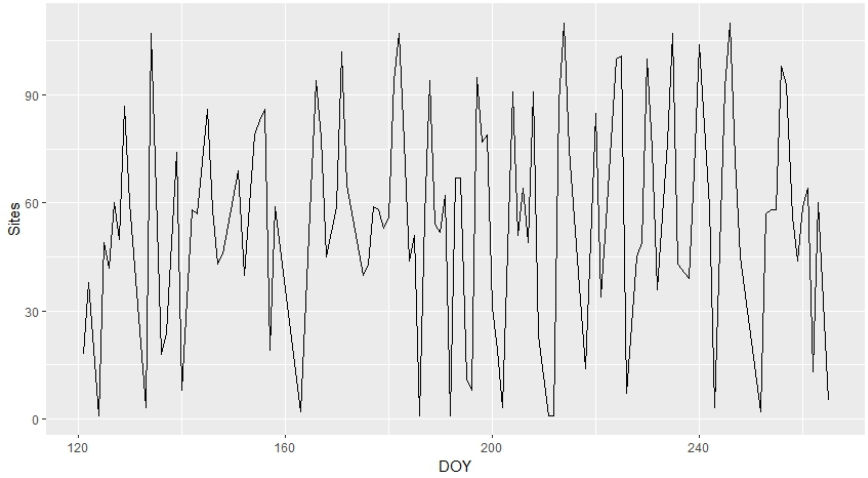


Figure 3.1: Number of sites/footprints per day of the year (DOY)

compensate for over-saturation. One approach that permits keeping the structure given by equation (3.1) is to use a fused penalty, similar to that employed by Li and Sang 2019.

In this chapter, we plan to use the spatial fused penalty structure presented by Li and Sang 2019, but using an L2 penalty instead of L1, to estimate a continuous field for the SVC in an expedited manner. Section 3.2 introduces the data that will be used in the application as well as the particular challenges it provides. Section 3.3 gives some background information on other SVC models as well as detail the modifications we perform for our particular application and the considerations we have to take to be able to perform successful tuning of the penalty hyperparameter. Section 3.4 is a simulation study to assess the accuracy of the coefficient estimations as well as a comparison with some of the other methods in the literature. Results are presented in Section 3.5 and discussed in Section 3.6.

3.2 Data

The SIF retrievals come from the Global Ozone Monitoring Experiment-2 (GOME-2) during 2007-2012, as detailed in Guan et al. 2016. The main area of study relates to Urbana-Champaign, Illinois, USA, and its surrounding areas (longitude: -88.6 to -87.8, latitude: 39.8 to 40.4). The predominant vegetation types are corn and soy, with forest and grass present at a much smaller percentage. There are other vegetation types that influence the total SIF values, but that are not taken into account directly and instead are deferred to the error term.

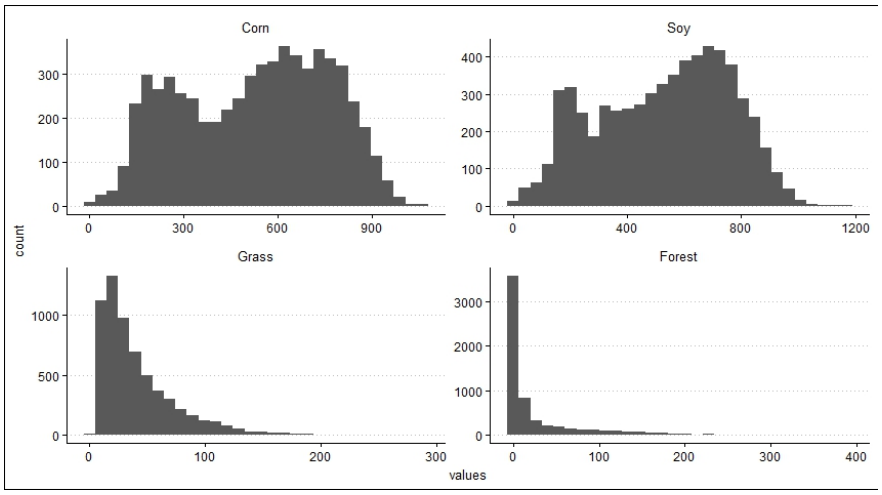


Figure 3.2: Histograms of the covariates for each land type for all DOY in study.

The data is only taken for one year for the months of May through September, which cover the end of Spring, the whole Summer, and the start of Fall. This is the time of the year when corn and soy grow and is harvested, so it makes sense to monitor their SIF to measure crop yield. However, not all days are measured sequentially due to the traversal patterns of the satellite. Furthermore, some days have less than 4 measurements, making them unusable for our estimation. Taking these days away from our study, we are left with 112 days of data. The number of sites per day of the year is shown in Figure 3.1 and as seen there are not many sites per day with the average being around 55.

A challenge that arises from this data is the scale of the covariates compared to the response variable. The observed SIF values range between 0 and 5, whereas the values of PAR times fPAR are between 300 and 2000, depending on the land type. The percentages of footprint that belong to each land type are mostly dominated by corn and soy, where their mean percentages are around 0.45, whereas for forest and grass, they are much smaller, around 0.05. This heavily unbalances the relation between SIF and the covariates and gives much more weight to corn and soy compared to forest and grass. This also changes the shape of how the covariates look, where corn and soy have a much wider distribution whereas grass and forest have longer tails (see Figure 3.2).

The observed correlations between SIF and corn and soy covariates are around 0.30 on average for each, whereas the correlations between SIF and forest and grass are closer to 0. However, as Figure 3.3 shows, the correlation is not always constant and seems to have a temporal effect. Both

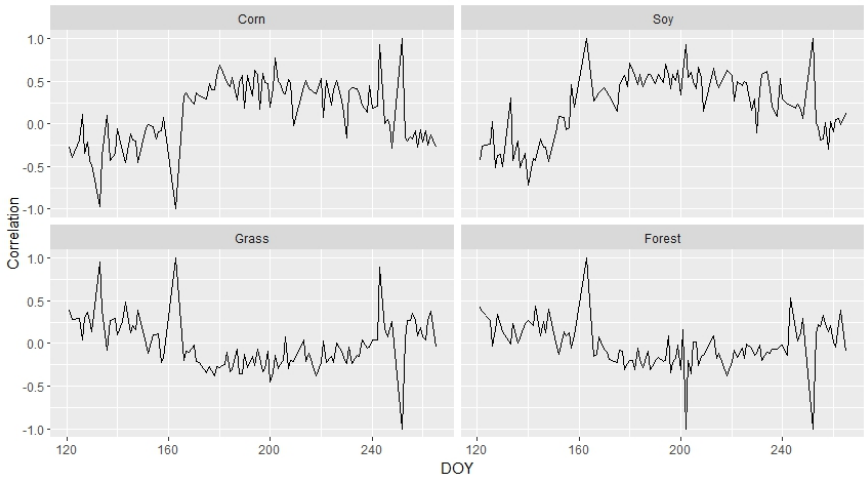


Figure 3.3: Daily correlations between the covariates and SIF for each different type of vegetation

corn and soy covariates seem to correlate more with SIF during the summer whereas grass and forest have lower correlations. This relationship is reversed for late spring and early fall, where it is grass and forest that have a higher correlation with SIF than corn and soy. SIF and the corn and soy covariates also show a similar seasonal trend, where their values increase as spring moves to summer and decrease as summer gives way to fall.

3.3 Spatially varying regression model for estimating SIF

3.3.1 Fitting the model

To fit the model we will take into consideration adding a penalty similar to that in Li and Sang 2019, where we shrink the differences between adjacent sites according to a minimum spanning tree (MST) structure following their advice to efficiently find said tree. However, unlike the mentioned paper, we will use an L2 penalty to induce spatial smoothness in the coefficient estimations.

Let $\boldsymbol{\beta}_j$ be the vector of size n containing all coefficients for the j-th land type. Additionally, define $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1^T & \cdots & \boldsymbol{\beta}_p^T \end{pmatrix}^T$ to be the concatenated vector of size np that contains all coefficients. Similarly, define the matrix $\boldsymbol{X}_j = \text{diag}(\boldsymbol{x}_j)$ to be a $n \times n$ diagonal matrix whose elements correspond to the covariates of the j-th land type and let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_p \end{bmatrix}$ be the $n \times np$ block matrix that groups all the previously defined diagonal matrices. Furthermore, let the incidence matrix defined by the MST be called $\boldsymbol{A}$. This is a $n-1 \times n$ matrix whose columns represent the locations

and each row represents an edge of the MST, with a 1 and -1 on the columns of the sites that share that edge, with the remaining values being 0. Furthermore, let $\mathcal{A} = \boldsymbol{I}_p \otimes \boldsymbol{A}$, where '$\otimes$' stands for the Kronecker product. With this, we can properly define the estimation problem:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \succcurlyeq 0} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2} \|\mathcal{A}\boldsymbol{\beta}\|_2^2 \right\} \tag{3.3}$$

where we use $\boldsymbol{\beta} \succcurlyeq 0$ to refer that all elements of $\boldsymbol{\beta}$ have to be non-negative. This constraint is important since $\boldsymbol{\beta}$ represents a measure of light that is by definition non-negative. Note that we can expand the equation above to include all the quadratic terms together:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \succcurlyeq 0} \left\{ \frac{1}{2} \boldsymbol{\beta}^T \left[\boldsymbol{X}^T\boldsymbol{X} + \lambda \mathcal{A}^T\mathcal{A} \right] \boldsymbol{\beta} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} \right\} \tag{3.4}$$

Without the non-negativity constraint we can find a closed-form solution for the problem, however including the constraint makes the estimation challenging. A natural choice is to reparameterize β to e^θ, $\theta \in \mathbb{R}$, however, this requires a different optimization strategy than taking the derivative of the objective function. One option is to apply a second-order Taylor expansion to e^θ, making the objective function linear with respect to θ and allowing us to take derivatives. This is done in a similar framework in Sass, Li, and Reich 2021, which also includes a spatially fused penalty. However, this method is computationally taxing as it requires multiple iterations of estimation to reach a stable estimate. This must be done for a fixed value of λ, and thus tuning for this parameter becomes even more computationally expensive. Additionally, using AIC to tune λ is unrealistic, as it continuously decreases as λ increases, over-penalizing the estimations. Instead, we resort to optimizing with quadratic programming, using the dual method implemented by Goldfarb and Idnani (1982; 1983).

The tuning parameter λ can be chosen using a metric such as AIC or the corrected AIC for small sample sizes. It is important to note that due to the nature of the problem, tuning λ using cross-validation is unfeasible since removing observations also removes those coefficients from the model fit. Additionally, for calculating AIC the degrees of freedom are approximated as $\text{trace}\left(\boldsymbol{X}\left[\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathcal{A}^T\mathcal{A}\right]\boldsymbol{X}^T\right)$, which includes the shrinkage effect of λ and the MST penalty but doesn't include the penalty imposed by the non-negativity constraint.

It is possible to generalize the model to include different tuning parameters λ_j, one for each

land type if preferred. Let $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_p)$ be a $p \times p$ diagonal matrix whose elements are the tuning parameters. Then, in equation (3.4) we can substitute $\lambda \mathcal{A}^T \mathcal{A}$ for $\Lambda \otimes \boldsymbol{A}^T \boldsymbol{A}$ to achieve different penalization for each land type. This can be of use when we want different degrees of spatial smoothness for each land type, but the estimation of the model will increase in time exponentially - for the purposes of this chapter we opted to only use one value of λ. Nevertheless, we scale the covariates to alleviate the stress on λ from different variabilities. Note that if we define $\Lambda = \lambda \boldsymbol{I}_p$ then we get back the original $\lambda \mathcal{A}^T \mathcal{A}$ of equation (3.4).

3.3.2 Incorporating zero coefficients into the model

As mentioned before, all the quantities in our model (except the Gaussian error) are non-negative. This is especially important since there exist footprints where their value of SIF is zero. This adds some deterministic information to our estimation, as a value of zero for SIF and strictly positive covariates implies that the coefficients for that particular site must all be zero as well. In other words, if $y(\boldsymbol{s}) = 0$, for some site $\boldsymbol{s}$, then $\beta_j(\boldsymbol{s}) = 0, \forall j = 1, ..., p$.

Let $s_0 = \{i : y(\boldsymbol{s}_i) = 0, i = 1, ..., n\}$ be the collection of indices whose sites correspond to one of the deterministic sites previously mentioned - we will refer to these sites as *zero coefficients sites*. We know $\beta_j(\boldsymbol{s}_i) = 0, i \in s_0, \forall j = 1, ..., p$ so we can use this information when estimating the remaining coefficients. If we know $i \in s_0$ and l corresponds to the index of a site adjacent to $\boldsymbol{s}_i$ in the MST, then the penalty term would become $(\beta_j(\boldsymbol{s}_i) - \beta_j(\boldsymbol{s}_l))^2 = \beta_j(\boldsymbol{s}_l)^2$ which reduces from a fused Ridge penalty to a simple Ridge penalty. This means we don't need to use the MST edge for cases that include a deterministic site and instead, we need an extra Ridge penalty for those sites that are adjacent.

To account for the sites indexed in s_0 we need to redefine the MST matrix $\boldsymbol{A}$ and create a new matrix that encodes the sites that receive a Ridge penalty. Let $\boldsymbol{A}_r$ be a reduced version of $\boldsymbol{A}$ where we remove the rows that have a 1 or -1 in columns indexed by s_0, i.e. we remove from $\boldsymbol{A}$ all rows that correspond to an edge that includes zero coefficients sites. Furthermore, let $\boldsymbol{Z}$ be a $n \times n$ diagonal matrix whose elements are $\boldsymbol{Z}_{ll} = 1$ if l indexes a site that is adjacent to a zero coefficients site, or 0 otherwise.

Finally, to deal with the issue of non-positive matrices, we remove the columns in $\boldsymbol{A}_r$ and $\boldsymbol{Z}$ that corresponds to s_0, as well as the rows in $\boldsymbol{X}$ and the entries in $\boldsymbol{y}$ that also correspond to s_0.

We will call these reduced dimensionality objects $\boldsymbol{A}_r', \boldsymbol{Z}', \boldsymbol{X}'$, and $\boldsymbol{y}'$ to distinguish them from the original ones. This will reduce the number of coefficients to estimate, so let $\boldsymbol{\beta}^+$ be the vector of coefficients that do not include those in s_0, which would be of length $n - |s_0|$. Let $\mathcal{A}_r = \boldsymbol{I}_p \otimes \boldsymbol{A}_r'$ and $\mathcal{Z} = \boldsymbol{I}_p \otimes \boldsymbol{Z}'$, then the estimation problem including the deterministic information becomes:

$$\hat{\boldsymbol{\beta}}^+ = \arg\max_{\boldsymbol{\beta} \succcurlyeq 0} \left\{ \frac{1}{2} \boldsymbol{\beta}^T \left[\boldsymbol{X}'^T \boldsymbol{X}' + \lambda \left(\mathcal{A}_r^T \mathcal{A}_r + \mathcal{Z} \right) \right] \boldsymbol{\beta} - \boldsymbol{y}'^T \boldsymbol{X}' \boldsymbol{\beta} \right\} \tag{3.5}$$

To obtain the original dimension for $\hat{\boldsymbol{\beta}}$, the estimation $\hat{\boldsymbol{\beta}}^+$ is expanded back to the original size by adding zeroes in the entries that originally corresponded to s_0.

3.3.3 Related models

It is common for environmental sciences to focus on the relationship between a response variable and many explanatory variables while taking into account the spatial dependence of the observations. Methods like Gaussian process regression (Cressie 1993) or spatial generalized linear regression models (Diggle, Tawn, and Moyeed 1998) model spatial dependence by utilizing a spatial random effect to the linear term of the regression model. However, these methods assume a constant effect of the explanatory variables throughout the whole spatial domain which limits the extraction of more interesting relationships between variables.

Allowing for spatially varying effects, the model of interest takes the general form $y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^T \boldsymbol{\beta}(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$, for a given site $\boldsymbol{s} \in S$. Assuming that for a particular problem, we have n samples and p explanatory variables, that leaves np coefficients to estimate which cannot be done directly without the use of further modeling choices. Frequentist approaches have done this by modeling the varying coefficients in a lower dimension or applying some constrained estimation, whereas Bayesian methods model the coefficients through a hierarchical model.

A simple exploration into allowing spatial flexibility of the regression coefficients was first presented by Casetti 1972; Casetti 1997 as a spatial expansion method. In this approach, the spatially varying coefficients (SVC) are represented by polynomials of a certain degree of the spatial coordinates, e.g. for $\boldsymbol{s} \in \mathbb{R}^2, \beta_j(\boldsymbol{s}) = \sum_{i=0}^r \alpha_i s_1^r + \gamma_i s_2^r, r > 0$. The selection of the polynomial order r is a key choice in the application. Although drastically reducing the number of coefficients to estimate and speeding computational time, the polynomial assumption of the underlying spatial

effect is too limited when detecting more complex fields.

Among the more popular methods for SVC models is the Geographical Weighted Regression (GWR; Brunsdon, Fotheringham, and Charlton 1996; Brunsdon, Fotheringham, and Charlton 1998; Fotheringham, Charlton, and Brunsdon 1998) which estimates $\beta(\boldsymbol{s})$ by assigning weights based on a distance decay kernel, such as the exponential kernel (Wheeler and Calder 2007; Wheeler and Waller 2009). For a given site, the coefficients are estimated as $\hat{\boldsymbol{\beta}}(\boldsymbol{s}) = \left[\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{y}$, where $\boldsymbol{W}(\boldsymbol{s})$ is a diagonal $n \times n$ matrix whose j-th element represents the weight assigned to the j-th sample. Spatial variation is measured with a single kernel bandwidth, which creates a limitation by fixing the scale of each coefficient to be the same. To overcome this limitation, the bandwidth can be allowed to vary for each explanatory variable, at the cost of computation time (Fotheringham, Yang, and Kang 2017).

Another popular approach is the more robust Bayesian alternative to SVC, introduced by Gelfand et al. 2003. In their approach, they assign a multivariate normal latent process to the vector of SVC. More specifically, the coefficients are given a multivariate normal prior with mean $\mathbf{1}_n \otimes \mu_\beta$ and covariance matrix given by $H(\phi) \otimes T$, where $\mathbf{1}_n$ is an $n \times 1$ vector of ones, $\mu_\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is the vector of fixed means, $H(\phi)$ is an $n \times n$ covariance matrix that models spatial dependency between locations, with hyperparameters ϕ, and T is a $p \times p$ matrix that models the correlation between coefficients. The terms in $H(\phi)$ can be specified through the user's choice of a stationary covariance function. Furthermore, estimation is typically done using MCMC with a slice Gibbs sampler, which is computationally heavy when n or p is large.

Another method that does not require using dimension reduction methods prior to fitting the model is the spatially clustered coefficient (SCC) regression from Li and Sang 2019. Similar to clustered Lasso (She 2010), SCC uses a clustered L1 penalty on the coefficients that are sufficiently near each other. This is done computationally efficiently by using the edge set of the minimum spanning tree defined on the spatial sites. The method is very fast with large data sets but is limited by only being able to estimate clustered spatial fields for the coefficient, which is not ideal if the underlying fields are assumed to be smooth.

3.4 Simulation Study

The simulation study serves two purposes: to evaluate how well the coefficients of different land types are estimated relative to each other when the true coefficients are modeled with different amounts of noise and to compare the estimation results of this method against other methods in the literature. For this, 1000 simulations will be done where each time 100 random locations are sampled within $[0, 1]^2$.

For each simulation, we generate the covariates so that they resemble the magnitudes of those in the real data. To then achieve a good simulation of SIF, the coefficients are simulated according to what their magnitudes would be so that when multiplied by the covariates and added together would give SIF values similar to the real ones. Coefficients and covariates are multiplied together and added with an extra Gaussian noise error term to create the simulated SIF; this noise is very small to help maintain the linear relation between the simulates SIF and covariates similar to the correlations with the real data. All quantities are resampled until everything, except the Gaussian noise, is strictly positive. Details are in the following paragraphs.

To simulate the magnitudes of the covariates observed in the real data, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are simulated from Gaussian Processes with a mean of $\mu = 650$ and utilizing the Matérn covariance function. Both covariance functions utilize a range of $\rho = 0.2$ and smoothness of $\nu = 1$, but $\boldsymbol{x}_1$ utilizes a standard deviation of 100 while $\boldsymbol{x}_2$ uses 50. This simulates the magnitudes of the observed data appropriately while giving some spatial correlation to the covariates. $\boldsymbol{x}_1$ has a higher variance since it does show higher and lower values than $\boldsymbol{x}_2$ in the dataset.

The remaining covariates $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$ have to have much smaller magnitudes and also present a heavily asymmetrical shape. For that reason, they are both simulated using Exponential distributions with means of 45 and 20, respectively.

The coefficients are all simulated from Gaussian processes with Matérn covariance functions and their magnitudes are so such that the resulting simulated SIF has the same range as the empirical one. $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are both simulated from a GP with a mean of 0.002, range of 0.2, smoothness of 5, and standard deviation of 0.0004. The last two remaining coefficient fields, $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$, have a much smaller mean of 0.0005, a standard deviation of 0.00002, the same smoothness at 5 and different ranges with 0.33 for $\boldsymbol{\beta}_3$ and 0.4 for $\boldsymbol{\beta}_4$.

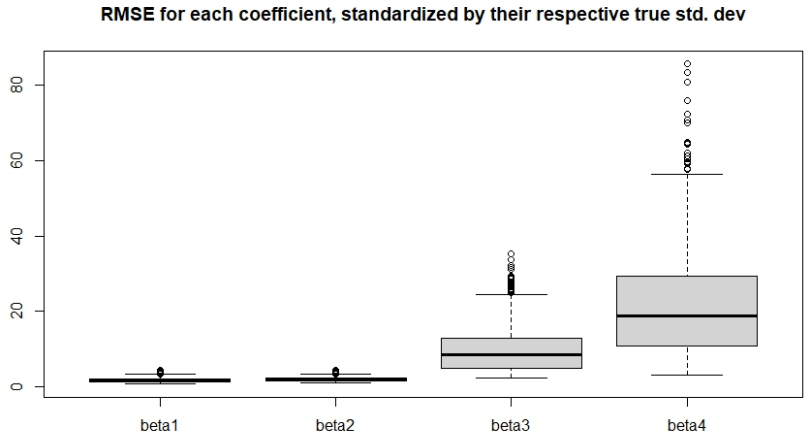


Figure 3.4: RMSE between the true and estimated $\boldsymbol{\beta}$,, divided by the true standard deviation, for each simulated land type

Lastly, we simulate SIF as $SIF = \sum_{j=1}^{p} \boldsymbol{x}_j \boldsymbol{\beta}_j + \varepsilon$ where $\varepsilon \sim N(0, 0.01)$. If all quantities, except ε, are positive then those become the simulated data for the simulation, otherwise, everything is resampled until all the quantities of interest are non-negative.

3.4.1 Coefficient Estimation

The first simulation study assesses how the model estimates each land type coefficient relative to each other. For this, we calculate the root mean square error (RMSE) between the real $\boldsymbol{\beta}_j$ and the estimated $\hat{\boldsymbol{\beta}}_j$ and divide it by the standard deviation used for that respective coefficient. As seen in Figure 3.4, the coefficients for the first two land types are estimated much better than those for the last two. This is not surprising given that there is more information about the first two coefficients on the simulated SIF than there is about the last two. Furthermore, the last coefficient shows a much higher standardized error than all the other three possibly due to its respective covariate being much smaller than all the rest.

3.4.2 Method Comparison

We compare our method against other common methods in order to evaluate how well they compare in terms of estimation error and correlation between the true and estimated $\boldsymbol{\beta}$. The methods chosen to compare against are the Spatial Expansion Method (SEM), Geographically Weighted Regression

(GWR), Bayesian Spatially Varying Coefficient Model (Bayes), and Spatially Clustered Coefficients (SCC). Since none of these methods allow for the addition of a non-negativity constraint, as allowed by their respective R packages, we will compare them against our method without the constraint (Ridge) and with the constraint (Ridge.QP). The quantities that will be used for the comparison are the RMSE between $\hat{\boldsymbol{\beta}}_j$ and $\boldsymbol{\beta}_j$ as well as the correlation between those two. We also consider Multiple Linear Regression (MLR) as an additional method when comparing RMSE.

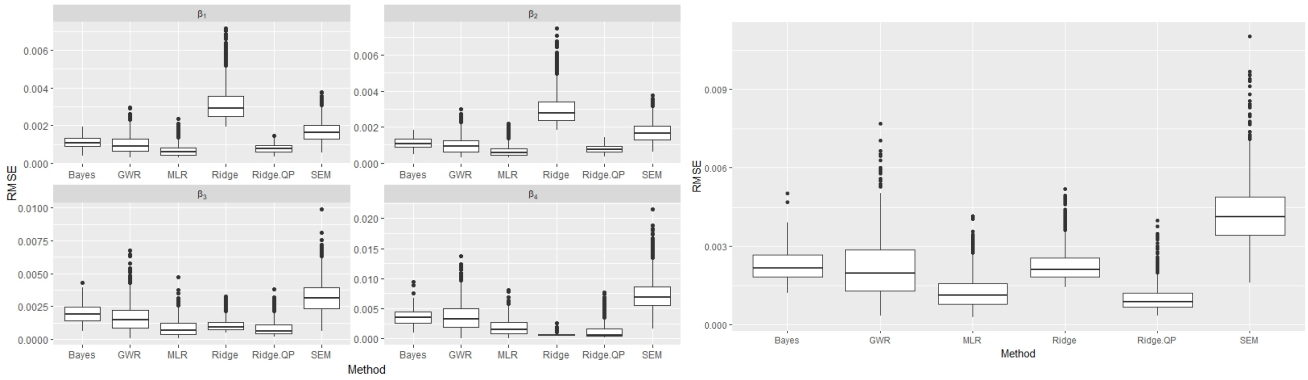
For SEM the coefficients are defined as a quadratic polynomial with respect to latitude, s_1, and longitude, s_2: $\beta_j(\boldsymbol{s}) = \alpha_{0,j} + \alpha_{1,j}s_1 + \alpha_{2,j}s_2 + \alpha_{3,j}s_1^2 + \alpha_{4,j}s_2^2 + \alpha_{5,j}s_1s_2$. This way, each land type would have 6 coefficients associated with it, for a total of 24 coefficients to estimate which is much lower than before. For GWR we utilize a Gaussian kernel for the distance decay and a single global bandwidth was estimated through cross-validation. This is all implemented in the R package 'spgwr' (Bivand and Yu 2022).

For the Bayesian alternative, we define the spatially varying coefficients as a mixture of a fixed effect and a spatial random effect, i.e. $\tilde{\beta}_j(\boldsymbol{s}) = \beta_j + \beta_j(\boldsymbol{s})$. Let $\tilde{\boldsymbol{\beta}}$ be the $np \times 1$ vector of spatially varying coefficients as previously defined, and let $\boldsymbol{X}$ be the same $n \times np$ matrix defined in Section 3.3, then we can write our model as $\boldsymbol{y} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. Let $\boldsymbol{\mu}_\beta = \begin{pmatrix} \beta_1 & \beta_2 & \dots & \beta_p \end{pmatrix}^T$ be the $p \times 1$ vector that holds all the fixed effects. Then, for some covariance matrix $H(\phi)$ that models spatial dependency between locations, with hyperparameters ϕ, and another matrix T that models the correlation between coefficients, we have:

$$\tilde{\boldsymbol{\beta}} \sim N\left(\mathbf{1}_n \otimes \boldsymbol{\mu}_\beta, H(\phi) \otimes T\right)$$

For this simulation study, $H(\phi)$ follows a Matérn covariance function with spatial decay parameter ρ and smoothness ν. A uniform prior from 1% to 120% of the maximum distance is used for ρ, whereas ν is fixed at 1 to ease computations. For the covariance matrix T, we define an inverse Wishart prior for it, with p degrees of freedom and a $p \times p$ identity matrix as the scale matrix. Finally, the precision $\tau = \frac{1}{\sigma^2}$ is given an inverse Gamma prior with shape 4 and scale 1 as a non-informative prior.

To fit the model we use the built-in functions from the package 'spBayes' (Finley, Banerjee, and Carlin 2007; Finley, Banerjee, and E.Gelfand 2015). This consists of utilizing MCMC to obtain the



(a) (b)

Figure 3.5: RMSEs for (a) each $\boldsymbol{\beta}_j$ and (b) overall, for each method.

posterior densities for the random fields of the coefficients of each land type. For this, we obtain 55,000 posterior samples with a burn-in of 5000 and thinning every 150 samples. Afterward, $\hat{\boldsymbol{\beta}}$ is obtained as the posterior mean, out of 334 posterior samples. Unlike the rest of the methods which are repeated 1000 times, the Bayesian method is done only 100 times due to its excessive runtime.

When applying SCC we follow the steps taken by Li and Sang 2019. This consists of augmenting the matrix $\mathcal{A}$ by p rows, each one which consists of $\frac{1}{n}$ repeated n times in the p-th position block-wise and the remaining entries as zeroes. This new matrix, called $\tilde{\mathcal{A}}$, is a $np \times np$ invertible matrix and so we re-parametrize the model using $\boldsymbol{\theta} = \tilde{\mathcal{A}}\boldsymbol{\beta}$. This reduces the lasso penalty to just $\|\boldsymbol{\theta}\|_1$ on all entries except the last p. This model can then be fit by making use of the 'glmnet' package in R (Friedman, Hastie, and Tibshirani 2010) and then transforming back the results to obtain $\boldsymbol{\beta} = \tilde{\mathcal{A}}^{-1}\boldsymbol{\theta}$. Tuning for the shrinkage parameter λ is done using the corrected AIC.

The RMSE for all methods (except SCC) is presented in Figure 3.5. SCC is excluded because due to clustering estimations, RMSE is much higher than all other methods and impacts the scale of the plots. These simulations show that clearly, SCC is not able to predict enough different clusters to be similar to other methods that are able to estimate a smooth process. This is especially true for the two coefficients with the stronger signal, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. However, for the other two coefficient fields, $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$, it does similar to other methods. However, this happens because it is estimating much of those fields as 0 (in fact, out of 1000 simulations only one did not have all $\boldsymbol{\beta}_4$ be estimated as 0). Since all the methods except Ridge.QP has negative estimates when the true values are all non-negative, the fact that they are all estimated as 0 will indeed reduce the error. However, when

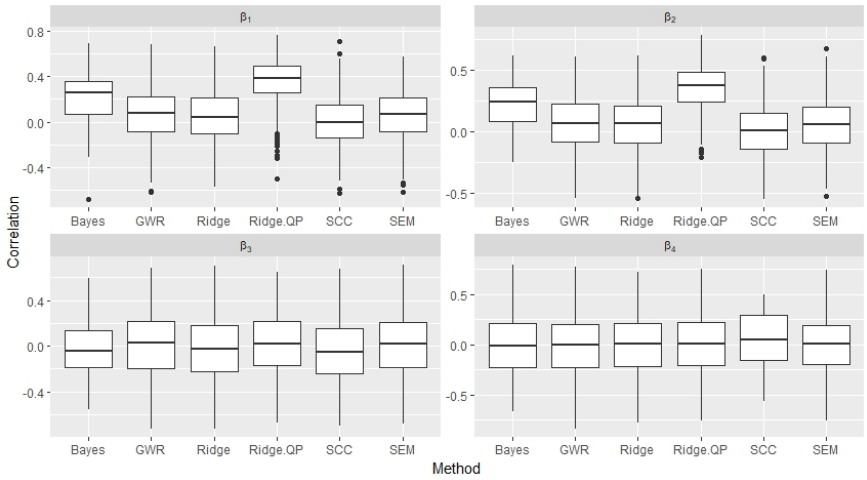


Figure 3.6: Correlations for each $\boldsymbol{\beta}_j$

all estimates are set to 0, these will fail to capture the underlying spatial pattern of interest.

Furthermore, we can see that our method (Ridge.QP) is far better than the rest. The only exception seems to be MLR, which performs slightly better than Ridge.QP for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. However, when looking at $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$, as well as the overall RMSE, our method is superior to all the others. Of course, our interest lies mainly in the estimation of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, which represent the crops of interest, and so the results we get are still promising, especially considering that MLR is unable to provide spatially varying coefficients estimate.

Finally, Figure 3.6 shows the correlations between $\hat{\boldsymbol{\beta}}_j$ and $\boldsymbol{\beta}_j$ for all considered methods. Surprisingly, most methods have correlations around 0 and thus seem unable to pick out a consistent direction in the estimation. The exceptions are Ridge.QP and to a lesser extent the Bayesian method, but only for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, where they seem to consistently have a positive correlation between the estimates and true values.

Overall, this simulation study would suggest that Ridge.QP is able to perform the best when estimating all coefficients, both in terms of prediction error and correlation. The estimations for $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$ are not very good overall for any method, but this is fine as both of them represent land types that are not of major interest.

If we take a look at just one simulation result (Figure 3.7), we can see that our method lies very close to the identity line between the estimates and true values, meaning that we create a faithful reconstruction of the field. The only other method that seems similar is the Bayesian method. We

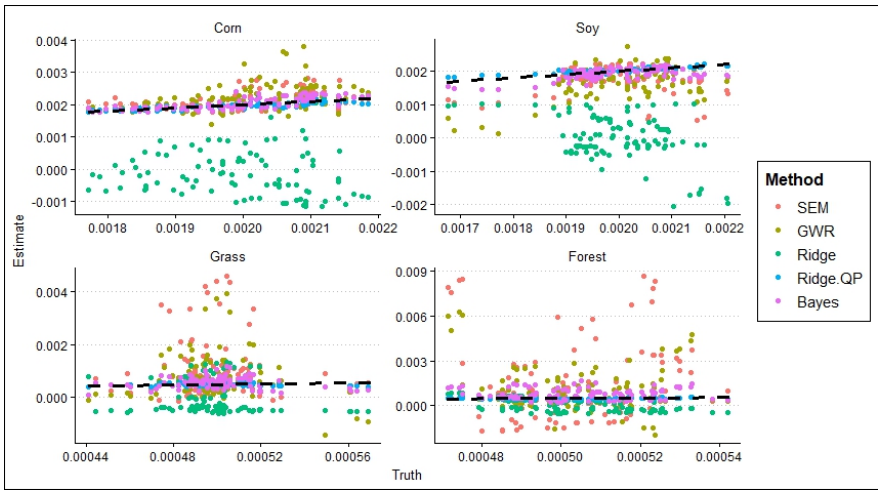


Figure 3.7: Scatterplot for one simulation between $\boldsymbol{\beta}_j$ and $\hat{\boldsymbol{\beta}}_j$, for each land type, showing the results from all methods except SCC. The dashed line represents the identity line.

can also see how the correlations between estimates and true values are also low, in particular with $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$. For similar reasons to those discussed above, SCC is removed from the plots to be able to appreciate the differences among all other methods.

3.5 Estimation of Solar Induced Fluorescence yield

There are 112 days of measurements taken, ranging from May to September. The number of footprints per day ranges from just 1 observation to 110. Before doing any analysis, all days with 3 or fewer observations were also excluded from the analysis. This reduced the number of days with data to analyze to 102. The days are modeled independently of each other. Some of the days analyzed had a lot of sites where $y(\boldsymbol{s}) = 0$, making it so that the MST was not used for inducing similarities and instead, the coefficients left to estimate were estimated with just a Ridge penalty.

The penalty parameter λ was tuned using a sequence of 1000 values from e^{-3} to e^{25}. To further ease the tuning, each covariate was scaled by its standard deviation, and the resulting coefficients were then transformed back to the original scale.

To efficiently find the MST, we first obtain a Delaunay triangulation of the sites for a given day. Then we find the MST (with Euclidean distance as the edge weight) within this triangulation rather than using the whole connected graph. This allows for a much more efficient and quicker calculation of the MST. To solve the quadratic programming problem we used the package 'quadprog' in R

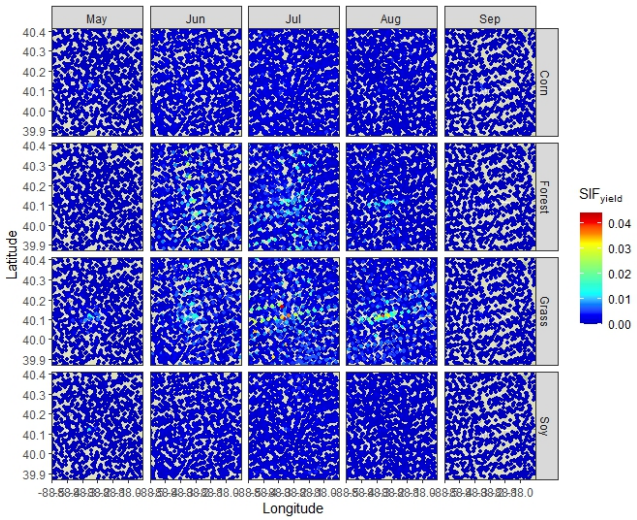


Figure 3.8: Estimated SIF yields for each land type, per month

(Cleve Moler dpodi/LINPACK) 2019), which utilizes the dual method from Goldfarb and Idnani 1982; Goldfarb and Idnani 1983.

The estimated SIF yields for each land type, for each month, are presented in Figure 3.8. The results in this plot are difficult to observe due to the presence of some high estimations of SIF yield for grass and forest lands during the months of Summer, especially in July. Nonetheless, we can observe that the areas of high SIF yield for grass and forest correspond to the center of the region which is mostly urban. This makes sense given that only near the urban areas of Urbana and Champaign can we find parks with this kind of foliage, the remaining surrounding areas mostly consist of farmlands.

The results are truncated in Figure 3.9 to show only estimations less than 0.002 (a range comparable to the results obtained in Wang et al. 2020). With this reduced range it is easier to perceive a seasonal trend in the estimation where SIF yield increases from the end of spring to summer and then decreases as summer transitions into fall. A spatial trend can be seen with grass and forest where during May and September there is a concentration of high yields near the urban areas. However, the yield estimations for corn and soy do not seem to possess a clear spatial trend, possibly due to most of the area containing farmland.

The temporal trend can be seen more clearly in Figure 3.10, where all land types share the same pattern previously discussed. The effect seems stronger for grass and forest, but this is primarily due to the high estimations during the summer months (hence why the shaded region representing

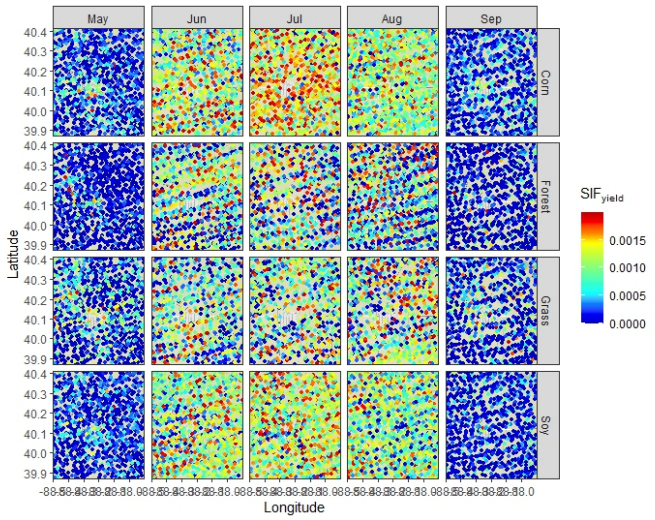


Figure 3.9: Estimated SIF yields for each land type, per month - truncated only to small values

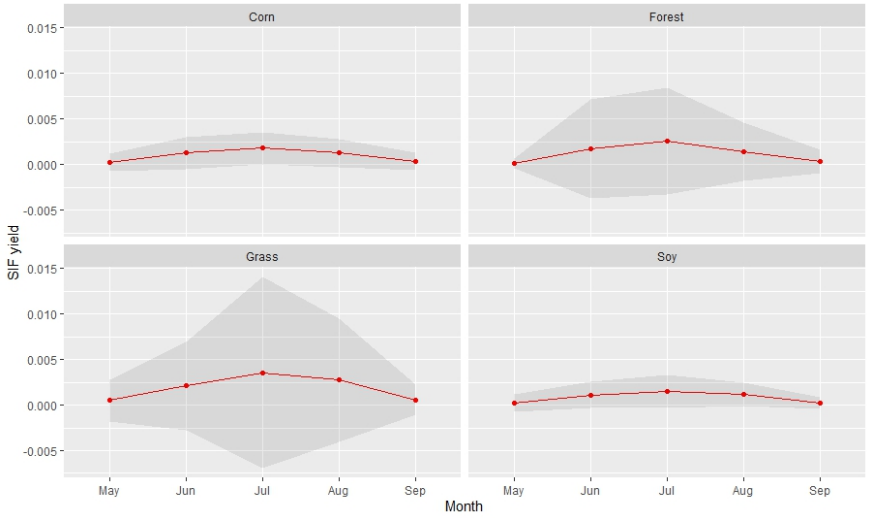


Figure 3.10: Median SIF yield, in red, for each land type, per month. The shaded area represents one standard deviation

one standard deviation is also much larger). Conversely, the shaded region for corn and soy looks much more uniform, suggesting that there are no major spatial areas where corn and soy produce higher or lower SIF yields.

3.6 Conclusions

Using the decomposition of SIF proposed by Wang et al. 2020 and the relationship to crop yield given by Guan et al. 2016, we are able to use satellite spectrometry for large-scale crop monitoring. However, the means to decompose SIF into vegetation-specific SIF yields that then link to crop yield have been very time-consuming. For this, we have proposed an alternative method to the ones currently used that is fast while also producing accurate estimates that respect the SIF decomposition. This leads to a spatially varying coefficient model where said coefficients represent vegetation-specific SIF yields of interest.

Our method is an alternative to Li and Sang 2019, where we use a fused Ridge penalty instead of a fused Lasso penalty to induce spatial smoothness in the estimations. Furthermore, estimates have to be strictly non-negative given their interpretation, which adds a further constraint to our estimation problem. Since the fused Ridge penalty adds a penalty parameter that must be tuned, this is done using AIC with an approximation to calculate the degrees of freedom under the additional non-negativity constraint.

Simulations show that our method is suitable when recovering information about the coefficients relating to corn and soy, the two most important coefficients considered. This is also true when comparing against other spatially varying coefficient models. Nonetheless, as the simulations show, the underlying data has a very weak signal, and thus estimation overall is particularly difficult. However, the simulated data faithfully tried recreating the relationships of the actual observed data, and even under these conditions, our method is able to retrieve information about corn and soy fields.

Results show that SIF yield increases from spring to summer and then decreases from summer to fall for all four vegetation types. This is concurrent with expectations on how vegetation works during the summer. Results for forest and grass are particularly noisy, but this is not unexpected given how low their signal is compared to corn and soy. The latter two vegetation land types

produce results comparable to previous studies. Nonetheless, there doesn't seem to be a strong spatial relationship among any of the four land types. However, this could possibly be due to the region of study, since Champaign-Urbana is known for consisting mostly of farmlands dedicated to corn and soy.

Overall, the method proposed in this chapter produces good estimations for the SIF yields of the two vegetation types of interest. Nevertheless, it would be interesting to observe how this model behaves in areas that have more representation of forest and grass and are not mostly dedicated to farmlands. Additionally, results show that the temporal trend in the estimations is respected, but it would be of future interest to include a temporal component in the model that respects the SIF decomposition and also produces results in real time.

Chapter 4

Data Fusion of Temperature Datasets Using INLA

4.1 Introduction

A very classic problem in spatial statistics is combining information that is gathered at differing spatial resolutions. this happens due to the advances in remote sensing and satellite imagery which allows for data to be gathered at different resolutions. This is normally referred to as 'areal' data, since the information retrieved is aggregated over a particular area given by the resolution of the measuring instrument.

Opposed to this is another classical type of spatial data, referred to as geostatistical data. This is considered a spatial process indexed over a continuous space. This information is typically gathered by instruments that are fixed in a particular known location for a given time, like weather stations, weather balloons, etc. For the remainder of this work we will refer to this data as 'point' data, in contrast to areal data. Nonetheless, this does not mean that we are talking about point data in the classical spatial statistics sense, where the location is random.

When there is a mismatch or inconsistency between the spatial units at which data are observed or measured and the spatial units at which the analysis or inference is desired, we are presented with what is known as the change of support problem. This problem is particularly relevant in the field of geostatistics, where spatial data are commonly collected at one set of locations or spatial

resolutions but are needed or desired at a different set of locations or spatial resolutions for modeling, prediction, or decision-making purposes.

One particular case of the change of support problem that is very common to environmental sciences is trying to combine information from different resolutions. Typically, we are interested in using both areal and point data to estimate a process at a smaller resolution. This is the main goal of this project, where we try to combine the areal data concerning surface temperatures obtained from several instruments on board of NOAA satellites and the point data from a collection of historical and near-real-time radiosonde and pilot balloon observations.

The main goal of combining these data sets is to produce a more complete temperature map of the entire world. The point data is mostly in land and in the northern hemisphere. Areal data is collected throughout the globe, but with a great quantity of missing areas due to cloud coverage and other issues that arise from satellite imaging. This leaves an incomplete picture of global temperature that we hope to complete by combining both data sets in a modeling framework.

This work is part of a bigger project that seeks to create a complete data set of world surface temperature from 1990-1993 for use in other research projects. Typically, reanalysis data is used to fulfill this need, but it either lacks uncertainty calculations or they take too much time to compute. The goal of this project is to find a method that is able to interpolate the observed data to the entire globe in a time-efficient manner and that also takes into account the differences in handling areal and point data. This project intends to do that by looking at how INLA is able to treat areal and point data differently.

The following project is divided into several sections. Section 4.2 introduces where the observed areal and point data comes from, as well as the sampling limitations that are present in them that challenge inference. Section 4.3 discusses how areal and point data are modeled differently and how INLA respects this distinction, as well as an alternative hierarchical model approach that is faster to fit but loses accuracy. Section 4.4 shows the results of several simulations to study how different sampling scenarios affect the predictions, as well as how much the prediction is affected by using the alternative model. Finally, results are shown in Section 4.5 with a discussion made in Section 4.6.

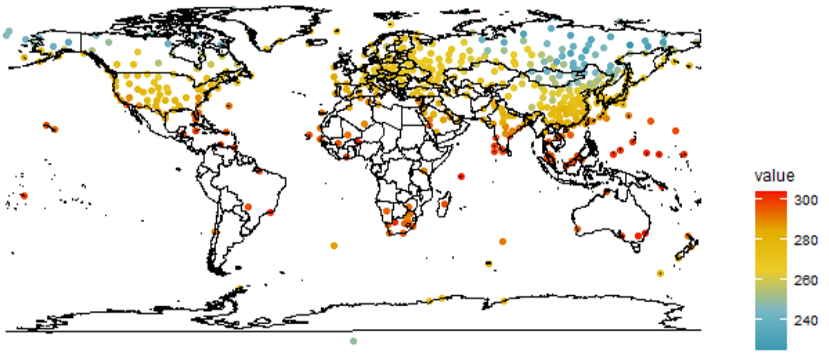


Figure 4.1: Point temperature data obtained from NOAA's Integrated Global Radiosonode Archive (IGRA) for January 1st, 1990.

4.2 Data

Point and areal world surface temperature data are collected daily from 1990 to 1993. For this particular project, we will be mostly focusing on January 1st, 1990. Point data is obtained from the Integrated Global Radiosonde Archive (IGRA), which is provided by the National Oceanic and Atmospheric Administration (NOAA) (Durre et al. 2016). This data consists of radiosonde and pilot balloon observations from more than 2,800 globally distributed stations that date back to 1905. As seen in one example for January 1st, 1990 in Figure 4.1, most of the data is located in land and in the northern hemisphere. This holds true for the remaining days in the study, which posses an interesting question in how well the data will mix in those regions where we have less or no observations.

Areal data is recovered from NOAA's TIROS Operational Vertical Sounder (TOVS), which is a suite of three instruments that were flown on the NOAA-6 through NOAA-14 Polar-orbiting Operational Environmental Satellites. These instruments were primarily designed for atmospheric sounding are sensitive to surface temperatures (Anyamba and Susskind 1998). Data is recovered in blocks of 1° by 1°, meaning that it creates a grid of size 180×360.

Figure 4.2 displays the areal temperature data retrieved by TOVS. The gray blocks are those were cloud cover made it impossible for the instruments to obtain a reasonable temperature reading. For this particular day, the amount of missing data represented 52.28% of the grid. However, this percentage is variable for other days and is not always as high. Nonetheless, this is the main reason why we must find a way to combine this data with point data to produce a more complete snapshot

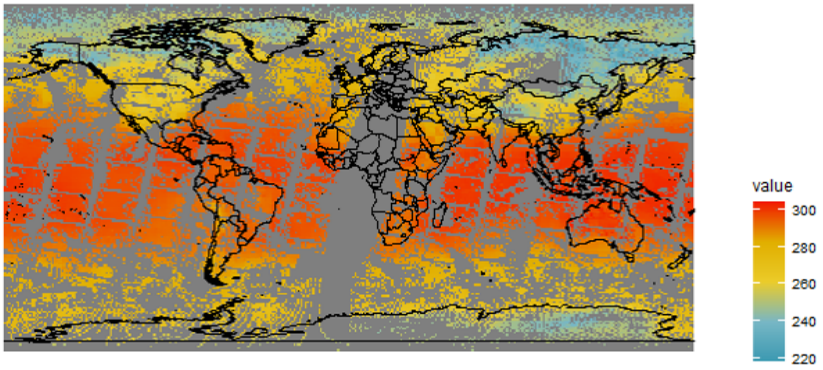


Figure 4.2: Areal temperature data obtained from NOAA's TIROS Operational Vertical Sounder (TOVS) for January 1st, 1990. The gray areas correspond to missing data due to cloud cover.

of daily surface temperature.

4.3 Models and Estimation

We suppose we are studying an underlying spatial process from which we observe continuous observations with some measuring error. Let the underlying spatial process be $S(\boldsymbol{x})$, $\boldsymbol{x} \in D \subset \mathbb{R}^2$ where we assume that $S(\boldsymbol{x})$ is a mean-zero Gaussian process with some covariance function. For a finite set of sites, say $\boldsymbol{x}_i \in D$, $i = 1, 2, ..., n_p$, we model the point data as:

$$y(\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) + S(\boldsymbol{x}_i) + \varepsilon(\boldsymbol{x}_i) \tag{4.1}$$

where $\varepsilon(\boldsymbol{x}_i) \sim N(0, \sigma^2)$ independently for each $i = 1, 2, ..., n_p$.

Areal data is defined as averages in a block $A_j \subset D$, $j = 1, 2, ..., n_a$, such that

$$y(A_j) = \frac{1}{|A_j|} \int_{A_j} (\mu(\boldsymbol{x}) + S(\boldsymbol{x}))dx \tag{4.2}$$

where $|A_j| > 0$ denotes the area of block A_j. For this project, we will assume that all blocks are observed on a regular grid, and thus have the same shape and area.

4.3.1 Estimation with INLA

The models are fit using the INLA approach (Rue, Martino, and Chopin 2009) with the SPDE approach (Lindgren, Rue, and Lindström 2011), following the handling of the change of support

problem portrayed in Moraga et al. 2017. All of this can be easily applied using the R package R-INLA (Lindgren and Rue 2015).

As mentioned in Chapter 2, INLA uses a combination of analytical approximations and numerical integration to do Bayesian inference in models that have a latent Gaussian process. Furthermore, using the SPDE approach, modeling can be done in continuous space, but the inference uses the sparse precision matrices of the GMRF defined on the triangular mesh of the spatial domain. The representation of the continuously indexed field, $S(\boldsymbol{x})$, through the discretely indexed GMRF is done by means of a finite basis function defined on the mesh:

$$S(\boldsymbol{x}) = \sum_{m=1}^{M} \phi_m(\boldsymbol{x}) S_m \tag{4.3}$$

where S_m are mean-zero Gaussian weights, $\phi_m(\cdot)$ denotes a piecewise polynomial basis function on each triangle, and M is the number of vertices in the mesh.

The way R-INLA approximates $S(\boldsymbol{x})$ for point data is by calculating the weighted mean of the GMRF estimates in the vertices of the triangle that includes $\boldsymbol{x}$. The weights are given by the barycentric coordinates, i.e. they are proportional to the areas of each of the three subtriangles defined by the point $\boldsymbol{x}$ and the vertices of the triangles. Areal data, $S(A)$ is estimated by taking the average of all the GMRF values in block A. This makes it necessary for each areal block to include at least one vertex of the mesh.

Since INLA uses the SPDE approach, this means that the only stationary covariance function available to fit models is the Matérn covariance function, as defined in Chapter 1. However, this approach allows for a more flexible class of non-stationary models. The details of this can be found in Lindgren, Rue, and Lindström 2011, which explains how it is possible to define the non-stationarity in the SPDE instead of in the covariance function itself. For this model we allow the variance to vary with latitude, such that the variance increases in a polynomial way as we move farther away from the equator.

Another advantage of using the SPDE approach is the gain in computation time by using a sparse precision matrix but also thanks to the dimension reduction obtained by only having to do the estimation in the triangle vertices. Normally, we would expect to have fewer nodes in the triangulation than data. However, this is not always the case when working with areal data. Since

we must consider that each block must have at least one vertex in the triangulation, this means that there is actually a dimension increase when fitting the GMRF. This is typically not an issue as we are still working with sparse precision matrices, unless the resolution is very large (i.e. small areal block) in which case computation time will be affected.

4.3.2 Alternative model

To handle the larger computation times encountered by having to create a very dense mesh we have to handle the model differently to be able to use a sparse mesh. A very simple alternative is to just use the areal data as if it were point data, but obviously this doesn't account for the change of support. This would mean that the uncertainty specific to areal data would be underestimated.

One way to respect the different uncertainties for point and areal data but also create a computationally faster method is to suppose a hierarchical model structure where both areal and point data are treated as observations from a common latent field, but with different uncertainty structures.

The first level consists of modeling both data sets as realizations of the same common latent field, $\mu(x) + S(\boldsymbol{x})$, but with independent and different error structures:

$$\begin{aligned} Y_p(\boldsymbol{x}) &= \mu(\boldsymbol{x}) + S(\boldsymbol{x}) + \epsilon_p, \quad \epsilon_p \sim N(0, \sigma_p^2), \\ Y_a(\boldsymbol{x}) &= \mu(\boldsymbol{x}) + S(\boldsymbol{x}) + \epsilon_a, \quad \epsilon_a \sim N(0, \sigma_a^2), \end{aligned} \tag{4.4}$$

In comparison to equations (4.1) and (4.2), point data is modeled the same way but areal data is now modeled as point data but with a different error structure that hopefully is able to capture the different uncertainty that is characteristic of this type of data.

For simplicity, we will model $\mu(\boldsymbol{x})$ as a constant μ, but it could be interesting in the future to include some more information that depends on $\boldsymbol{x}$. However, as for now, we do not have any more insight into how this mean structure could look like.

The second level of this hierarchical model describes the latent field, $S(\boldsymbol{x})$, as mean zero Gaussian process with some covariance function:

$$S(\boldsymbol{x}) \sim GP(0, \Sigma) \tag{4.5}$$

The covariance function is modeled using a fixed smoothness parameter, $\nu = 1$, and two unknown parameters for the spatially varying variance, $\sigma^2(\boldsymbol{x})$, and stationary range parameter, ρ. The variance is modeled as $\log(\sigma(\boldsymbol{x})) = \theta_1 + \theta_2 \cdot \text{scale}(x_2)^2$, where x_2 refers to latitude and the scale$(\cdot)$ function centers the latitude around it's mean and scales it to have a variance of 1.

Finally, the third level of the model would be given by the prior distributions on the parameters:

$$\begin{aligned} \mu &\sim N(0, 1000), \\ \log(1/\sigma_p^2), log(1/\sigma_a^2) &\sim \text{LogGamma}(1, 0.00005), \\ (\theta_1, \theta_2, \log\rho)^T &\sim N\left((0, 0, \log(142))^T, 100 \cdot \boldsymbol{I}_3\right). \end{aligned}$$

The log range parameter of the latent field is centered at $\log(142)$ since it represents about a third of the maximum distance of the spatial domain. Nonetheless, the variance is kept high to still be considered a mostly uninformative prior like all the others used.

4.4 Simulation Study

The simulation study is divided into two sections. The first one plans to evaluate how well INLA is able to combine point and areal data under different sampling scenarios and with two different cases for areal and point sample sizes.

The second part studies how much prediction error is incurred when using the alternative hierarchical model instead of the original formulation. Recall that this model is preferred for computation purposes, even though it is known that areal data is not being handled appropriately. We know that a very fine triangulation of the region will produce better results, but at a higher computational cost. To be able to use a more sparse mesh we need to forgo the requirement that each areal block must include at least one vertex, and that's why we use the hierarchical model formulation instead.

4.4.1 Sampling Scenarios

The main objective of this simulation study is to assess how well INLA is able to combine both point and areal data under different data sampling scenarios and with different sample sizes. For this, 100 'true' fields are simulated from mean-zero Gaussian process with Matérn covariance function.

The covariance function uses $\nu = 1, \sigma^2 = 1$ and $\rho = 142$. For simplicity in the simulation, the mean function is taken to be fixed at $\mu = 0$. Each field is generated on a 100×200 regular grid, producing 20,000 true points.

This model will eventually be fitted for measuring global temperatures, which are values vastly different from what we're simulating, but for this simulation study we are more interested in seeing the effects of how the data is sampled, as this is the most important characteristic of the real data.

We will study five different sampling scenarios under two different ways of splitting the point and areal sample sizes; the sampling scenarios are summarized in Table 4.1. Point data is sampled at random from the original field in one of two cases: either fully at random over the entire domain or constrained inside land masses. When sampled inside land masses, they are further sampled such that the northern latitudes have more points than the southern ones. This replicates the pattern observed in the real data where the point data is mostly in land and with more prevalence in Northern America and Europe. In either case, point data is also sampled with an added measurement error as seen in Equation (4.1).

Scenario	Areal Data	Point Data
1	All data	Sampled at random
2	All data	Sampled in land
3	Missing data at random	Sampled at random
4	Missing data in stripes	Sampled at random
5	Missing data in stripes	Sampled in land

Table 4.1: Different sampling scenarios for the areal and point data. Areal data can either be fully present, or some cells might be missing at random or in stripped patterns. Point data can either be sampled at random over the entire domain, or constrained to only land masses, with more prevalence in northern latitudes.

In the case of areal data, we generate a 90×180 grid, where each block averages 1,2 or 4 points from the original field, following the form of (4.2). When 'sampling' areal data we're actually considering how we're sampling blocks to be considered as missing data. Scenarios 1 and 2 use the whole grid of areal data. Scenario 3 samples the missing blocks at random over the entire spatial domain. Scenarios 4 and 5 sample the missing blocks inside 'stripes'. These stripes are generated using a sinusoidal function on longitude and latitude that is tuned to produce diagonal vertical stripes (as seen in Figure 4.3). The blocks that are removed correspond to the highest values of the sine wave, according to another user-tuned threshold. All the scenarios that have missing data have

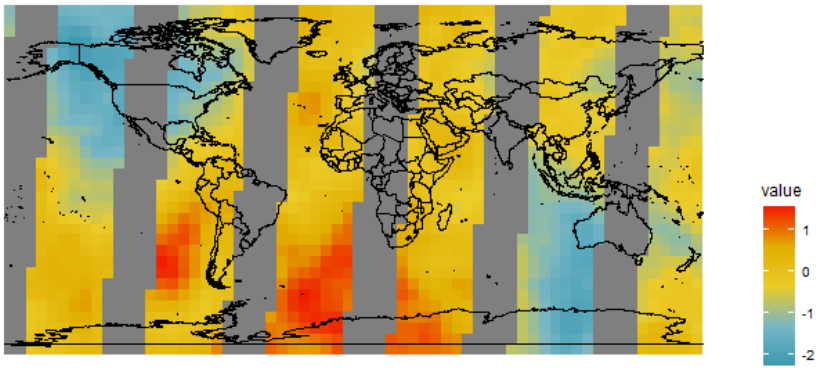


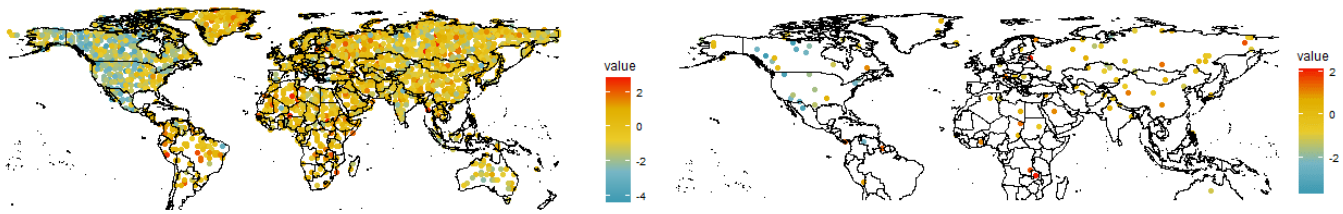
Figure 4.3: Example of a simulated areal dataset with missing blocks in a vertically diagonal stripe pattern.

the same amount of removed data.

The sample sizes are considered in two different splits of how much point data we want to have in comparison to areal data. Areal data is always kept fixed at 16,200 blocks (without missing data). As for point data, one case considers having the same number of point data as areal data, whereas the other case is more similar to the real data and considers that point data represents about 5% of the total data; these splits are referred to as the '50-50' and '95-05' splits, respectively. Figure 4.4 shows an example of point data sampled in land under both sample size splits.

The sampling scenario that resembles the most the real data is Scenario 5 under the 95-05 split. All other Scenarios are considered as they represent 'better' versions of what we observe in the real data and we wish to compare how the predictions under Scenario 5 compare to the rest. Similarly, the 50-50 split also represents a more optimistic situation where we have equal amounts of data from both sources.

The model that properly treats areal data as such was fitted for each combination of sampling scenario and sample size split. The metrics observed for prediction accuracy were the root mean square error and the correlation between predicted values and true values at the original 100×200 resolution. The predicted values consist of the posterior means of the fitted value. RMSE results are shown in Figure 4.6, whereas the correlation results are presented in Figure 4.6.



(a) (b)

Figure 4.4: Point data sampled in land, with more data in northern latitudes, under the (a) 50-50 split and the (b) 95-05 split of areal to point data sample sizes.

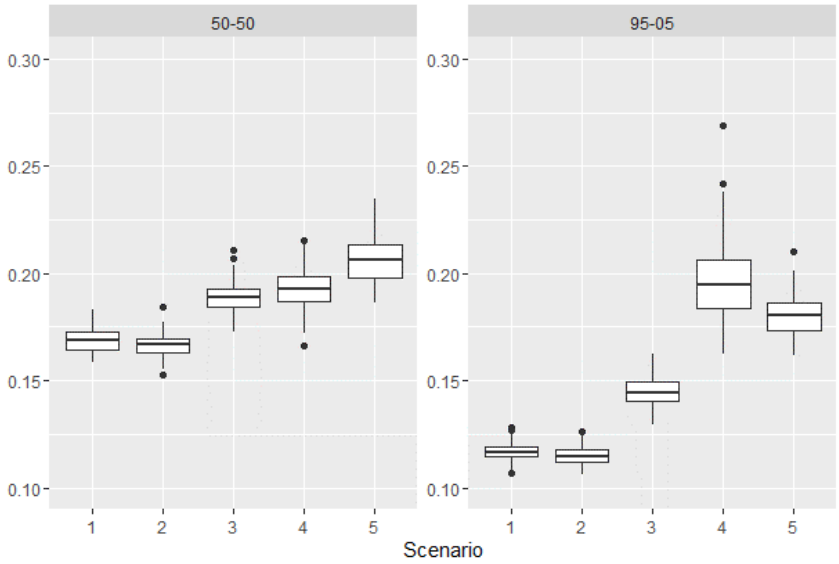


Figure 4.5: Root mean square error between the true and predicted values of $S(\boldsymbol{x})$ for each of the five different sampling scenarios and the two sample size splits.

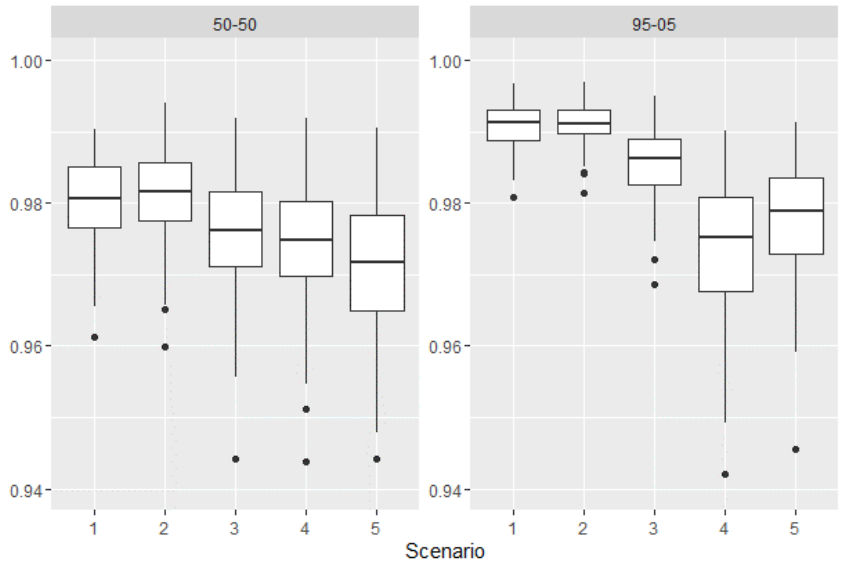
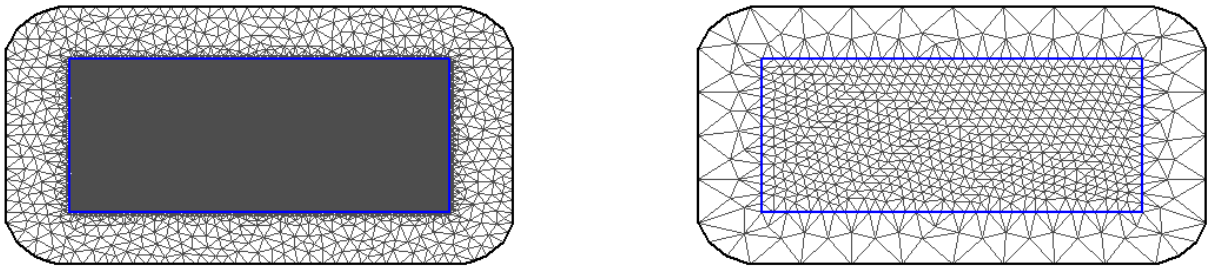


Figure 4.6: Correlation between the true and predicted values of $S(\boldsymbol{x})$ for each of the five different sampling scenarios and the two sample size splits.

From both plots we can observe that high correlations relate to low RMSEs and vice-versa. In general, we observe that results in the 95-05 split are better than in the 50-50 split, which at first seems counter-intuitive as one would expect more data to be better for the analysis, but we must consider that the only source of noise in these simulations come from point data. Further simulations with reduced error in the point data (not shown) indeed show under the presence of missing areal data, predictions are better for the 50-50 split than the 95-05 split. Furthermore, under the presence of all the areal data, both splits coincide that having point data throughout the entire region is better than clustered in land.

When introducing missing blocks in areal data, it is clear that having them be missing at random (i.e. Scenario 3) is better than having them missing in the striped pattern (Scenarios 4 and 5). As for the sampling of point data with striped areal missing data, the results show an interesting reversal between the different splits. Under the 50-50 split, having points sampled in the whole domain is better than just sampling them in land, but the reverse is true for the 95-05 split. This once again seems to be due to noise being present only in point data, since when this noise is removed (not shown), the trend of the 50-50 split is also observed in the 95-05 split.

Overall, the simulation study seems that suggest that under the presence of noisy point data (or at least noisier than areal data), having less point data is better, irregardless of whether these



(a) (b)

Figure 4.7: Delaunay triangulation of the study region (inside the blue polygon) using a (a) dense mesh and a (b) sparse mesh. The dense mesh has 18066 vertices, whereas the sparse mesh consists of only 708, much less than the total sample size (16302).

points are sampled in the whole domain or just in land. Nevertheless, it is clear from both splits that the best case scenario would be having a full areal data set, but in practice this is not feasible.

4.4.2 Using the hierarchical model

This study aims to quantify the loss in prediction accuracy when using the hierarchical model (4.4), as opposed to the original formulation in (4.1) and (4.2). Recall that the hierarchical model is known to not be correct, but can be fitted using a much sparser mesh and thus will be much faster to compute. Other simulations done (not shown) revealed that as the number of areal blocks increased, it didn't matter if they were treated as areal or point data, as long as the mesh was kept the same. This study hopes to asses how a sparser mesh would affect the predictions, with using the alternative method as a counterbalance to just treating areal observations as simple point data.

For this simulation we use only one of the simulated fields and sample point and areal data in a manner equivalent to Scenario 2 under the 95-05 sample size split. We then proceed to fit the model using both formulations, with the second formulation using a much sparser mesh. Both meshes can be seen in Figure 4.7, where it is very difficult to observe the small triangles inside the study domain (delineated by the blue line) for the dense mesh.

The estimated fields are shown in Figure 4.8. Visually, there is almost no difference between

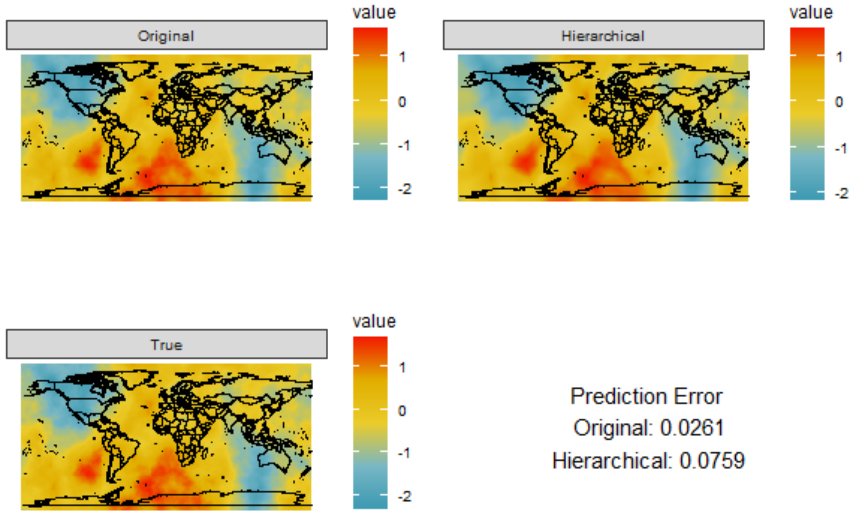


Figure 4.8: Estimated fields for the original model formulation that accurately treats areal data and for the hierarchical model formulation that treats areal as point data with a different error structure. The true field and the prediction errors (RMSE) are also included for reference.

both estimated fields and the truth, except for some smoothing in the results. However, the root mean square error between the true values and the estimated ones are almost three times as high when treating areal data as point data in the alternative model, than when treating it appropriately in the original formulation. Nonetheless, the error is still relatively low.

In terms of correlation, the correlation between the estimation and the true field for the original formulation was 0.999, whereas the correlation with the alternative model was 0.995, which is not a terrible loss. Nonetheless, the computation time for the original method was around 35 minutes, whereas the alternative model took about 8 seconds, which is a drastic difference.

Figure 4.9 shows the posterior standard deviations for the fitted values under the two different modeling approaches. As seen in the hierarchical model plot, there are spots scattered around the domain that correspond to higher standard deviations. These lie close to the vertices of the sparse mesh. Overall, the mean standard deviation for the original formulation is about 0.002, whereas for the alternative model it is 0.014.

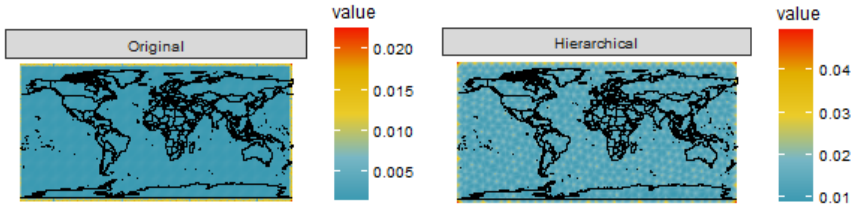


Figure 4.9: Posterior standard deviations for each fitted value for the original model formulation as well as the alternative hierarchical model.

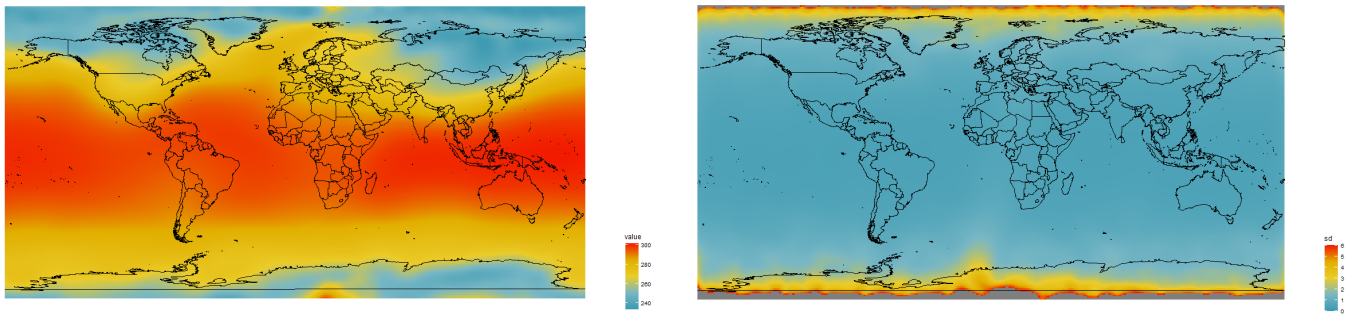
4.5 Interpolation of global temperature

We will only present the results pertaining to January 1st, 1990 here. The data is fitted using both the original formulation in equations (4.1) and (4.2), as well as the modified hierarchical model (4.4). The original formulation uses a dense mesh, similar to the one in Figure 4.7a, whereas the alternative model uses the same sparse mesh in Figure 4.7b.

Results for the model that uses the dense mesh are presented in Figure 4.10. The estimates of the field are the posterior means of the fitted values at the centroids of each areal block and they are given in Kelvin. This model took approximately 45 minutes to run. Figure 4.11 shows the results for the alternative model with the sparse mesh. These results were obtained in about 30 seconds.

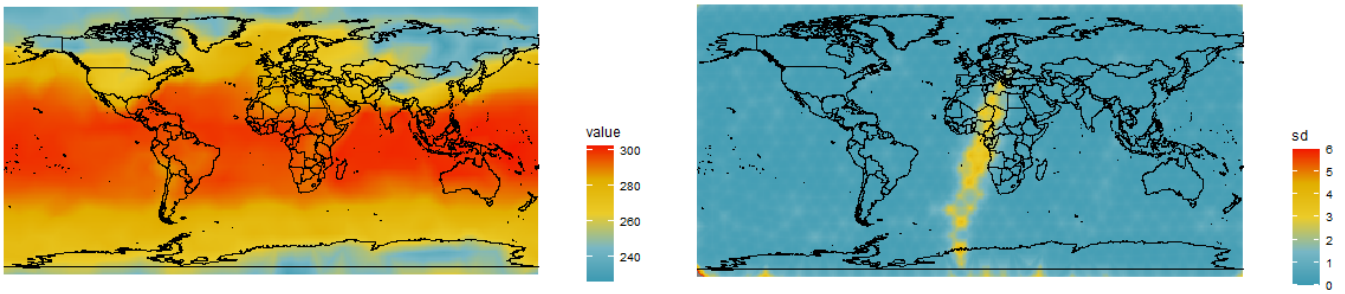
Both estimated fields look very similar in general terms. They seem to do well appropriately capturing the regions of cold and hot climate for this particular time of the year. Interestingly, the original model seems to estimate a small region of hot temperatures in Antarctica that is unexpected.

The difference in meshes is more noticeable when looking at the posterior standard deviations. Standard deviations in Figure 4.10b are overall smaller than those in 4.11b. Nevertheless, there is higher uncertainty in regions where areal data was missing (although, due to the presence of high values in the border, this is difficult to see in the results for the original model). Figure 4.11b also



(a) (b)

Figure 4.10: Results obtained by treating point and areal data accordingly. The field estimate (a) correspond to the posterior means and the standard deviation (b) to the posterior standard deviations at each location.



(a) (b)

Figure 4.11: Results obtained by treating point and areal data as point data but with different error structures. The field estimate (a) correspond to the posterior means and the standard deviation (b) to the posterior standard deviations at each location.

shows small regions of high uncertainty scattered around the map that correspond to the vertices of the sparse mesh.

Both results show higher uncertainty around the border. This is a common problem with the SPDE approach that is mitigated by having larger triangles outside the study region. However, this is still difficult to mitigate when the inner triangles are very small, as the outer triangles grow in size starting from the size of the inner ones. This means that the triangles that are closest to the mesh are still very small (as seen in Figure 4.7a) and thus produce higher border uncertainties.

4.6 Conclusion

INLA is capable of accurately combining areal and point data through the SPDE approach. However, this is heavily dependent on the mesh, as areal data requires at least one mesh vertex inside of each areal block. This becomes a issue when areal blocks become smaller as computations take much longer. We propose an alternative hierarchical model that relates areal and point data to the same latent process but allows flexibility in modeling the error of each type of data. This allows the GMRF to be fit in a more sparse network of vertices and drastically reduce computation times.

This project presents an interesting question moving forward. Are we more interested in the higher accuracy by treating areal and point data accordingly, at the cost of longer computation times, or can we sacrifice some of that accuracy to have the results out faster? Simulations show that there is a loss in accuracy by using the faster method, but this loss is decreased as the number of areal blocks become larger.

Through a simulation study that changes the way areal and point data are obtained we have seen that our results would be benefited from having less missing areas coming from TOVS. They also seem to suggest that point data is not as essential as areal data, but this is only true if point data is particularly noisy. As radiosonde and weather balloons become better measuring instruments we can assume that measurement error would be decreased and their presence would help with the final predictions much more.

There is one major problem when using this method to interpolate world data, being that the mesh is constructed in a 2D surface, as opposed to a sphere. Lindgren, Rue, and Lindström 2011 shows that it is possible to create a mesh in a sphere, but as far as we know this is still not

implemented into R-INLA. This means that the estimated covariance structure is dependent on the projection of the world we use. This is most likely the reason why regions near each of the poles seem to have higher temperatures than expected, as the covariance structure identifies these regions to be farther away than what they should be.

As previously mentioned, this is part of a bigger project that is trying to find the best way to handle the change of support problem in terms of both accuracy and computation time when it comes to integrating observed datasets. Additionally, these methods should also be able to provide some quantification of prediction error, which in this case can be handled by the posterior standard deviations. So far, the results for world temperature are promising and sensible. In the future, we would like to apply a similar methodology to aerosol data, which pose an extra set of challenges with how that data was recovered in the early 90s.

References

Abbott, Richard J and Christian Brochmann (2003). "History and evolution of the arctic flora: in the footsteps of Eric Hultén". In: *Molecular ecology* 12.2, pp. 299–313.

Allouche, Omri, Asaf Tsoar, and Ronen Kadmon (2006). "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)". In: *Journal of Applied Ecology* 43.6, 1223–1232. DOI: 10.1111/j.1365-2664.2006.01214.x.

Anderson, Lynn L, Feng Sheng Hu, and Ken N Paige (2011). "Phylogeographic history of white spruce during the last glacial maximum: uncovering cryptic refugia". In: *Journal of Heredity* 102.2, pp. 207–216.

Anderson, Lynn L et al. (2006). "Ice-age endurance: DNA evidence of a white spruce refugium in Alaska". In: *Proceedings of the National Academy of Sciences* 103.33, pp. 12447–12450.

Anderson, Patricia M and Linda B Brubaker (1994). "Vegetation history of northcentral Alaska: a mapped summary of late-Quaternary pollen data". In: *Quaternary Science Reviews* 13.1, pp. 71–92.

Anyamba, Ebby and Joel Susskind (1998). "A comparison of TOVS ocean skin and surface air temperatures with other data sets". In: *Journal of Geophysical Research: Oceans* 103.C5, pp. 10489–10511.

Aoki, K et al. (2019). "Approximate Bayesian computation analysis of EST-associated microsatellites indicates that the broadleaved evergreen tree Castanopsis sieboldii survived the Last Glacial Maximum in multiple refugia in Japan". In: *Heredity* 122.3, pp. 326–340.

Bakka, Haakon et al. (2018). "Spatial modeling with R-INLA: A review". In: *WIREs Computational Statistics* 10.6, e1443. DOI: https://doi.org/10.1002/wics.1443. eprint: https:

//onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1443. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1443.

Bigelow, Nancy H. et al. (2003). "Climate change and Arctic ecosystems: 1. Vegetation changes north of 55°N between the last glacial maximum, mid-Holocene, and present". In: *Journal of Geophysical Research: Atmospheres* 108.D19. DOI: https://doi.org/10.1029/2002JD002558. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002JD002558. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD002558.

Bivand, Roger and Danlin Yu (2022). *spgwr: Geographically Weighted Regression*. R package version 0.6-35. URL: https://CRAN.R-project.org/package=spgwr.

Botkin, Daniel B. et al. (Mar. 2007). "Forecasting the Effects of Global Warming on Biodiversity". In: *BioScience* 57.3, pp. 227–236. ISSN: 0006-3568. DOI: 10.1641/B570306. eprint: https://academic.oup.com/bioscience/article-pdf/57/3/227/26898698/57-3-227.pdf. URL: https://doi.org/10.1641/B570306.

Brown, Jason L and L Lacey Knowles (2012). "Spatially explicit models of dynamic histories: examination of the genetic consequences of Pleistocene glaciation and recent climate change on the American Pika". In: *Molecular Ecology* 21.15, pp. 3757–3775.

Brubaker, Linda B. et al. (2005). "Beringia as a glacial refugium for boreal trees and shrubs: new perspectives from mapped pollen data". In: *Journal of Biogeography* 32.5, pp. 833–848. DOI: https://doi.org/10.1111/j.1365-2699.2004.01203.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2699.2004.01203.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2699.2004.01203.x.

Brunsdon, Chris, A Stewart Fotheringham, and Martin E Charlton (1996). "Geographically weighted regression: a method for exploring spatial nonstationarity". In: *Geographical analysis* 28.4, pp. 281–298.

Brunsdon, Chris, Stewart Fotheringham, and Martin Charlton (1998). "Geographically weighted regression". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.3, pp. 431–443.

Budde, KB et al. (2013). "The ancient tropical rainforest tree Symphonia globulifera L. f.(Clusiaceae) was not restricted to postulated Pleistocene refugia in Atlantic Equatorial Africa". In: *Heredity* 111.1, pp. 66–76.

Casetti, Emilio (1972). "Generating models by the expansion method: applications to geographical research". In: *Geographical analysis* 4.1, pp. 81–91.

— (1997). "The expansion method, mathematical modeling, and spatial econometrics". In: *International regional science review* 20.1-2, pp. 9–33.

Clark, James S (2005). "Why environmental scientists are becoming Bayesians". In: *Ecology letters* 8.1, pp. 2–14.

Cleve Moler dpodi/LINPACK), S original by Berwin A. Turlach R port by Andreas Weingessel ¡Andreas.Weingessel@ci.tuwien.ac.at¿ Fortran contributions from (2019). *quadprog: Functions to Solve Quadratic Programming Problems*. R package version 1.5-8. URL: https://CRAN.R-project.org/package=quadprog.

Cornejo-Romero, Amelia et al. (2017). "Alternative glacial-interglacial refugia demographic hypotheses tested on Cephalocereus columna-trajani (Cactaceae) in the intertropical Mexican drylands". In: *PloS one* 12.4, e0175905.

Cressie, Noel (1993). *Statistics for spatial data*. John Wiley & Sons.

Davis, Margaret B and Ruth G Shaw (2001). "Range shifts and adaptive responses to Quaternary climate change". In: *Science* 292.5517, pp. 673–679.

Dawson, Terence P. et al. (2011). "Beyond Predictions: Biodiversity Conservation in a Changing Climate". In: *Science* 332.6025, pp. 53–58. ISSN: 0036-8075. DOI: 10.1126/science.1200303. eprint: https://science.sciencemag.org/content/332/6025/53.full.pdf. URL: https://science.sciencemag.org/content/332/6025/53.

De Lafontaine, Guillaume et al. (2013). "Stronger spatial genetic structure in recolonized areas than in refugia in the E uropean beech". In: *Molecular Ecology* 22.17, pp. 4397–4412.

Diggle, Peter J, Jonathan A Tawn, and Rana A Moyeed (1998). "Model-based geostatistics". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.3, pp. 299–350.

Durre, Imke et al. (2016). "Integrated Global Radiosonde Archive (IGRA), Version 2". In: *NOAA National Centers for Environmental Information*. DOI: 10.7289/V5X63K0Q.

Espíndola, Anahí et al. (2012). "Predicting present and future intra-specific genetic structure through niche hindcasting across 24 millennia". In: *Ecology Letters* 15.7, pp. 649–657.

Excoffier, Laurent and Peter E Smouse (1994). "Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony." In: *Genetics* 136.1, pp. 343–359.

Feurdean, Angelica et al. (2013). "Tree migration-rates: narrowing the gap between inferred post-glacial rates and projected rates". In: *PLoS One* 8.8, e71797.

Finley, Andrew O., Sudipto Banerjee, and Bradley P. Carlin (2007). "spBayes: An R Package for Univariate and Multivariate Hierarchical Point-Referenced Spatial Models". In: *Journal of Statistical Software* 19.4, pp. 1–24. URL: https://www.jstatsoft.org/article/view/v019i04.

Finley, Andrew O., Sudipto Banerjee, and Alan E.Gelfand (2015). "spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models". In: *Journal of Statistical Software* 63.13, pp. 1–28. URL: https://www.jstatsoft.org/article/view/v063i13.

Fotheringham, A Stewart, Martin E Charlton, and Chris Brunsdon (1998). "Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis". In: *Environment and planning A* 30.11, pp. 1905–1927.

Fotheringham, A Stewart, Wenbai Yang, and Wei Kang (2017). "Multiscale geographically weighted regression (MGWR)". In: *Annals of the American Association of Geographers* 107.6, pp. 1247–1265.

Franklin, Janet (2010). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. DOI: 10.18637/jss.v033.i01. URL: https://www.jstatsoft.org/v33/i01/.

Fuglstad, Geir-Arne et al. (2019). "Constructing priors that penalize the complexity of Gaussian random fields". In: *Journal of the American Statistical Association* 114.525, pp. 445–452.

Gao, JIE et al. (2012). "Demography and speciation history of the homoploid hybrid pine Pinus densata on the Tibetan Plateau". In: *Molecular ecology* 21.19, pp. 4811–4827.

Gavin, Daniel G et al. (2014). "Climate refugia: joint inference from fossil records, species distribution models and phylogeography". In: *New Phytologist* 204.1, pp. 37–54.

Gelfand, Alan E et al. (2003). "Spatial modeling with spatially varying coefficient processes". In: *Journal of the American Statistical Association* 98.462, pp. 387–396.

Goldfarb, Donald and Ashok Idnani (1982). "Dual and primal-dual methods for solving strictly convex quadratic programs". In: *Numerical analysis*. Springer, pp. 226–239.

— (1983). "A numerically stable dual method for solving strictly convex quadratic programs". In: *Mathematical programming* 27.1, pp. 1–33.

Graham, Catherine H et al. (2010). "Dynamic refugia and species persistence: tracking spatial shifts in habitat through time". In: *Ecography* 33.6, pp. 1062–1069.

Guan, Kaiyu et al. (2016). "Improving the monitoring of crop productivity using spaceborne solar-induced fluorescence". In: *Global change biology* 22.2, pp. 716–726.

Guanter, Luis et al. (2007). "Estimation of solar-induced vegetation fluorescence from space measurements". In: *Geophysical Research Letters* 34.8.

Guanter, Luis et al. (2014). "Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence". In: *Proceedings of the National Academy of Sciences* 111.14, E1327–E1333.

Guisan, Antoine and Wilfried Thuiller (2005). "Predicting species distribution: offering more than simple habitat models". In: *Ecology letters* 8.9, pp. 993–1009.

Hampe, Arndt and Alistair S Jump (2011). "Climate relicts: past, present, future". In: *Annual Review of Ecology, Evolution, and Systematics* 42, pp. 313–333.

Hao, Qian et al. (2018). "The critical role of local refugia in postglacial colonization of Chinese pine: joint inferences from DNA analyses, pollen records, and species distribution modeling". In: *Ecography* 41.4, pp. 592–606.

Hewitt, Godfrey (2000). "The genetic legacy of the Quaternary ice ages". In: *Nature* 405.6789, pp. 907–913.

Hopkins, D.M., P.A. Smith, and J.V. Matthews (1981). "Dated wood from Alaska and the Yukon: Implications for forest refugia in Beringia". In: *Quaternary Research* 15.3, pp. 217–249. ISSN: 0033-5894. DOI: https://doi.org/10.1016/0033-5894(81)90028-4. URL: https://www.sciencedirect.com/science/article/pii/0033589481900284.

Keppel, Gunnar et al. (2012). "Refugia: identifying and understanding safe havens for biodiversity under climate change". In: *Global Ecology and Biogeography* 21.4, pp. 393–404.

Knowles, Lacey L and Diego F Alvarado-Serrano (2010). "Exploring the population genetic consequences of the colonization process with spatio-temporally explicit models: insights from coupled

ecological, demographic and genetic models in montane grasshoppers". In: *Molecular Ecology* 19.17, pp. 3727–3745.

Lafontaine, Guillaume de, Julie Turgeon, and Serge Payette (2010). "Phylogeography of white spruce (Picea glauca) in eastern North America reveals contrasting ecological trajectories". In: *Journal of Biogeography* 37.4, pp. 741–751.

Lafontaine, Guillaume de et al. (2014). "Beyond skepticism: uncovering cryptic refugia using multiple lines of evidence". In: *New Phytologist* 204.3, pp. 450–454.

Lafontaine, Guillaume de et al. (2018). "Invoking adaptation to decipher the genetic legacy of past climate change". In: *Ecology* 99.7, pp. 1530–1546.

Lemey, Philippe et al. (2009). "Bayesian phylogeography finds its roots". In: *PLoS Comput Biol* 5.9, e1000520.

Lemey, Philippe et al. (2010). "Phylogeography takes a relaxed random walk in continuous space and time". In: *Molecular biology and evolution* 27.8, pp. 1877–1885.

Lemmon, Alan R and Emily Moriarty Lemmon (2008). "A likelihood framework for estimating phylogeographic history on a continuous landscape". In: *Systematic biology* 57.4, pp. 544–561.

Li, Bo, Douglas W Nychka, and Caspar M Ammann (2010). "The value of multiproxy reconstruction of past climate". In: *Journal of the American Statistical Association* 105.491, pp. 883–895.

Li, Furong and Huiyan Sang (2019). "Spatial homogeneity pursuit of regression coefficients for large datasets". In: *Journal of the American Statistical Association*.

Li, Long et al. (2013). "Pliocene intraspecific divergence and Plio-Pleistocene range expansions within Picea likiangensis (Lijiang spruce), a dominant forest tree of the Qinghai-Tibet Plateau". In: *Molecular Ecology* 22.20, pp. 5237–5255.

Lindgren, Finn and Håvard Rue (2015). "Bayesian Spatial Modelling with R-INLA". In: *Journal of Statistical Software* 63.19. DOI: 10.18637/jss.v063.i19.

Lindgren, Finn, Håvard Rue, and Johan Lindström (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4, 423–498. DOI: 10.1111/j.1467-9868.2011.00777.x.

Luoto, M. and R. K. Heikkinen (2008). "Disregarding topographical heterogeneity biases species turnover assessments based on bioclimatic models". In: *Global Change Biology* 14.3, pp. 483–

494. DOI: https://doi.org/10.1111/j.1365-2486.2007.01527.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2486.2007.01527.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2007.01527.x.

Magri, Donatella et al. (2006). "A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences". In: *New Phytologist* 171.1, pp. 199–221. DOI: https://doi.org/10.1111/j.1469-8137.2006.01740.x. eprint: https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.2006.01740.x. URL: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.2006.01740.x.

Manolopoulou, Ioanna and Brent C Emerson (2012). "Phylogeographic ancestral inference using the coalescent model on haplotype trees". In: *Journal of Computational Biology* 19.6, pp. 745–755.

Marion, Glenn et al. (2012). "Parameter and uncertainty estimation for process-oriented population and distribution models: data, statistics and the niche". In: *Journal of Biogeography* 39.12, pp. 2225–2239.

Marske, Katharine A, Richard AB Leschen, and Thomas R Buckley (2012). "Concerted versus independent evolution and the search for multiple refugia: comparative phylogeography of four forest beetles". In: *Evolution: International Journal of Organic Evolution* 66.6, pp. 1862–1877.

McLachlan, Jason S., James S. Clark, and Paul S. Manos (2005). "MOLECULAR INDICATORS OF TREE MIGRATION CAPACITY UNDER RAPID CLIMATE CHANGE". In: *Ecology* 86.8, pp. 2088–2098. DOI: https://doi.org/10.1890/04-1036. eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/04-1036. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/04-1036.

Meirmans, Patrick G. and Shenglin Liu (2018). "Analysis of Molecular Variance (AMOVA) for Autopolyploids". In: *Frontiers in Ecology and Evolution* 6, p. 66. ISSN: 2296-701X. DOI: 10.3389/fevo.2018.00066. URL: https://www.frontiersin.org/article/10.3389/fevo.2018.00066.

Moraga, Paula et al. (2017). "A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE". In: *Spatial Statistics* 21, pp. 27–41.

Mosblech, Nicole A. Sublette, Mark B. Bush, and Robert van Woesik (2011). "On metapopulations and microrefugia: palaeoecological insights". In: *Journal of Biogeography* 38.3, pp. 419–429. DOI: https://doi.org/10.1111/j.1365-2699.2010.02436.x. eprint: https://onlinelibrary.

wiley.com/doi/pdf/10.1111/j.1365-2699.2010.02436.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2699.2010.02436.x.

Napier, Joseph D, Guillaume de Lafontaine, and Melissa L Chipman (2020). "The evolution of paleoecology". In: *Trends in ecology & evolution* 35.4, pp. 293–295.

Napier, Joseph D. et al. (2019). "Rethinking long-term vegetation dynamics: multiple glacial refugia and local expansion of a species complex". In: *Ecography* 42.5, pp. 1056–1067. DOI: https://doi.org/10.1111/ecog.04243. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.04243. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.04243.

Napier, Joseph D et al. (2020). "Ice-age persistence and genetic isolation of the disjunct distribution of larch in Alaska". In: *Ecology and evolution* 10.3, pp. 1692–1702.

Nogués-Bravo, David (2009). "Predicting the past distribution of species climatic niches". In: *Global Ecology and Biogeography* 18.5, pp. 521–531.

Pagel, Jörn and Frank M Schurr (2012). "Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics". In: *Global Ecology and Biogeography* 21.2, pp. 293–304.

Parducci, Laura et al. (2012). "Glacial Survival of Boreal Trees in Northern Scandinavia". In: *Science* 335.6072, pp. 1083–1086. ISSN: 0036-8075. DOI: 10.1126/science.1216043. eprint: https://science.sciencemag.org/content/335/6072/1083.full.pdf. URL: https://science.sciencemag.org/content/335/6072/1083.

Petit, Rémy J. et al. (2003). "Glacial Refugia: Hotspots But Not Melting Pots of Genetic Diversity". In: *Science* 300.5625, pp. 1563–1565. ISSN: 0036-8075. DOI: 10.1126/science.1083264. eprint: https://science.sciencemag.org/content/300/5625/1563.full.pdf. URL: https://science.sciencemag.org/content/300/5625/1563.

Porto, Tiago Jordão, Ana Carolina Carnaval, and Pedro Luís Bernardo da Rocha (2013). "Evaluating forest refugial models using species distribution models, model filling and inclusion: a case study with 14 Brazilian species". In: *Diversity and Distributions* 19.3, pp. 330–340.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly (2000). "Inference of population structure using multilocus genotype data". In: *Genetics* 155.2, pp. 945–959.

Provan, Jim and KD Bennett (2008). "Phylogeographic insights into cryptic glacial refugia". In: *Trends in ecology & evolution* 23.10, pp. 564–571.

Randin, Christophe F. et al. (2009). "Climate change and plant distribution: local models predict high-elevation persistence". In: *Global Change Biology* 15.6, pp. 1557–1569. DOI: https://doi.org/10.1111/j.1365-2486.2008.01766.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2486.2008.01766.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2008.01766.x.

Ren, Guangpeng et al. (2017). "Genetic consequences of Quaternary climatic oscillations in the Himalayas: Primula tibetica as a case study based on restriction site-associated DNA sequencing". In: *New Phytologist* 213.3, pp. 1500–1512.

Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, 319–392. DOI: 10.1111/j.1467-9868.2008.00700.x.

Sass, Danielle, Bo Li, and Brian J Reich (2021). "Flexible and fast spatial return level estimation via a spatially fused penalty". In: *Journal of Computational and Graphical Statistics* 30.4, pp. 1124–1142.

Schurr, Frank M et al. (2012). "How to understand species' niches and range dynamics: a demographic research agenda for biogeography". In: *Journal of Biogeography* 39.12, pp. 2146–2162.

Shafer, Aaron BA et al. (2010). "Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America". In: *Molecular ecology* 19.21, pp. 4589–4621.

She, Yiyuan (2010). "Sparse regression with exact clustering". In: *Electronic Journal of Statistics* 4.none, pp. 1055 –1096. DOI: 10.1214/10-EJS578. URL: https://doi.org/10.1214/10-EJS578.

Simpson, Daniel et al. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors". In: *Statistical science* 32.1, pp. 1–28.

Stein, Michael L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer.

Stewart, John R. et al. (2010). "Refugia revisited: individualistic responses of species in space and time". In: *Proceedings of the Royal Society B: Biological Sciences* 277.1682, pp. 661–671. DOI: 10.1098/rspb.2009.1272. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.2009.1272. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2009.1272.

Svenning, Jens-Christian et al. (2011). "Applications of species distribution modeling to paleobiology". In: *Quaternary Science Reviews* 30.21-22, pp. 2930–2947.

Thuiller, Wilfried et al. (2009). "BIOMOD–a platform for ensemble forecasting of species distributions". In: *Ecography* 32.3, pp. 369–373.

Tsuda, Yoshiaki et al. (2016). "The extent and meaning of hybridization and introgression between Siberian spruce (Picea obovata) and Norway spruce (Picea abies): cryptic refugia as stepping stones to the west?" In: *Molecular Ecology* 25.12, pp. 2773–2789.

Urban, Michael A et al. (2013). "A hierarchical Bayesian approach to the classification of C3 and C4 grass pollen based on SPIRAL δ13C data". In: *Geochimica et Cosmochimica Acta* 121, pp. 168–176.

Wang, Cong et al. (2020). "Satellite footprint data from OCO-2 and TROPOMI reveal significant spatio-temporal and inter-vegetation type variabilities of solar-induced fluorescence yield in the US Midwest". In: *Remote Sensing of Environment* 241, p. 111728.

Wang, Qian et al. (2016). "Arctic plant origins and early formation of circumarctic distributions: a case study of the mountain sorrel, Oxyria digyna". In: *New Phytologist* 209.1, pp. 343–353.

Warren, Emile et al. (2016). "Joint inferences from cytoplasmic DNA and fossil data provide evidence for glacial vicariance and contrasted post-glacial dynamics in tamarack, a transcontinental conifer". In: *Journal of Biogeography* 43.6, pp. 1227–1241.

Wheeler, David C and Catherine A Calder (2007). "An assessment of coefficient accuracy in linear regression models with spatially varying coefficients". In: *Journal of Geographical Systems* 9.2, pp. 145–166.

Wheeler, David C and Lance A Waller (2009). "Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests". In: *Journal of Geographical Systems* 11.1, pp. 1–22.

Whittle, P. (1954). "On Stationary Processes in the Plane". In: *Biometrika* 41.3/4, pp. 434–449. ISSN: 00063444. URL: http://www.jstor.org/stable/2332724.