Distribution Agreement

Signature:

________________________________ ____________

Delante E. Moore 17 October 2024

Surveilling Spatially Concentrated Disease Patterns:
A Bayesian Approach to Understanding Tuberculosis in the United States

By

Delante E. Moore
Doctor of Philosophy
Biostatistics

Lance Waller, Ph.D. Emory University
Advisor

Natalie Dean, Ph.D. Emory University
Committee Member

Neel Gandhi, Ph.D. Emory University
Committee Member

Robert Lyles, Ph.D. Emory University
Committee Member

Benjamin Risk, Ph.D. Emory University
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D., Emory University
Dean of the James T. Laney School of Graduate Studies

Date

Surveilling Spatially Concentrated Disease Patterns:
A Bayesian Approach to Understanding Tuberculosis in the United States

By

Delante E. Moore
BS Electrical Engineering, United States Military Academy, NY, 2005
MS Industrial Engineering, University of Miami, FL, 2014
MS Biostatistics, Emory University, GA, 2024

Advisor:
Lance Waller, Ph.D. Emory University

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2024

Abstract

Surveilling Spatially Concentrated Disease Patterns: A Bayesian Approach to Understanding Tuberculosis in the United States

By Delante E. Moore

Despite advances in medical science, tuberculosis(TB) continues to be a significant health concern across the world. In the United States, local prevalence of TB is often highest in refugee communities due both to higher background prevalence among politically unstable populations and due to a long latent period between infection and active disease. In a public health surveillance setting, this pattern can result in relatively sharp increases in observed local prevalences. By pinpointing areas of concentrated TB prevalence and examining the factors contributing to these patterns, this research aims to enhance the strategic planning and implementation of TB control measures.

Bayesian hierarchical disease mapping models offer a robust means to borrow information across small areas in order to better interpret complex spatial patterns of disease prevalence, enabling public health professionals to identify high-risk areas and tailor interventions more effectively. However, such methods often result in strong spatial smoothing that might dampen the ability to accurately estimate local spikes in prevalence.

Here, we conduct a detailed comparison of Bayesian Hierarchical Models, customizing approaches to better capture TB's spatial heterogeneity. Aim 1 explores performance of hierarchical Bayesian disease mapping models in the context of spatially concentrated TB prevalence. Through a motivating study centered on TB prevalence in Metro Atlanta, we demonstrate the utility of these models in identifying high-risk areas and enhancing our understanding of TB's spatial dynamics but also raise new questions regarding model performance. Aim 2 shifts the focus to model adequacy, using localized approaches to assess the fit of disease surveillance models, particularly in capturing sharp spatial transitions. We evaluate specific epidemiological characteristics of different regions. Finally, Aim 3 focuses on the development of an enhanced Bayesian Spatially Varying Coefficient Model (BSVCM) integrated with Locally Adaptive Smoothing (LAS). This model was specifically designed to address the limitations of over-smoothing in traditional models while also capturing spatial outliers. The LAS-enhanced BSVCM demonstrated significant improvements in modeling accuracy, particularly in regions with high spatial variability, making it a valuable tool for more accurate disease risk mapping and public health interventions. Taken together, the aims explore how customizing Bayesian hierarchical models can provide deeper insights into the epidemiology of TB in the presence of sharp changes in local prevalence.

Dedication

To my inspiration Mackenzie Noelle Moore: may you take from this work the motivation to never give up on your dreams, regardless of their vastness or timing.

“I have learned that success is to be measured not so much by the position that one has reached in life as by the obstacles which he has overcome while trying to succeed."

Booker T. Washington, 1900

Surveilling Spatially Concentrated Disease Patterns:
A Bayesian Approach to Understanding Tuberculosis in the United States

By

Delante E. Moore
BS Electrical Engineering, United States Military Academy, NY, 2005
MS Industrial Engineering, University of Miami, FL, 2014
MS Biostatistics, Emory University, GA, 2024

Advisor:
Lance Waller, Ph.D. Emory University

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2024

Acknowledgments

I would like to start by thanking the Department of Mathematical Sciences, United States Military Academy, who provided funding for the first three years of study. Dr. Donald Outing, a longtime mentor, who encouraged my to pursue Ph.D. studies at a time when I was perfectly content "pushing troops". Brigadier General (R) Tina Hartley and Colonel Michael Scioletti who saw in me what I did not see in myself and selected me as a senior faculty member and later Academy Professor. I am very thankful to them and others teachers, coaches, and mentors in the Army who helped make this a reality for me.

Next let me express my deepest gratitude to my advisor, Dr. Lance Waller, for his unwavering support and guidance throughout my journey. Dr. Waller, you not only inspired my love for all things Bayesian and spatial but also set a gold standard for excellence in biostatistics. Your ability to make complex concepts accessible and engaging has left a lasting impact on me, and I am truly fortunate to have been part of your research group. Your mentorship has been invaluable, and I look forward to future collaborations.

I extend my heartfelt thanks to Dr. Amita Manatunga for believing in me even when I doubted myself. Your encouragement and the occasional "kick in the pants" were exactly what I needed to keep going. Your support has been instrumental in my success, and I am deeply grateful.

A special thanks to Dr. Renee Moore, Dr. Robert Lyles, and Dr. John Helfelt for their words of encouragement and extra assistance when I needed it most. Your kindness and expertise helped me navigate the challenges of this program.

I must also acknowledge the incredible community of PhD students who have been with me every step of the way. Jacob Englert, Sydney Busch, Emily Wu, Wyatt Madden, Rachel Parker, Dr. Raphael Muirden, Hannah Waddell, Thomas Hsiao, and the entire family of fellow students—thank you for getting me through this program, sometimes kicking and screaming. You will always hold a special place in my heart.

I would be remiss if i did not express my deep appreciation to Angela Guinyard, Mary Abosi, and Porshia Arnold. Your sound advice and willingness to engage in those much-needed "grown-folk" talks after challenging days have been a lifeline for me. The support and wisdom you provided during this journey will never be forgotten. Thank you for always being there when I needed it most.

Finally, to my wife, Kimberly—words cannot express my gratitude for your unwavering support and love. You have been my rock, my confidante, and my greatest cheerleader. Without you, this journey would have been impossible. Thank you for standing by my side through it all.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Background

The preamble to the Constitution of the United States states that ...“We the People of the United States, in Order to form a more perfect Union, ... , promote the general welfare, and secure the blessings of liberty to ourselves and our posterity, do ordain and establish this Constitution for the United States of America” [1]. Providing for the health and welfare of the citizenry were of such fundamental concern to the framers that one of our most significant founding documents identifies it as a role of the government in its opening strophe. One method to achieve this is monitoring or surveilling population health status and trends.

Disease surveillance provides a rich source of information about when and where major health problems occur in populations over long periods of time, and large areas. From Lemuel Shattuck’s landmark 1850 Report of the Massachusetts Sanitary Commission that proposed the creation of a permanent public health at the state and local levels in order to gather statistical information on public health conditions that ultimately led to the development of the decennial census and standardization of nomenclature for diseases and causes of death, to modern modern initiatives like the HealthMap Project that leverages machine learning and big data to enable near real time forecasting infectious diseases, the ways and means of executing public health surveillance have evolved over time [109, 21].

Similarly, the definition and objectives of surveillance have evolved as well. Until the 1960's, surveillance in regard to public health was the close monitoring and, if necessary, the quarantining of persons who, due to some exposure, were at risk of developing an infectious disease in order to prevent the spread of disease to others [108]. In a contemporary context, surveillance in regards to public heath involves the ongoing systematic collection, analysis, and interpretation of data, closely integrated with the timely dissemination of these data to those responsible for preventing and controlling disease and injury with a stated goal of providing information that can be used for health action by public health personnel, government leaders, and the public to guide public health policy and programs [110, 25]. One method to achieve these goals is by using mathematical models to quantify, measure, and track infectious diseases over space and time. The current chapter intends to provide some basic information about mathematical modeling of infectious diseases as well as different methods to model them over time and space.

In this work, we seek to reconcile current approaches to infectious disease mapping with the classical methods of modeling infectious diseases given the challenges associated with using aggregated surveillance data. In this chapter, we briefly introduce infectious disease and disease mapping models that we build upon throughout this work and describes the organization of the rest of this paper.

1.2 Spatial Modeling of Infectious Disease Surveillance Data

Spatial modeling of infectious diseases has long played a role in the arena of public health as public health professionals learned how geographical information could be used to explore patterns of diseases and the relationships between diseases and risk factors. As early as 1840, Robert Cowan used a map to highlight the relationship between overcrowding and scarlet fever in Glasgow, Scotland [85]. Perhaps the most famous early use of spatial modeling of a disease was John Snow, who drew dots on a map of the Soho neighborhood of London which he subsequently used to trace the source of the cholera outbreak to fecal-contaminated water supplied by the Broad Street pump [105]. These early attempts to model infectious diseases were based on the basic premise that maps can be used to contain, control, and reduce the spread of infectious diseases. They typically also have a goal of attempting to define the most vulnerable populations at risk for obtaining disease.

The first real Geospatial Information System (GIS), the Canada Geographic Information System in the mid-1960s, led to the increased availability of spatial data that further enabled improvements in the analysis of spatially referenced public health data [76, 33, 98, 120]. Spatially referenced data are any data with a direct or indirect reference to a specific location or geographical area. The location may include a set of coordinates such as longitude and latitudes or administrative areas, such as census tracts, zip codes, congressional districts, etc. In terms of infectious diseases, this data is typically presented as counts of people infected with a particular disease in a small area such as a county. Researchers can then use this data to estimate spatio-temporal patterns in disease risk, explain the variation in the risk by covariate factors, identify high-risk clusters of disease, and conduct surveillance of disease outbreaks [76].

Several statistical methods exist to analyze spatially referenced infectious disease data. The choice of statistical method depends on the type or category of infectious disease data that is available. Blangiardo defined spatially referenced data as the realized values of stochastic process indexed by space formally defined as:

$$Y(s) = y(s), s \in D$$

Where D is a fixed subset of ($\mathbb{R}^d$). In an public health setting, d is typically equal to 2 and represents a 2 dimensional coordinate plane (map). $y(s)$ is an individual observation from the actual data and s is the spatial unit of measurement. If D is a continuous surface then the problem can be specified as a spatially continuous random process, and if D is a countable collection of 2-dimensional spatial units then the problem is defined as a discrete random process. Blangiardo categorized spatially referenced data as either areal, point-referenced, or point-pattern [17]. This distinction is crucial for selecting appropriate statistical models; specifically, point-referenced data, where measurements at fixed locations are random variables, are typically analyzed using geostatistical methods, whereas point-pattern data, in which the locations themselves are random variables in space, require point process techniques to account for the inherent randomness of the spatial locations.

Areal data refers to a situation where $y(s)$ is the aggregation of values over areal units partitioning the overall study area D. Areal units are typically irregular in shape and has boundaries that are defined based on an administrative area such as census tracts, zip codes, counties, etc. For example, state level public health departments and the CDC typically aggregate infectious disease counts by counties.

Point referenced data consists of data measured at specific location y(s) where the spatial domain s at specific locations s varying continuously over the spatial domain D. The location s is commonly represented by two-dimensional vector longitude and latitude. The actual data are represented by observations $y = (y(s_1), y(s_2), ..., y(s_n))$ at locations $(s_1, s_2, ..., s_n)$. For example, we can get measurements of pollutants from air pollution monitors positioned throughout a particular city. The most common goal of point-referenced data in a public health setting is to interpolate y(s) to locations where data measurements are not available.

Point pattern data is the collection of information about whether the event of interest occurred or not at random locations. The spatial domain D represents the set of points where the events of interest occurred. For example, we might be interested in locations of homeless encampments or addresses of persons with HIV. In these examples, the location S in $(\mathbb{R}^2)$ is random and the measurements $y(s)$ are taken as binary value 0, 1 based on whether the event has occurred or not. The main question of interest with point pattern data is whether the event of interest is random or clustered in the spatial domain D.

1.3 Existing Methods and Estimation Approaches for Infectious Disease Model Parameters

1.3.1 Maximum Likelihood Estimation (MLE)

Over the years, a variety of approaches have been used to conduct inference of the model parameters epidemic models. Frequentists tend to apply the Maximum Likelihood approach. Maximum likelihood estimation (MLE) is a method used for fitting a parametric statistical model to data and providing estimates for the model parameters. In general, for a given set of data and an underlying probability model, the method of maximum likelihood selects values of the model parameters such that the parameters maximize the likelihood function, or equivalently, log-likelihood function. Maximum likelihood estimation, which has well-understood large sample properties, provides a unified approach for estimating parameters of interest. Probability statements within a MLE approach rely on indirect statements based on confidence intervals and p-values. Calculation of the confidence levels based on an MLE may require development of appropriate theoretical results and the usual conditions that require asymptotic normality of maximum likelihood estimators are often violated. [52, 74].

1.3.2 Hierarchical Bayesian Approximation

Bayesian inference is the process of fitting a model to data and summarizing the results via the posterior distribution of the parameters and unobserved quantities. In contrast to maximum likelihood estimation, in a Bayesian model, parameters are treated as random variables and relevant data are considered as given. Put simply, in Bayesian inference, before observing the data, the possible means and standard deviations have some prior distribution, which can range from very uncertain to highly informed. After we observe the data, we update our probabilities based on information observed in the data, summarized via the likelihood function. Thomas Bayes formalized the method to describe how uncertainty changes when new data are taken into account as:

$$P(\alpha \mid D) = \frac{P(D \mid \alpha)P(\alpha)}{P(D)}$$

where:

$$P(D) = \int d\alpha P(D \mid \alpha)P(\alpha)$$

Here, α is the set of model parameters, D is the observed data, $P(\alpha \mid D)$ is the posterior distribution, $P(\alpha)$ is the prior probability or what you believed before you saw the evidence, $P(D \mid \alpha)$ is the likelihood likelihood summarizing information in the data regarding α, and $P(D)$ is the "marginal likelihood" or the likelihood of seeing that evidence under any circumstances [74, 9].

In infectious disease modeling, we are often interested in describing data from individuals within groups. In situations like this, the parameters of the model will have meaningful dependencies on each other. For example, we know that different demographic and socio-economic groups have different underlying biases that will impact the probability that they become infected. In this case, whether a person becomes infected depends only on the parameters specific to that individual, but those individual parameters are influenced by the different groups that the person belongs to. This chain of dependencies among the different parameters in the population describes a hierarchical model [73].

1.3.3 Bayesian Hierarchical Disease Mapping

Researchers often utilize disease maps to investigate the spatial variation of disease risk across different geographical regions. These maps display "small-area" risk estimates such as area-level period prevalence or relative risk (RR) of disease occurrence. Relative risk refers to the ratio of the probability of developing a disease in a particular area to the probability of developing the disease in a reference region, which may be the entire study region or a distinct area. Disease maps are instrumental in highlighting spatial disparities in disease risk, generating hypotheses about disease etiology, and informing the allocation of resources [117]. The history of disease mapping dates back to the 1800s, with early examples including maps of yellow fever in the United States and cholera in Europe, which were pivotal in identifying risk factors [122]. One of the most renowned examples is John Snow's 1855 cholera map (Figure 1.1) [104], which depicted the distribution of cholera deaths in London at the residence level and played a crucial role in linking the outbreak to a contaminated water supply.

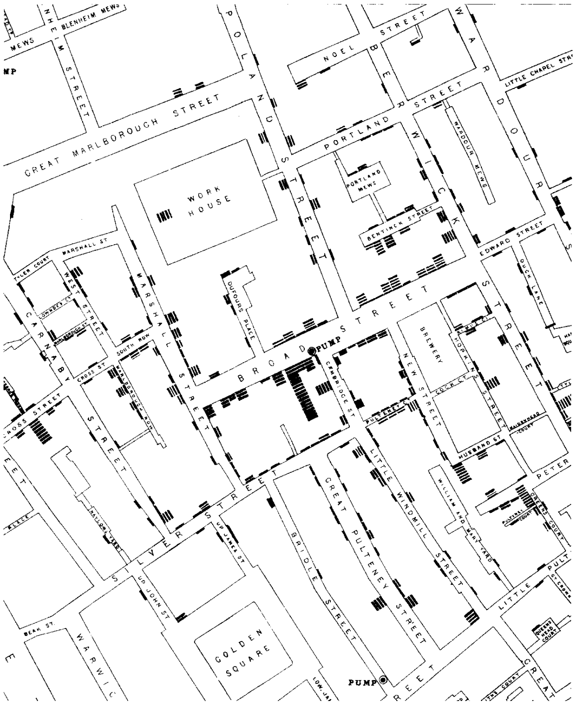


Figure 1.1: John Snow's Cholera Map, [104]

Statistical methods for disease mapping typically utilize either point-level data or areal data. Point-level data provide the exact geographical coordinates of disease cases and non-cases, similar to the data used in John Snow's cholera map. However, point-level data require detailed information on the location of each case, which can be resource-intensive to collect and is often not readily available. Therefore, this discussion focuses on methods for areal data, where the study region is divided into discrete areas. In areal data maps, the data consist of counts of disease cases and the at-risk population for each area. Areal data are more commonly accessible from routine data sources, such as cancer registries or census data, and they protect individual privacy since exact locations are not required [92]. Disease maps of the United States, for example, can show subdivisions by state or by counties within states. The ideal map partition should minimize ecological bias, where associations observed at the aggregate level do not hold at the individual level due to heterogeneity within areas. To ensure that disease risk estimates at the area level closely reflect individual-level risk, it is advisable to use maps with the finest possible subdivisions [37]. However, choosing a coarse grid can result in risk estimates that average out high and low-risk areas.

While finer grids are preferred for disease maps, they can lead to increased random variation in the estimates, particularly in the context of rare diseases or sparse populations [119]. This increased variability can make it challenging to determine whether extreme risk estimates are indicative of true elevated risk or simply a result of random fluctuation. For instance, in maps showing standardized morbidity/mortality ratios (SMRs), the SMR for an area is calculated by dividing the observed number of disease cases by the expected number of cases. When the disease is rare or the population is small, the variance of the SMR can be large, making it difficult to identify true hot spots of disease incidence.

Advanced statistical methods can help address the issue of random variation in disease maps. One common approach is to introduce assumptions about the similarity of risk estimates between neighboring areas, allowing for more stable estimates. Bayesian hierarchical models, or Bayesian "smoothing" models, have been widely used in disease mapping to enhance the precision of risk estimates by borrowing information from adjacent areas. Bayesian smoothing models work by stabilizing disease maps, allowing each area to "borrow strength" from other areas, thereby reducing the impact of random variation. For example, local spatial smoothing models assume that disease risk in a given area is similar to the risk in neighboring areas, which helps to smooth out extreme values by averaging them with the risks of nearby regions. Global smoothing, on the other hand, allows for borrowing strength across non-contiguous areas, assuming that area-level risks deviate from a mean level without a specific spatial pattern.

Despite the benefits of spatial smoothing, it can introduce challenges when detecting clusters or hot spots. Spatial smoothing may mask true clusters by assuming a smooth risk surface, which can lead to a loss of accuracy in identifying isolated high-risk areas [34, 120, 58]. While spatial smoothing can remove extra variation and produce more stable estimates, it may not capture abrupt changes in risk that characterize isolated hot spots.

In summary, disease maps with finer subdivisions are generally preferred because they provide estimates that better approximate individual-level risk. However, maps of crude risk estimates can be unstable when dealing with rare diseases or small populations. Bayesian disease mapping models, which incorporate spatial smoothing, offer a way to enhance the stability of these estimates by reducing random variation. While spatially smoothed disease maps may present a more cohesive view of clustering, their assumptions may obscure the detection of true hot spots. Therefore, the choice of model and smoothing technique should be carefully considered based on the specific objectives of the disease mapping study.

1.4 Motivating Example: Georgia Tuberculosis Surveillance Data

This chapter focuses on addressing problems of spatial aggregation in infectious diseases. Spatial models attempt to identify structural factors that create conditions for the prevalence of certain outcomes [69]. By approaching the analysis of infectious diseases from a spatial perspective, we can identify the characteristics of groups that influence variation in rates of a particular outcome of interest. Unlike many traditional infectious disease models, spatial models begin with the assumption that relationships between variables do not operate similarly in all places or at all levels of interaction [5]. Further, the use of spatial models is based on the premise that geographical variation in the outcome of interest accurately models the influence of aggregate-level factors such as socio-economic status on the individual, and thus, specifically links structure to rates of individual-level outcomes [70].

Due to the limited availability of small-area data, identifying the spatial boundaries of an infectious disease becomes challenging, leading to an instance of the modifiable areal unit problem (MAUP) [79]. MAUP refers to the fact that aggregating data into different sizes or geographical units for spatial analysis can introduce issues that prevent a disease mapping model from fully reflecting the spatial characteristics of an infectious disease [93]. Although this approach does not completely solve the problem, we will employ spatially varying coefficient models to help alleviate it. These models allow parameters to vary across space, thereby reducing the scale effect of MAUP by capturing local variations in the data. This, in turn, will enable researchers to identify potential future disease clusters and allocate the necessary surveillance and mitigation resources to appropriate areas. We test our models on tuberculosis surveillance data collected from the CDC's National Tuberculosis Surveillance System (NTSS) as well as the Georgia Department of Public Health's State Electronic Notifiable Disease Surveillance System (SendSS).

1.4.1 Overview of Tuberculosis

Tuberculosis (TB) is an infectious disease caused by the bacteria Mycobacterium tuberculosis (MTb) that typically manifests through respiratory symptoms, but can also affect other parts of the body such as the kidneys, spine, and brain. The World

Health Organization reports that over 10 million people became infected with TB in 2020 resulting in an estimated 1.5 million deaths. This made TB the second most deadly infectious disease that year after only COVID-19 [94]. Furthermore, in 2020, the United States saw 7,174 cases, resulting 526 deaths, and $503 million in associated heath care costs. Besides the loss of life and financial costs, each case requires a patient to take approximately 180 days of medication, in addition to the burden of multiple x-rays, lab tests, and contact tracing [35, 42]. In response to this loss, the World Health Organization has set a goal to reduce TB incidence by 90% and TB-related deaths by 95% by 2035. Similarly, the CDC established the Division of TB Elimination, whose mission is to promote health and quality of life by preventing, controlling, and ultimately eliminating TB in the United States.[62].

The MTb bacteria is transmitted from person to person via airborne particles mostly through an infected person coughing, speaking, or sneezing in the vicinity of an uninfected individual. One of the most challenging aspects of TB is its latency. Tuberculosis infections can be classified as either “latent” (non-transmissible TB infection) or “active” (generally symptomatic, transmissible TB disease). It is estimated that approximately 23% of the worlds population has latent TB infection (LTBI), which is an asymptomatic condition that is not infectious. However, only five to ten percent of persons with LTBI will progress to active TB disease during their lifetime, with approximately half of those developing the disease within the first two years of exposure [36]. The remaining 90 - 95% of exposed persons remain infected, but the infection remains dormant or latent [94]. The rates of LTBI in the United States are much lower, than those worldwide at closer to 5%, but LTBI still remains a significant risk factor [36].

Complicating issues concerning LTBI, the Human Immunodeficiency Virus (HIV) epidemic exacerbates active TB incidence due to the manner in which HIV suppresses one's immune system [94]. Persons co-infected with HIV and MTb are twenty to thirty times more likely to progress from LTBI to active TB disease [18]. Additionally, TB is one of the leading causes of death among HIV infected individuals both globally and in the United States, with 19% of HIV-positive TB patients dying compared to 3% of HIV-negative patients according to a recent study [18].

In addition to LTBI and HIV co-infection, the proliferation of drug-resistant strains of TB also pose significant challenges to efforts to manage and control infection. Drug resistance typically develops when patients fail to take medications regularly, undergo therapies involving a single drug, have poor drug absorption, take improper combinations of active medications in a treatment regimen, and/or fail to adhere to the prescribed treatment regimen. Increased use of antibiotics to treat infectious diseases caused multidrug resistant (MDR) and extensively drug resistant (XDR) stains of TB to propagate. Drug resistance typically results in fewer and more expensive treatment options, longer treatment durations, and generally poorer outcomes from the remaining available treatments [87, 57]. The WHO defines Pre-XDR as TB caused by Mycobacterium tuberculosis strains that fulfil the definition of MDRand which are also resistant to any fluoroquinolone. Additionally, the WHO defines XDR as TB caused by Mycobacterium tuberculosis strains that fulfil the definition of MDR and which are also resistant to any fluoroquinolone and at least one of the following: levofloxacin or moxifloxacin, bedaquiline and linezolid. [43, 44, 127]. Drug resistance also leads to increased risk of death as evidenced by a 2012 study where even when controlling for several potential confounders, patients with a drug resistant strain of TB had a higher risk of death during during TB treatment in comparison to drug-susceptible TB cases [29]. Disease resistance, latent infections, and the HIV epidemic highlight how critical is it to monitor and control the spread of TB and adds to complexities of TB disease surveillance.

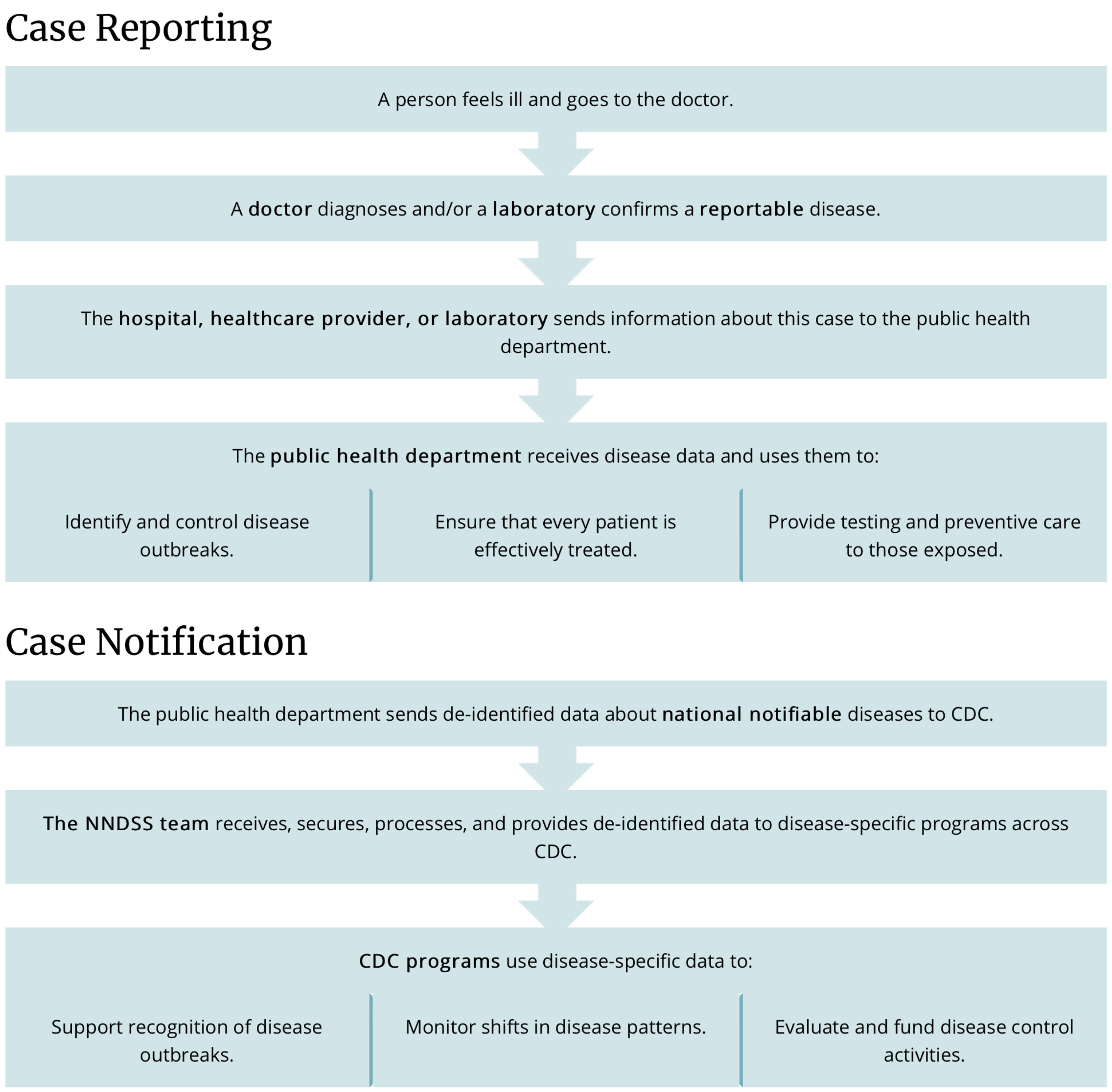
1.4.2 Tuberculosis Surveillance in the United States

Although the prevalence of TB in the United States is among the lowest in the world, the nature of the disease still poses a problem, as many geographic locations within the United States have TB incidence that mirrors that of high burden countries. This requires the United States to develop a dual approach of maintaining and strengthening current TB control priorities, while increasing efforts to identify and treat latent TB infection in populations at risk for TB disease [116].

The CDC defines a notifiable disease as one in which regular, frequent, and timely information regarding individual cases is considered necessary for the prevention and control of the disease [2]. TB became a nationally notifiable disease in 1951 with reporting being mandated in all 50 states and the District of Columbia [82]. Data are collected by local and state health departments and submitted electronically to the CDC, Division of Tuberculosis Elimination. Using case surveillance methods, active TB cases are reported to the National Tuberculosis Surveillance System (NTSS) [45].

Case surveillance begins with local, regional, state, and territorial public health agencies. These agencies collect data on confirmed diagnoses of TB to understand how the frequency of TB cases varies across small areas. This assists researchers and public health officials in identifying TB trends and facilitates the tracking of outbreaks. The reporting process generally occurs in two phases. First, hospitals, healthcare providers, and laboratories report positive lab results or information on people diagnosed with TB to appropriate health departments. Then state and local health departments send deidentified data about confirmed cases of TB to CDC [2]. Figure 1.2 illustrates the CDC's case reporting and notification process. For TB cases, this process is executed via the Report of Verified Case of Tuberculosis (RVCT) [2].

The RVCT allows the CDC to standardize case information collected from all juris-



Case Reporting

A person feels ill and goes to the doctor.

A **doctor** diagnoses and/or a **laboratory** confirms a **reportable** disease.

The **hospital, healthcare provider, or laboratory** sends information about this case to the public health department.

The **public health department** receives disease data and uses them to:

Identify and control disease outbreaks.

Ensure that every patient is effectively treated.

Provide testing and preventive care to those exposed.

Case Notification

The public health department sends de-identified data about **national notifiable** diseases to CDC.

The NNDSS team receives, secures, processes, and provides de-identified data to disease-specific programs across CDC.

CDC programs use disease-specific data to:

Support recognition of disease outbreaks.

Monitor shifts in disease patterns.

Evaluate and fund disease control activities.

Figure 1.2: Flowchart showing the CDC case surveillance reporting procedure for a nationally notifiable disease, such as tuberculosis. The process includes reporting by healthcare providers to local and state health departments, with data ultimately submitted to the CDC for analysis and public health response.

dictions into a unified surveillance system for TB cases from all 50 states, the District of Columbia, and U.S. territories. Information collected on the form includes case categorization, demographic, clinical, laboratory, treatment, risk factors, and outcome characteristics. This system categorizes TB cases according to occupation status, site of disease, and sputum culture conversion. Demographic information includes the sex, age, origin at birth, and race/ethnicity of the patient. Clinical information includes tuberculin skin test (TST) result at time of TB diagnosis, interferon gamma release assay (IGRA) result at time of TB diagnosis, culture result, and initial drug resistance. Risk factors included previous diagnosis of TB, excess alcohol use, drug use, incarceration at time of diagnosis, and residential status at a long-term care facility at time of diagnosis. Treatment outcomes included method of TB treatment, and completion of TB treatment. Method of TB treatment was categorized as directly observed therapy (DOT), self-administered therapy (SAT), both DOT and SAT, or unknown/missing [41]. Cases may be verified using a laboratory result, the clinical case definition, or a provider diagnosis. Follow up reports collect information obtained after the initial case report has been submitted. Data are transmitted to CDC electronically, using CDC-supplied software [82]. See Appendix **??** for the complete RVCT form.

1.4.3 Tuberculosis Surveillance in Georgia

Georgia follows all baseline CDC protocols for reporting cases. Specifically, the state of Georgia, by law, requires all physicians, laboratories, and other health care providers to immediately report clinical and laboratory-confirmed TB cases under their care to Georgia public health authorities. The TB Epidemiology Section of the Georgia Department of Public Health (GDPH) is responsible for the systematic collection of all reported TB cases in the state [91]. The GDPH states that their TB surveillance program allows public health officials and health care providers to follow up with patients, administer directly observed therapy (DOT), monitor TB treatment until completion, evaluate and screen individuals exposed to a TB case, and control TB outbreaks [91].

Hospital infection control personnel as well as other local public health officials electronically report TB cases through the State Electronic Notifiable Disease Surveillance

System (SendSS). If officials are not able to report cases electronically, they may also report TB cases by calling, mailing, or faxing a report to public health authorities. Information contained in the report is similar to that collected by the NTSS and includes demographic, clinical, and risk factor information about reported TB cases as well as their contacts. This information is transmitted to the U.S. Centers for Disease Control and Prevention (CDC) and becomes part of the national TB surveillance database. State and local public health officials use this surveillance data to guide policy and decision making, set priorities for program interventions, evaluate program performance for the prevention and control of TB in Georgia, and educate key stakeholders and the general public on TB [91].

In 2021, Georgia reported 221 new TB cases for a rate of 2.1 cases per 100,000 in 2020. This is slightly lower than the national case rate of 2.4 cases per 100,000 population in 2021. Even though the most recent reports (2021) indicate that Georgia has the sixth highest number of new TB cases, it only had the fourteenth highest rate per 100,000 population among the fifty states who reported information [40].

Public health officials in Georgia also routinely conduct contact investigations among persons exposed to a TB case to identify secondary TB cases and contacts with latent TB infection (LTBI). During a contact investigation, public health staff conduct in-person interviews to ask recent contacts whether they have TB-like symptoms, administer a TB skin test (TST) or interferon gamma release assay (IGRA), repeat the TST or IGRA 8-10 weeks after the last exposure to the index (first) TB case if the initial TST or IGRA is negative, and have a chest radiology exam performed if the TST or IGRA is positive. Georgia identifies persons with LTBI as having a positive TST or IGRA but are asymptomatic and have a normal chest radiology exam. In 2019, the GDPH identified 3,120 people who were thought to have been in contact with someone with an active TB infection. Of the 3,120 people identified, public health officials were able to test 2,361 persons for TB. From this number, 464 had LTBI and 42 had TB disease [91]. Additionally, among the laboratory confirmed cases in Georgia in 2020, there were 17 cases with some form of drug resistance and there were 20 deaths as a result of TB [91].

1.5 Organization

This dissertation is structured to guide the reader through the technical foundations, methodological advancements, and applied analyses that underpin the study of spatial disease mapping, with a particular emphasis on tuberculosis (TB) prevalence.

Chapter 2 lays the groundwork by providing essential background on several technical topics that are frequently referenced throughout the dissertation. This chapter delves into the theoretical aspects of modern disease mapping, focusing on the frameworks and assumptions that guide current research. Additionally, it introduces Hamiltonian Monte Carlo (HMC), an advanced and efficient Markov chain Monte Carlo technique that plays a crucial role in the posterior inference processes utilized in the subsequent chapters. The detailed discussion of HMC is necessary, as it is the primary computational technique used across all models presented in this work.

Chapter 3 addresses Aim 1, where the focus shifts to evaluating the performance of hierarchical Bayesian disease mapping models in the context of TB prevalence that is spatially concentrated. Using a case study centered on Metro Atlanta, this chapter illustrates how these models can effectively identify high-risk areas, providing valuable insights into the spatial dynamics of TB. While demonstrating the practical utility of these models, the chapter also highlights some challenges and raises critical questions about their performance in specific contexts.

Chapter 4 explores Aim 2 by shifting the focus to model adequacy, particularly in the context of localized disease surveillance. This chapter investigates the ability of disease mapping models to capture sharp spatial transitions, which are often critical in understanding the spread of diseases like TB. The chapter evaluates the epidemiological characteristics of different regions and assesses how well the models perform in representing these localized variations.

Chapter 5 discusses Aim 3 and presents a comparative analysis of local spatial regression models, specifically Bayesian Hierarchical Models (BHM)—including the Intrinsic Conditional Autoregressive (ICAR) and Horseshoe models—and Geospatial Weighted Regression Models (GWR). This chapter examines the strengths of each approach in capturing spatial variation in the associations between TB prevalence and local covariates. The comparison aims to provide a deeper understanding of complex spatial patterns and to identify which models are best suited for various analytical scenarios.

Finally, Chapter 6 concludes the dissertation by summarizing the key findings and discussing potential directions for future research. Throughout the research presented here, several new discoveries and challenges emerged that could not be fully addressed within the scope of this dissertation. These include potential model extensions, the development of customized methods for marginal likelihood estimation, and the exploration of alternative computational techniques. This final chapter aims to reflect on the broader implications of the work and to propose avenues for continued exploration in the field of spatial disease mapping.

Chapter 2

Technical Background on Disease Mapping Topics

2.1 Theoretical Foundations of Linear Mixed Models

2.1.1 Introduction

Traditionally, statistical analyses often rely on linear models to capture relationships between variables. While effective in capturing linear trends, these models often falter when faced with data characterized by inherent variability due to unobserved factors or repeated measurements on the same subjects. The limitations become particularly pronounced when working with data structures where observations are nested within higher-level units such as individuals within groups or repeated measurements over time.

Linear Mixed Models (LMMs) provide a potent solution to these multifaceted challenges. These models offer a flexible and robust framework for modeling relationships between variables, while simultaneously accommodating the intrinsic variability introduced by different levels of grouping in the data. By assimilating fixed effects, which capture population-level relationships, and random effects, which encapsulate individual-level deviations from these population-level trends, linear mixed models furnish a more nuanced and accurate representation of complex real-world data [19]. In the sections below we define LMMs, followed by a detailed description of the concept of Best Linear Unbiased Prediction (BLUP). Throughout, we contrast so-called "old style" and "new style" interpretations of random effects, and contrast their roles in both estimation and prediciton.

2.1.2 Definition

A general equation for a linear mixed model (LMM) can be written as:

$$Y_i = X_i\beta + Z_i u_i + \epsilon_i \tag{2.1}$$

Where:

- Y_i is the vector of observations.
- X_i is the design matrix for the fixed effects.
- β is the vector of fixed effect coefficients.
- Z_i is the design matrix for the random effects.
- u_i is the vector of random effect values.
- ϵ_i is the vector of (within i) random errors.

Assumptions include:

- $E(\underset{\sim}{\epsilon_i}) = 0$ and $Var(\underset{\sim}{\epsilon_i}) = \sigma^2 \underset{\sim}{V_i}$; where $\underset{\sim}{V_i}$ is a matrix that defines the structure of the variance of the residuals (ϵ_i) and σ^2 is the error variance.

- $E(\underset{\sim}{u_i}) = 0$ and $Var(\underset{\sim}{u_i}) = G$; where G is the covariance matrix of the random effects.
- $\underset{\sim}{\epsilon_i}$'s and $\underset{\sim}{u_i}$'s are mutually independent

2.1.2.1 Random Effects in Linear Models

To grasp the significance of LMMs, one must appreciate how they differ from traditional linear models. Traditional linear regression places exclusive emphasis on fixed effects, assuming that all observations are independent but not identically distributed since $E(Y)$ varies across observations. This assumption is often inadequate when dealing with data characterized by correlation, clustering, or hierarchical structures, as frequently encountered in longitudinal studies, repeated measurements, and data collected across varying levels of aggregation [75].

Random effects introduce an intricate hierarchy into the modeling process. This hierarchical structure encapsulates the insight that observations within the same group or subject are more likely to be correlated with each other than with observations from different groups or subjects. Often, random effects associated with a particular layer of a hierarchical model are independent of one another. However, when their predicted values are incorporated within each layer, observed outcomes are independent conditioned on these values but correlated within each layer in the joint model. These relationships provide a convenient structure for incorporating level-specific correlations such as repeated measures on the same study subject, observations taken at the same time, or in nearby spatial locations.

Random effects, u_i, capture the variability in the data that is not explained by the fixed effects. They are assumed to be normally distributed with mean 0 across experimental units, i.e.,

$$u_i \sim N(0, G) \tag{2.2}$$

Where:

- G is the covariance matrix of the random effects, capturing the variability of each random effect and the covariance between them.

The residuals, ϵ_i, represent the variability not explained by either the fixed or random effects. They are also assumed to be independent and normally distributed:

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2 \underset{\sim}{V_i}) \tag{2.3}$$

where:

- σ^2 is the error variance.
- $\underset{\sim}{V_i}$ is the matrix defining the variance structure of the residuals.

The combined variability in the outcome, considering both random effects and residuals, is:

$$Var(y_i) = Z_i G Z_i' + \sigma^2 \underset{\sim}{V_i} \tag{2.4}$$

where:

- Z_i' is the transpose of the design matrix for the random effects.

In more complex designs, the analysis can include nested or crossed random effects. Nested effects occur when one grouping factor is completely contained within another (e.g., students within schools). Crossed effects occur when grouping factors overlap but are not nested (e.g., students and test examiners).

The composition and quantity of random effects, along with their interconnections (whether crossed or nested), significantly impact the analysis and interpretation of linear mixed models. Moreover, the selection of covariance structures, such as deciding if matrix G is diagonal or includes more complex patterns, is crucial. This is because the structure of G can be instrumental in shaping the expected correlations among observations. Such correlations may arise due to repeated measurements or from temporal or spatial relationships among the data points, thereby influencing the analytical outcomes and interpretations of these models.

In summary, random effects in linear mixed models allow for the incorporation of group-specific variability into the model. Their proper specification and understanding are crucial for correct modeling and interpretation of hierarchical or grouped data. By accounting for such correlations, LMMs engender more accurate parameter estimation, robust hypothesis testing, and enhanced reliability in the statistical inferences derived from the model. Ultimately, LMMs form a versatile statistical framework designed to address the challenges posed by hierarchical and correlated data. Comprising both fixed and random effects, LMMs provide a flexible means to model relationships between variables while accounting for the variability introduced by various levels of grouping. The inclusion of random effects acknowledges the correlations anticipated within the hierarchical grouping structure, enhancing parameter estimation and enabling robust statistical inference.

2.1.2.2 Prediction versus Estimation

Traditional statistical theory tends to focus almost exclusively on estimation, emphasizing its centrality in the field. However, as illuminated by linear mixed models, the importance of prediction cannot be overlooked [20]. The primary goal of statistical estimation is determining the values of unknown parameters in a statistical model. In the context of linear mixed models, these parameters can encompass the fixed effects, like treatment effects, or the variances and covariances of the random effects. Estimation is driven by a desire to generalize about the population from which samples are drawn. Instead of pinpointing predictions for individual outcomes, estimation concentrates on deducing information about the entire population or specific groups within that population. The results of estimation are typically expressed in terms of point estimates (which offer the best guesses for parameter values) and interval estimates (providing a range of plausible values for the parameters).

The main objective of statistical prediction is to produce forecasts or deduce values of the response variable for new observations, particularly when given specific values for the predictor variables. Predictions using LMMs involve using predicted values of random effects to provide statistical predictions of individual-level outcomes. In a linear mixed model, one might predict the response of a particular subject based on fixed effects, such as treatment conditions, conditional on the value of an individual's random effects, like individual-specific random intercepts. The predictions factor in both the average response across all subjects and the deviations specific to individual subjects. In essence, while estimation focuses on understanding the underlying structure and parameters of the model generating the observed data, prediction zeroes in on forecasting new data observations. Further, in LMMs, the predictions of the random effects enable "best guesses" applicable to specific subjects.

2.1.3 Best Linear Unbiased Prediction

Best Linear Unbiased Prediction (BLUP) is a method used to make predictions of random effects in linear mixed models. While the fixed effects are estimated using least squares or maximum likelihood methods, BLUP provides a way to predict the random effects. It is termed "best" because it minimizes the prediction error variance, "linear" because it is a linear function of the observed data, and "unbiased" because, on average, it correctly estimates the true value of the random effect. BLUP plays an instrumental role in providing both estimation-based and prediction-based inference within the linear mixed model framework. It not only aids in understanding the structure of the observed data but also in forecasting unobserved outcomes. By optimizing the balance between these crucial aspects, BLUP ensures that the resulting predictions are both accurate and unbiased across varying levels of the hierarchy and that the estimates of the model's fixed effects maintain clear statistical properties as well. BLUP plays a pivotal role in capturing the nuances of individual-level deviations from population-level trends, thereby contributing to the comprehensive analysis of hierarchical data [99].

The marriage of LMMs with BLUP imparts precision and efficiency to data analysis. BLUP provides a robust means to predict random effects, contributing to accurate modeling of within-group variability. In the context of LMMs, this prediction is seamlessly integrated within the broader inferential framework, enabling researchers to construct more accurate and nuanced models that capture the complex interplay between fixed and random effects. BLUP not only enhances the quality of parameter estimates but also facilitates the prediction of unobserved values within and across hierarchical structures.

In the paper entitled, *That BLUP is a good thing: the estimation of random effects*, George Robinson nicely summarizes the relationship between the linear mixed model and the BLUP. In the work, given the linear mixed model, Robinson (1991) describes the best linear unbiased prediction (BLUP) for the random effects u_i as [99]:

$$\hat{u_i} = GZ_i'\Sigma_i^{-1}(y_i - X_i\hat{\beta}) \tag{2.5}$$

where Σ is the variance-covariance matrix of y_i and is defined as:

$$\Sigma_i = Z_i G Z_i' + \sigma^2 \underset{\sim}{V_i}, (i = 1, ..., n) \tag{2.6}$$

and the estimated fixed effects $\hat{\beta}$ are provided by:

$$\hat{\beta} = (\sum_{i=1}^{n} X_i' \hat{\Sigma_i}^{-1} X_i)^{-1} \sum_{i=1}^{n} X_i' \hat{\Sigma_i}^{-1} y_i \tag{2.7}$$

The prediction of a new observation $\hat{y}^*$ conditional on $\widehat{u_i}$ is given by:

$$\widehat{y_i^*} = X_i^* \hat{\beta} + Z_i^* \hat{u}_i \tag{2.8}$$

.

The foundation of BLUP-based estimation lies in the calculation of BLUP-based predictions of u_i and BLUP-adjusted estimates for β, represented as $\hat{u}_i$ and $\hat{\beta}$, respectively. These estimates are governed by the mixed model equations, as initially formulated by Henderson in 1950. Mixed models have been foundational in quantitative genetics and animal breeding. Henderson played a pivotal role in the development and application of mixed model equations, especially in animal breeding. His work in the mid-20th century established methods for estimating variance components and making predictions using mixed models. Here are the basic mixed model equations as formulated by Henderson:

$$\begin{bmatrix} X' \underset{\sim}{V}^{-1} X & X' \underset{\sim}{V}^{-1} Z \\ Z' \underset{\sim}{V}^{-1} X & Z' \underset{\sim}{V}^{-1} Z + G^{-1}\sigma^2 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X' \underset{\sim}{V}^{-1} y \\ Z' \underset{\sim}{V}^{-1} y \end{bmatrix} \tag{2.9}$$

The simultaneous equations utilize these matrices to derive BLUP predictions of random variable u and estimates of fixed effects β, thereby establishing their interrelation with the observed data y. The mixed model equations offer a systematic approach to ascertain these parameter estimates conditional on the predicted random effects.

Comparatively, BLUP estimates diverge from those obtained through Generalized Least Squares (GLS) when random effects are treated as fixed effects. While GLS aims to provide the most precise estimates of fixed effects by minimizing the variance of the errors, BLUP acknowledges the stochastic nature of random effects and aims to provide the best linear unbiased predictions (BLUP) for them. This distinction is particularly pronounced in the case of the dairy cow example detailed below [99], where BLUP effectively accommodates the variability in sire effects based on available information. In essence, while both BLUP and GLS are integral tools in the realm of statistical modeling, they cater to different needs, with BLUP excelling in its ability to generate reliable predictions for random effects, and GLS standing out in its capability to estimate fixed effects.

To illustrate the practical application of BLUP, we consider a scenario involving first lactation yields of dairy cows. In this case, herd effects are treated as fixed (β), while sire additive genetic merits constitute the random effects (u_i). The variance-covariance matrix $\sigma^2 I$ encapsulates deviations of yield from a model primarily explained by sire and herd effects. Meanwhile, the known multiple of the identity matrix is represented by the genetic variation matrix G. By utilizing BLUP estimates in this context, we capitalize on available information regarding each sire's daughters' lactation yields, leading to more informative and precise estimates compared to the fixed sire effects [99].

A critical aspect of BLUP estimation lies in assessing the precision of the obtained estimates. The variance-covariance matrix of BLUP estimates is determined by σ^2 multiplied by the inverse of the left-hand-side matrix in the mixed model equations [99]. This inverse matrix encapsulates variance estimates for various components, such as herd and sire effects, providing a quantification of the estimate precision. Notably, increased information availability for a specific sire translates to diminished estimate variation, ultimately yielding more dependable and finely-honed estimates of its genetic merit.

While the illustrative model detailed in Robinson (1991) used small numbers with few significant digits to make the example easier to follow, it underscores the significance of accurate variance parameter estimation. The residual variance σ^2 is crucial for BLUP for several reasons:

1. **Weighting Observations**: BLUP uses the inverse of the total variance-covariance matrix of y_i to weight observations when predicting random effects. The matrix is determined by both the structure imposed by random effects (via the matrix $Z_i G Z_i'$, where G is the covariance matrix of the random effects) and the residual variance σ^2.
2. **Shrinkage**: The magnitude of the residual variance relative to the variance of the random effects determines the amount of "shrinkage" applied to the random effect predictions. If the residual variance is large relative to the random effect variance, the BLUPs will be shrunken more towards the population mean. This shrinkage property ensures that the predictions are stable and less influenced by random noise.
3. **Model Fit and Diagnostics**: The estimated σ^2 can be used to assess the goodness of fit of the model. Large residuals (compared to what is expected under the assumed σ^2) can indicate model misspecification, outliers, or influential points.

Ultimately, the residual variance in the context of BLUP and mixed models quantifies the variability in the data unexplained by the fixed and random effects. It plays a pivotal role in weighting observations, determining the degree of shrinkage of the random effect predictions, and assessing the model's fit.

2.1.3.1 Links to Other Statistical Theory

Robinson (1991) connects the BLUP approach to other branches of statistical theory. It is fascinating to note that BLUP has its roots in animal breeding and genetics, specifically tailored for ranking and selection. The BLUP has proved especially useful when ranking things based on traits you cannot easily observe, treated as random effects. Even though BLUP theory has roots in animal breeding experiments, it is now used in many other areas. Robinson's (1991) discussion focuses on several key themes: the retrieval of inter-block information through BLUP, a comparative analysis of Bayesian and Classical approaches to BLUP-based inference, and an exploration of BLUP's relationship with regression to the mean. Through this examination, we gain a comprehensive understanding of how BLUP-based inference interacts with broader statistical concepts.

As noted above, a crucial aspect of statistical analysis is the estimation of variance parameters, where the Restricted Maximum Likelihood (REML) method stands out, particularly with unbalanced data. REML is recognized for its effectiveness in estimating variance parameters in such scenarios, as it provides unbiased estimates of variance components, which are essential for precise and dependable statistical results. Unlike the Maximum Likelihood (ML) method, which tends to underestimate variance components, REML corrects for the bias introduced by estimating fixed effects. When estimating variance components, if fixed effects are included in the model, their estimation consumes degrees of freedom and can bias the estimates of the variance components. REML addresses this issue by first estimating the fixed effects, then using the residuals (the remaining variation after accounting for fixed effects) to estimate the variance components, thereby accounting for the loss of degrees of freedom due to the estimation of fixed effects. This results in more accurate and unbiased estimates of the variance components, facilitating robust statistical inferences. This unbiased estimation of variance components is crucial for making robust statistical inferences, particularly when dealing with intricate and varied data structures.

Next, Robinson (1991) discusses how BLUP offers a useful tool for addressing outliers. BLUPs function as a useful tool for addressing outliers. Drawing parallels between BLUPs and residuals, Robinson introduced a method to detect and address outliers using these predictions.

In cases where dispersion parameters require estimation from data, we enter a multifaceted statistical landscape. Within this framework, the merits of the REML approach are discussed as a potent solution to this challenge. Concurrently, we consider approximate Bayesian procedures as an alternate route for capably estimating fixed and random effects in the presence of dispersion parameter estimation.

Shifting his focus, Robinson (1991) presents Empirical Bayes methods as a robust toolkit for estimating distributions of random effects. Through this lens, we understand how these methodologies empower the estimation of realized values of random effects by leveraging Bayes' theorem. With the application of empirical Bayes techniques, we identify an avenue to refine the accuracy and consistency of random effects estimation, enriching the quality of statistical inferences.

In the sections below, we expand on Robinson's (1991) overview and provide details on the following modern estimation approaches for predicting random effects and predicting fixed effects. Specifically, we offer additional detail on Restricted Maximum Likelihood, Penalized Maximum Likelihood, Bayesian Hierarchical Models, and Bayesian Lasso Methods.

2.1.3.2 Further Commentary

Robinson's work prompted further discussions and insight into statistical modeling, especially the thoughtful comments by Harville (1991) and Speed (1991). Specifically, Harville (1991) touches on the relevance of the Best Linear Unbiased Prediction (BLUP) in animal breeding. He also relates BLUP to other established prediction methods like kriging and the Kalman filter. Although BLUP was originally designed from a frequentist viewpoint, Harville (1991) discusses its interpretation within a Bayesian context as well (as detailed below). He presents BLUP as a broader tool for prediction, covering its technical aspects and foundational assumptions. Harville (1991) emphasizes using BLUP to predict unseen variables from observable data and introduces key statistical parameters. He also dives into the relationship between BLUP and other predictive methods, comparing empirical BLUP with empirical Bayesian techniques.

Importantly, Harville addresses the common misunderstanding that mixed-effects models are limited to fixed effects and variance components. He highlights the importance of recognizing both fixed and random effects in these models. Harville then looks into Bayesian prediction methods, noting their computational demands and the need for further research. He suggests the possibility of a unified prediction approach, recommending that models be viewed as joint distributions for framing specific prediction problems. Assumptions, non-informative priors, and evaluation criteria play a vital role in this context. Overall, Harville's response provides a detailed analysis of Robinson's work, focusing on BLUP, its relationship with Bayesian methods, and the potential for an integrated approach to prediction problems [59].

Speed's (1991) response to the Robinson (1991) praises its clarity in explaining a complex topic, ability to bridge gaps in the subject, and its provocative nature. Speed (1991) commends Robinson (1991) for highlighting the role of BLUPs and offers additional insights to strengthen the case for explicit recognition of BLUPs. Speed (1991) discusses Robinson's (1991) Bayesian derivation with an emphasis on the posterior mode being equivalent to BLUP estimates. The possibility of using proper priors and standard Bayesian formulas for deriving BLUPs is mentioned and we expand on this in Section 4.3 below. Speed (1991) also discusses different formulas for random effects, including an "obvious plug-in" expression involving matrices V, Z, and G. The response touches on solving BLUP equations, suggesting an iterative approach based on rearranging equations and updating variance components.

Speed (1991) further highlights the connection between REML and BLUP, suggesting that REML equations equate observed with expected sums of squares of BLUPs. Additionally, the concept of penalized least squares, which can turn standard least squares into a case of BLUP, is discussed in relation to its initial formulation by Charles Roy Henderson in 1950 [61]. Speed then explores several relationships between the BLUP and and other theorectical concepts relating to the LMM. These include relationships between smoothing splines and BLUPs, where Speed (1991) corrects the terminology in the spline literature and connecting Generalized Maximum Likelihood with REML. Speed further relates linear smoothers to BLUPs, with similarities in their underlying theory and formulas. Lastly, Speed (1991) addresses the issue of interval estimates involving BLUPs along with concerns about interpreting Bayesian posterior intervals in this context. The response concludes with a play on words, suggesting that, just as "To a Bayesian, all things are Bayesian," a summary of the paper could be "To a non-Bayesian, all things are BLUPs". Overall, the response provides additional insights, connections, and considerations related to the themes discussed in the Robinson (1991) article about BLUPs and their applications [106].

2.1.4 “Old” versus “New” Style Random Effects

Robinson’s article describes a technique for incorporating predictions of random effects into estimates of fixed effects. BLUP can be used in a variety of settings, including animal breeding, ore reserve estimation, and insurance premium calculation. The article also discusses the relevance of BLUP to the foundations of statistics. The book *Richly Parameterized Linear Models* by James Hodges provides a succinct overview of the evolution in the conceptualization and application of random effects in statistical models [64].

Hodges states that many things are now called random effects that do not fit the the definition originally provided by Scheffe in 1959 [103]. This distinction has several implications in conceptual as well as practical ways.

Old-style random effects are those that are thought to be draws from a population. For example, in a study of nerve fiber density, subjects are considered to be a random sample from the population of all non-diabetic adults. The random effects in this model are the subject main effect, the method-by-subject interaction, and the location-by-subject interaction. These random effects describe how the average nerve density varies between subjects, between methods, and between locations.

New-style random effects are those that are not thought to be draws from a population. They are used to implement smoothing or shrinkage in the model. For example, in a penalized spline model for global mean surface temperature, Hodges (2013) uses a random effect to smooth the curve and reduce the variance. In a spatial model of stomach cancer mortality, Hodges (2013) uses a random effect to capture the spatial clustering of cases.

The distinction between old-style and new-style random effects is important for inference, prediction, and simulation. For old-style random effects, we can make inferences about the population from the sample. For new-style random effects, we can only make inferences about the particular data set that we have.

The old style of random effects estimation typically relies on methods such as the method of moments or REML. Both approaches maintain the assumption that the random effects follow a normal distribution with constant variance. Here, random effects are interpreted as deviations from the overall population mean. The assumption of normality can lead to biased estimates when the underlying distribution deviates from normality. Additionally, the reliance on homoscedasticity assumptions might be problematic when dealing with heteroscedastic data.

In contrast, the new style of random effects involves modern techniques such as Penalized Maximum Likelihood methods such as Lasso or Ridge. They may also be estimated by Bayesian methods. These methods incorporate regularization into the estimation process to prevent overfitting, stabilize estimates, and address multicollinearity. Additionally, random effects can be estimated using Bayesian methods, which also apply regularization to enhance the estimation process. Further, Bayesian methods provide flexible alternatives for estimating random effects by leveraging the advantages of Bayesian inference. Simulation studies and real data applications have demonstrated that this new style of random effects tends to outperform the old style in terms of estimation accuracy, model fit, and predictive ability, especially in scenarios with limited sample size or heterogeneous data [65].

The distinction between old-style and new-style random effects is important for understanding the interpretation of statistical models. As noted above, old-style random effects can be used to make inferences about the population, while new-style random effects can only be used to make inferences about the particular data set that we have. The choice of the distribution for a new-style random effect is a modeling decision that should be made based on hierarchies or association, the specific data set, and the goals of the analysis.

2.1.4.1 Restricted Maximum Likelihood (REML)

In discussing the methodologies used for estimating variance components in linear mixed models, attention is often directed towards the Restricted Maximum Likelihood (REML) approach, as noted above. This method, known for its iterative nature, is particularly effective in models with fixed effects. It addresses the tendency of Maximum Likelihood (ML) estimates to underestimate variance components by not considering the reduction in degrees of freedom during the estimation of random effect parameters. By excluding fixed effects from the likelihood function, REML provides more accurate estimates of variance components.

Corbeil (1976) presents the REML in detail. REML is formulated as the maximizer to the following log likelihood:

$$^*(\gamma) = -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}\log(|X'\Sigma^{-1}X|) - \frac{1}{2}(Y - X\hat{\beta})'\Sigma^{-1}(Y - X\hat{\beta}) \tag{2.10}$$

Here γ is a vector that contains all the variance components. This approach is "restricted" compared to standard maximum likelihood estimation because it modifies the likelihood function. Specifically, REML adjusts for the degrees of freedom consumed by estimating fixed effects (represented by $\hat{\beta}$ in the equation). This adjustment is not present in the standard maximum likelihood method, making REML more suitable for unbiased estimation of variance components, particularly in models with several fixed effects.

Corbeil's (1976) method develops estimators that do not include fixed effects in their likelihood, maximizing over a restricted parameter set. This generalizes earlier methods for balanced data and random models, extending applicability to unbalanced data and mixed models with various fixed and random effects. The process involves partitioning the likelihood function, segregating the portion independent of fixed effects, and maximizing this to obtain REML estimators for variance components [32, 13].

Corbeil (1976) emphasizes that REML estimators are invariant to the fixed effects in the model and do not depend on estimates of these fixed effects. In balanced data scenarios, REML estimators align with traditional analysis of variance (ANOVA) estimators, a property not shared by ML estimators, offering advantages due to the optimal properties of ANOVA estimators in balanced data.

In transitioning from traditional random effects estimation methodologies like REML to more contemporary approaches, we observe a significant evolution in statistical modeling techniques. REML, with its roots firmly in classical statistical theory, primarily focuses on obtaining unbiased estimates of variance components under the assumption of normally distributed errors. This method has been a cornerstone in handling random effects, particularly in linear mixed models, offering a robust framework for dealing with complex data structures. The transition from REML represents a paradigm shift from the 'old style' of random effects estimation to a 'new style' that emphasizes flexibility, adaptability, and the capacity to manage more complex and high-dimensional data structures. This shift not only reflects the advancements in computational power and data availability but also the growing need for models that can accurately capture the intricacies of modern datasets while maintaining robustness and interpretability.

2.1.4.2 Penalized Maximum Likelihood

One popular method of incorporating new-style random effects within a linear mixed model is through penalized maximum likelihood. New style random effects utilize the penalized maximum likelihood. Penalized maximum likelihood is an advanced statistical approach that seeks to bridge the gap between traditional maximum likelihood estimation and regularization techniques. In the context of mixed models, where random effects play a pivotal role in accounting for unobserved heterogeneity, the introduction of penalized maximum likelihood offers a novel way to predict the values of random effects and estimate fixed effects conditional on these predictions. Instead of merely maximizing the likelihood of the observed data, this method introduces penalties to the likelihood function, essentially providing a trade-off between the fit of the model and the complexity of the random effects. This approach is particularly useful in situations where there is a risk of overfitting due to a large number of random effects or when data is sparse for certain levels. By incorporating penalties, the method enhances robustness, prevents overfitting, and improves the generalization of random effects in complex hierarchical models [31]. For a generalized linear model, where the relationship between covariates and expected outcomes is given by:

$$g(E(y_i)) = X\beta \tag{2.11}$$

with g representing the link function, the log-likelihood is:

$$l(\beta) = \sum_{i=1}^{n} y_i x_i' \beta - b(x_i' \beta) \tag{2.12}$$

where b represents a function of the linear predictor $x_i'\beta$. The function $b(\cdot)$ is related to the canonical parameter of the exponential family distribution that is assumed for the response variable.

The penalized log-likelihood becomes:

$$l_p(\beta) = l(\beta) - \lambda P(\beta) \tag{2.13}$$

Here, λ is a non-negative regularization parameter, and $P(\beta)$ denotes the penalty function. These newer methods extend the fundamental principles of maximum likelihood estimation by incorporating penalty terms, which impose regularization on the model parameters. This regularization, often seen in the form of L1 (Lasso) or L2 (Ridge) penalties, enables the handling of high-dimensional data and mitigates issues like overfitting, a common challenge in traditional models. These penalty functions are given by:

1. L1 penalty (Lasso):
$$P_{L1}(\beta) = ||\beta||_1 = \sum_{j=1}^{p} |\beta_j| \tag{2.14}$$
2. L2 penalty (Ridge):
$$P_{L2}(\beta) = ||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2 \tag{2.15}$$

The penalized maximum likelihood estimates maximize $l_p(\beta)$.

Disease Mapping Models provide a case study for the use of Penalized Maximum Likelihood. Disease mapping refers to the process of estimating disease risk or prevalence over a spatial domain, often on a fine spatial scale like neighborhoods, towns, or counties [77]. One of the main goals of disease mapping is to identify areas with unusually high or low disease risks which might not be apparent when using raw disease counts alone. However, direct estimates based solely on observed disease counts in each area can be highly unstable, especially for rare outcomes observed in subsets with small numbers of observations.

To deal with these issues, we employ smoothing or shrinkage techniques. Penalized Maximum Likelihood (PML) is a powerful method for achieving this. Instead of estimating the disease risk purely based on the observed counts, PML introduces a penalty term in the likelihood which shrinks the estimates towards a global mean or a local average. This process “borrows strength” from neighboring regions, thereby reducing random noise in the estimates [77].

The penalty term typically involves some measure of the roughness of the estimated risk surface over the domain. Smoother risk surfaces are preferred because they are more likely under the assumption that disease risk changes gradually over space. In fact, the smoother the surface, the more information a local estimate will borrow from its neighboring values, resulting in a more stable and reliable estimate. The penalty term usually depends on a smoothing parameter, which determines the amount of smoothing applied: a larger smoothing parameter results in more shrinkage and a smoother risk surface.[31].

To further illustrate, imagine a state with 10 cities. You want to estimate the risk of a particular disease in each city. Raw counts for each city might be:

City	Observed Cases	Population
A	100	100,000
B	5	500
C	25	5,000
...	—	—
J	50	5,000

City B has a raw risk of 1% (5 cases out of 500 people), which is higher than City A's risk of 0.1% (100 cases out of 100,000 people). However, with such a small population in City B, this estimate is very uncertain. Using PML, the risk estimate for City B would shrink towards the overall average risk for all towns or possibly towards the risk estimates of neighboring towns.

After applying PML, the risk in City B might be revised downwards, say to 0.8%, reflecting both the observed data and the borrowed information from other towns. This smoothed estimate would be more stable and reliable than the raw estimate.

Penalized Maximum Likelihood is a crucial tool in disease mapping, ensuring that estimates are both spatially smooth and more reliable. By borrowing strength from neighboring regions, PML can produce more stable and credible risk maps, aiding in the identification of areas with genuinely high or low disease risks.

2.1.4.3 Bayesian Framework

Further expanding on the concept of new style of random effects, the Bayesian framework provides for very flexible specification and inference. Within the Bayesian framework, every parameter, including random effects, is assigned a prior distribution. For random effects, commonly used priors are the Gaussian (normal) distributions, though others can be used based on domain knowledge or modeling considerations. The prior encapsulates the researcher's beliefs about the distribution of the random effects before observing the data [55].

As with traditional random effects models, the likelihood function characterizes how the observed data is generated given the parameters, including the random effects. In the context of new style random effects, this likelihood might be adjusted or penalized to incorporate additional constraints or penalties via the random effects. In other words, the random effects distribution is the means by which these constraints and penalties are added. From a Bayesian perspective, these additional constraints or penalties can be seen as incorporating a "structural" prior on the random effects distribution. This prior acts similarly to the penalties mentioned above, serving to constrain the random effects in a way that aligns with the underlying assumptions of the model. On the other hand, a classical statistician might view such a prior as a constraint on the likelihood, influencing the estimates of the random effects in a way that enhances the overall fit and parsimony of the model.

In the Bayesian paradigm, Bayes' Theorem combines the prior distribution with the likelihood to form the posterior distribution. The posterior distribution describes the updated beliefs about the random effects after observing the data. In essence, it is a compromise between what the data tells us (through the likelihood) and what we believed *a priori* (through the prior).

In Bayesian inference, our goal is to compute the posterior predictive distribution of random variables representing model parameters given the observed data. As an example, consider the regression model:

$$y_i = X_i\beta + \epsilon_i \tag{2.16}$$

where $\epsilon_i \sim N(0, \sigma^2 \underset{\sim}{V_i})$.

In a Bayesian approach, prior distributions are set for the parameters:

For the coefficients (fixed effects), we set:

$$\beta \sim N(0, \Sigma_\beta) \tag{2.17}$$

For the error variance, we assume:

$$\sigma^2 \sim \text{Inverse-Gamma}(\alpha, \gamma) \tag{2.18}$$

Assuming that the priors for $\underset{\sim}{\beta}$ and σ^2 are independent, the posterior distribution, given observed data, is then:

$$p(\beta, \sigma^2 | y) \propto p(y|\beta, \sigma^2) p(\beta) p(\sigma^2) \tag{2.19}$$

This reflects the updated belief about the parameters after observing the data. To obtain estimates or predictions, one would sample from this posterior distribution. With random effects and potentially high-dimensional data, exact analytical solutions for the posterior distribution can be elusive. This is where Markov Chain Monte Carlo (MCMC) methods, like the Gibbs sampler or the Metropolis-Hastings algorithm, become invaluable. They allow for sampling from the posterior distribution to make statistical inferences about the random effects and other parameters. When new-style penalties are incorporated, specialized algorithms or adjustments might be necessary to efficiently sample from the posterior [63].

One of the most promising features of Bayesian random effects models is the natural shrinkage that occurs, especially with hierarchical models. This can be represented in the equations of the regression model by extending them to include a random effects term:

$$y = X\beta + Zu + \epsilon \tag{2.20}$$

Here, y is the response variable, X and Z are matrices of fixed and random effects covariates respectively, β is a vector of fixed effects coefficients, u is a vector of random effects, and ϵ is a vector of errors. The inclusion of the random effects term, Zu, allows the model to account for the variability between different levels or groups in the data.

Random effects, especially those defined for levels with few observations, are shrunk towards a global mean, reducing the risk of overfitting. This shrinkage effect is conceptually similar to the penalization in "new style" random effects models, where the idea is to avoid overfitting and derive more stable estimates, especially in the presence of sparse data. [53].

The Bayesian framework for random effects focuses on capturing group-specific variations. As such, the model complexity arises from the hierarchical structure, and interpretations are often about differences between groups or levels. This framework is ideal for data with inherent group structures like longitudinal studies, nested designs, or any scenario where it is crucial to account for non-independence within groups.

2.1.4.4 Bayesian LASSO

Bayesian models are particularly convenient when accounting for unobserved heterogeneity in data, especially hierarchical or clustered data. Such models capture variations that are not explained by the fixed effects by introducing random effects that can vary across levels of a grouping factor (like subjects, regions, or time points). The Bayesian LASSO (Least Absolute Shrinkage and Selection Operator) represents an intriguing fusion of Bayesian inference and regularization techniques, primarily geared towards variable selection and shrinkage in regression models. When cast into the realm of random effects, the Bayesian LASSO offers a potent tool to handle high-dimensional data structures, account for unobserved heterogeneity, and mitigate the risks of overfitting [95].

As noted, LASSO, in its traditional form, implements regularization by adding a penalty term to the likelihood function, which is proportional to the absolute values of the coefficients. This L1 penalty, interpreted from a Bayesian perspective as a Laplace prior on the regression coefficients, leads to a "double-exponential" distribution. This distribution is characterized by a sharp peak at zero and heavier tails than a Gaussian distribution. This structure enforces sparsity, setting some coefficients to zero, while allowing others to be large if the data supports it, as described by Park and Casella (2008).

In the Bayesian LASSO approach, as applied by Park and Casella, this penalty is specifically targeted at the fixed effect parameters, drawing smaller estimates towards zero. This method resembles early concepts of random effects distributions, as it is applied across a collection of parameters. In the context of Generalized Linear Mixed Models (GLMMs), this LASSO-type distribution can similarly be applied to a group of random effects, as seen in models like disease mapping. Here, the predicted random effects are more aggressively shrunk towards zero compared to the typical Gaussian (L2) random effect distributions. The application of the LASSO-type distribution to a collection of random effects also introduces a "structural prior," resulting in stronger shrinkage than that seen with Gaussian random effect distributions. Consider the linear regression setting:

$$y = X\beta + \epsilon \tag{2.21}$$

where $\epsilon \sim N(0, \sigma^2 I)$.

In the Bayesian Lasso, the L1 penalty of Lasso regression is integrated within a Bayesian framework by imposing a Laplace prior on the coefficients:

$$\beta_j | \lambda \sim \text{Laplace}(0, \lambda) \tag{2.22}$$

with density:

$$f(\beta_j | \lambda) = \frac{\lambda}{2} \exp(-\lambda |\beta_j|) \tag{2.23}$$

for $j = 1, ..., p$.

A common choice for a hyperprior on λ is:

$$\lambda^2 \sim \text{Exponential}(a) \tag{2.24}$$

where a is a predetermined hyperparameter.

The hierarchical formulation is then:

$$\begin{aligned} y | \beta, \sigma^2 &\sim N(X\beta, \sigma^2 I) \\ \beta_j | \lambda &\sim \text{Laplace}(0, \lambda) \\ \lambda^2 &\sim \text{Exponential}(a) \end{aligned}$$

When integrated into a random effects setting, a Bayesian model with LASSO priors assigned to the collection of random effects can induce sparsity and shrinkage among the random effects. Imagine a hierarchical model with several groups or clusters, where each group might have its set of predictors. By placing a Laplace prior on the group-specific coefficients, the Bayesian LASSO encourages a parsimonious representation. Some group-specific predictors might have their effects shrunk to zero, indicating they do not significantly vary across groups or do not offer additional explanatory power over the global (fixed) effects.

Furthermore, the random effects themselves can be subjected to the LASSO-type distribution. This can be particularly advantageous when dealing with a high number of random effects or when there is a suspicion that only a subset of random effects significantly contributes to the outcome.

Estimation in Bayesian LASSO models, especially with random effects, can be challenging due to the non-conjugacy of the Laplace prior. Advanced sampling techniques, such as Hamiltonian Monte Carlo (often implemented in software like Stan) or adaptive Metropolis-Hastings, are typically deployed. These techniques ensure efficient and accurate sampling from the posterior distribution, allowing for robust inferences about the random effects and other parameters under the LASSO constraint [95].

Bayesian LASSO with random effects presents a holistic approach to model high-dimensional data structures. By marrying the strengths of Bayesian inference with the regularization of the LASSO, it offers a robust framework to decipher complex relationships, enforce sparsity, and mitigate overfitting, all while accounting for unobserved heterogeneity.

2.1.4.5 Distinction Between Bayesian LASSO and L1 Norm Random Intercept

While both Bayesian LASSO and the L1 norm random intercept models utilize principles of shrinkage and sparsity, they are applied differently and serve distinct purposes in statistical modeling. The Bayesian LASSO is a technique primarily used for variable selection and shrinkage in the context of regression models. In contrast, the L1 norm random intercept model, as proposed by Besag et al. (1995) [13], applies similar principles but focuses on spatial shrinkage, particularly for handling random intercepts in hierarchical models.

The L1 norm random intercept model is designed to achieve spatial shrinkage, enabling the model to handle spatial data effectively. This approach is particularly beneficial in models where there are spatial dependencies, such as in geographical or environmental data. The L1 norm is applied to the random intercepts, allowing for sharper transitions than the typical L2-based conditional autoregressive regression (CAR) priors. This is in contrast to the overall shrinkage towards zero that the Bayesian LASSO seeks to achieve with its Laplace prior applied to regression coefficients [14].

In a hierarchical model with spatial components, employing an L1 norm random intercept can lead to better modeling of spatial variations. This is because it allows the intercepts to vary more flexibly across different spatial units while still enforcing a degree of sparsity. This sparsity is crucial in models with a large number of spatial units, as it helps in reducing the complexity of the model and avoiding overfitting.

The primary difference between the two approaches lies in their application and the type of shrinkage they induce. While the Bayesian LASSO imposes shrinkage on the regression coefficients, driving some of them to zero to achieve variable selection, the

L1 norm random intercept model focuses on achieving spatial shrinkage, allowing for more nuanced variations between spatially correlated intercepts.

In summary, while both Bayesian LASSO and L1 norm random intercept models utilize L1 penalties to induce sparsity, their applications differ significantly. The Bayesian LASSO is used for variable selection in regression models, whereas the L1 norm random intercept model is suited for spatial models, focusing on the random intercepts to handle spatial heterogeneity and dependencies. Understanding these distinctions is essential for appropriately applying these methods to various types of data and research questions.

2.1.5 Contrasting Prediction and Estimation

One of the foundational aspects of statistical modeling is understanding variance, be it for prediction or estimation. Both prediction variance and estimation variance provide insights into the accuracy and precision of our models, but they operate under different paradigms and carry distinct implications.

At its core, estimation is about determining the values of parameters (fixed or random effects) that best describe the observed data. Estimation variance quantifies the variability of these parameter estimates if the study were to be repeated numerous times. In the context of the "old-style" random effects, estimation variance is centered on the quantification of variance attributed to different hierarchical levels or groupings within a dataset. The aim was to capture and partition variability to understand the structure of the observed data. The primary focus on the variability of these estimates reflects the uncertainty about the true parameter value given the observed data.

On the other hand, prediction variance pertains to the variability expected when predicting new, unobserved outcomes based on the model. It combines the variability from the estimation of the parameters with the inherent variability of the new data points. This is particularly salient in the "new-style" random effects, where the emphasis shifted from merely estimating variance components to predicting outcomes based on a more in-depth understanding of the data's structure, including predictive values of the individual random effects (e.g., the random effect for each small area in disease mapping models). BLUP, for instance, thrives in this domain as it minimizes prediction error variance. It is all about forecasting the unknown, and prediction variance provides an idea of the spread or dispersion of these forecasts around their mean.

In contrasting the two, it's important to clarify that estimation variance pertains to the uncertainty in estimating fixed effect values in a model, whereas prediction variance encompasses this uncertainty as well as the variability in predicting values of random effects in new data points. This distinction is critical in practical applications. For example, a model may yield precise estimates (low estimation variance) for fixed effects using training data, but when predicting new observations, it faces the additional challenge of accurately predicting random effects. This can result in high prediction variance, indicating a wider spread or dispersion in forecasts for new, unseen data, as the model must account for both fixed effect estimation and random effect prediction.

Understanding both prediction and estimation variance is vital for researchers and practitioners. While the former tells us about the model's generalizability and ability to forecast new data points, the latter provides insights into the accuracy of the parameter estimates. As the field moved from "old-style" to "new-style" random effects, the balance between these two types of variance became even more significant, with the recognition that models need to be both accurate in estimation and effective in prediction for them to be truly valuable in applied settings.

2.1.6 Discussion

The intricate landscape of random effects theory underscores a pivotal distinction in statistical modeling: the contrast between prediction and estimation. Each method discussed in our review serves primarily one of these purposes, and understanding this dichotomy is fundamental for appropriate model selection.

BLUP stands as a classic testament to the power of prediction. By aiming to derive the best linear unbiased predictions, it leans heavily towards harnessing data to make accurate forecasts about unobserved entities or future observations. Its methodology prioritizes reducing prediction errors, which, while invaluable, might not always yield insights into the underlying structure of the data.

"Old-style" and "new-style" random effects serve to highlight the evolution in estimation techniques. The "old-style" method provides a more rigid structure for estimation and might lack flexibility in adapting to more complex data scenarios. Conversely, new-style random effects, by considering spatial and temporal dependencies, offer a nuanced approach to estimation, digging deeper into understanding the intrinsic data patterns and relationships.

Historically, random effects modeling was centered around the paradigm of estimation. The "old-style" random effects approach primarily aimed to quantify the variance attributed to different hierarchical levels or groupings within a dataset. The primary focus was on capturing and partitioning variability, resulting in models that, while robust for many classical datasets, adhered to rigid variance structures. In this realm, the end-goal was often to obtain estimates of variance components that could provide insight into the relative contribution of different factors or groupings to the overall variability in the outcome of interest.

With the advancement of computational power and the increased complexity of modern datasets, the focus shifted towards prediction. The "new-style" random effects models evolved to address this changing landscape. By embracing spatial, temporal, and other intricate dependencies, these models not only aim to estimate variance components but also predict outcomes based on a more detailed understanding of the data's structure. This shift towards prediction acknowledges the value of harnessing the rich information contained within datasets, allowing for more accurate forecasts and data-driven decisions.

In this shift from "old-style" to "new-style" random effects, the focus has shifted from simply estimating model parameters to also predicting random effects. In this approach, random effects are first predicted, and then the model parameters are estimated conditional on those predictions. This process is facilitated by the hierarchical structure of the models, which helps keep track of the conditioning and ensures that all estimates and predictions are made in the correct context. This progression from estimation to predictive modeling reflects a broader evolution in statistical modeling, where the focus has expanded from merely capturing the fixed effects and variance components to make informed forecasts that leverage all available information. This forward-looking approach is crucial in many modern applications, where understanding the current data structure and making accurate predictions for future data points can lead to actionable insights.

Penalized Maximum Likelihood (PML) further emphasizes the primacy of estimation, but with a crucial twist. When applied to spatial data, PML incorporates a smoothing element, ensuring that estimates are spatially coherent. By introducing a penalty term, it nudges the model towards achieving spatial continuity, offering a bridge between raw estimates and more refined, smoothed-out representations.

The Bayesian framework and Bayesian LASSO bring forth an amalgamation of both worlds. While they inherently focus on updating estimates based on prior beliefs and observed data, they also have predictive power. Particularly, Bayesian LASSO stands out by harnessing the sparsity-inducing power of LASSO within a Bayesian context, streamlining both estimation of parameters and prediction of outcomes.

In essence, the trajectory of random effects models accentuates the continuous interplay between prediction and estimation. While some methods prioritize one over the other, the true art lies in discerning which approach aligns best with the research question at hand. As data continues to grow in complexity, the models and methods we choose will inevitably need to strike a balance between these two pillars of statistical modeling.

In conclusion, the concept of random effects serves as both a reflection on the past and a glimpse into the future of statistical modeling. The models and methodologies discussed are not mere tools but represent the evolving thought processes of the statistical community. As we venture forward, these theories will not only guide our analyses but also shape the very questions we seek to answer.

2.2 Bayesian Modeling

As discussed in Chapter 1, statistical modeling often starts with a vector of observed data, denoted as $\mathbf{Y} = (Y_1, \ldots, Y_n)$, which we assume is generated from a probability distribution characterized by a set of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. The objective of statistical inference is to make the best possible estimates of these unknown parameters based on the observed data.

Under the classical likelihood framework, parameter estimation is typically carried out by maximizing the likelihood function, $L(\boldsymbol{\theta} \mid \mathbf{Y})$, defined as:

$$L(\boldsymbol{\theta} \mid \mathbf{Y}) = \prod_{i=1}^{n} f(Y_i \mid \boldsymbol{\theta}),$$

where $f(Y_i \mid \boldsymbol{\theta})$ represents the probability density function of the data given the parameters $\boldsymbol{\theta}$. The parameter estimates, $\hat{\boldsymbol{\theta}}$, are obtained by finding the values that maximize this likelihood function. In this framework, $\boldsymbol{\theta}$ is treated as fixed, and uncertainty about $\boldsymbol{\theta}$ is typically expressed through confidence intervals.

In contrast, the Bayesian approach treats the parameters $\boldsymbol{\theta}$ as random variables and assigns a probability distribution to them, known as the prior distribution, denoted by $f(\boldsymbol{\theta})$. The core of Bayesian inference lies in updating this prior distribution in light of the observed data $\mathbf{Y}$ to obtain the posterior distribution using Bayes' Theorem:

$$f(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{f(\mathbf{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\mathbf{Y})},$$

where $f(\mathbf{Y} \mid \boldsymbol{\theta})$ is the likelihood function and $f(\mathbf{Y})$ is the marginal likelihood, computed as:

$$f(\mathbf{Y}) = \int f(\mathbf{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Since $f(\mathbf{Y})$ is independent of $\boldsymbol{\theta}$, the posterior distribution is often expressed up to a normalizing constant as:

$$f(\boldsymbol{\theta} \mid \mathbf{Y}) \propto f(\mathbf{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}).$$

The posterior distribution provides a full probabilistic description of the uncertainty about the parameters after observing the data. Point estimates can be obtained by summarizing this distribution, typically using the posterior mean or median. Uncertainty in Bayesian inference is quantified through credible intervals, which provide a direct probabilistic interpretation [8].

The distinction between frequentist and Bayesian interpretations of probability is central to the philosophical differences between the two approaches. While frequentists interpret probability as the long-run relative frequency of an event, Bayesians interpret probability as a measure of belief or plausibility, which can be updated with new evidence. This difference in interpretation is fundamental to how each approach handles uncertainty and inference [72].

2.2.1 Prior Distributions

A key element of Bayesian inference is the specification of the prior distribution, $f(\boldsymbol{\theta})$. The prior encapsulates all prior knowledge or beliefs about the parameters before observing the data. Priors can be informative, incorporating strong prior knowledge, or non-informative (or weakly informative), designed to have minimal influence on the posterior. Examples of weakly informative priors include the normal distribution with a large variance, $N(0, \sigma^2)$, or the uniform distribution over a broad range. Non-informative priors can also take the form of improper priors, such as the uniform distribution over the entire real line, but care must be taken as these can lead to improper posterior distributions [72].

In certain cases, conjugate priors are used, which result in a posterior distribution that is in the same family as the prior. For example, if the likelihood is defined for independent observations following a normal distribution with unknown mean but known variance, a normal prior on the mean is conjugate, resulting in a normal posterior distribution. Conjugate priors simplify the analytical computation of the posterior but may not always be appropriate depending on the context of the model and data [28].

2.2.2 Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) methods are integral to modern Bayesian inference, especially when dealing with complex models where direct sampling from the posterior distribution is infeasible. MCMC algorithms allow us to generate samples from the posterior distribution by constructing a Markov chain that converges to the desired distribution. These samples can then be used to estimate posterior summaries such as means, variances, and credible intervals.

MCMC methods rely on the construction of a Markov chain, a sequence of random variables where the distribution of each variable depends only on the previous one. The key property of MCMC is that, under certain conditions, the chain will converge to a stationary distribution, which is the target posterior distribution in Bayesian analysis [50].

Two of the most commonly used MCMC algorithms are the Gibbs sampler and the Metropolis-Hastings algorithm.

2.2.2.1 Gibbs Sampling

First proposed by Geman and Geman (1984), Gibbs sampling is a special case of MCMC where the full conditional distributions of each parameter are known and can be sampled from directly. Suppose we have a set of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, and let $f(\theta_j \mid \boldsymbol{\theta}_{-j}, \mathbf{Y})$ denote the conditional distribution of θ_j given all other parameters $\boldsymbol{\theta}_{-j}$ and the observed data $\mathbf{Y}$. The Gibbs sampler proceeds by iteratively sampling from these conditional distributions:

$$\theta_1^{(t+1)} \sim f(\theta_1 \mid \theta_2^{(t)}, \ldots, \theta_p^{(t)}, \mathbf{Y}),$$

$$\theta_2^{(t+1)} \sim f(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_p^{(t)}, \mathbf{Y}),$$

$$\vdots$$

$$\theta_p^{(t+1)} \sim f(\theta_p \mid \theta_1^{(t+1)}, \ldots, \theta_{p-1}^{(t+1)}, \mathbf{Y}),$$

where t denotes the iteration number. This process is repeated until the chain converges, beyond which point subsequent samples can be used to approximate the posterior distribution [54].

2.2.2.2 Metropolis-Hastings Algorithm

Proposed by Metropolis (1953) and later extended by Hastings (1970), The Metropolis-Hastings algorithm is a more general MCMC method that can be applied when the full conditional distributions are not easily sampled from. At each step, the algorithm generates a candidate value θ^* from a proposal distribution $q(\theta^* \mid \theta^{(t)})$, where $\theta^{(t)}$ is the current value of the parameter. The candidate value is accepted with probability:

$$\alpha = \min\left(1, \frac{f(\theta^* \mid \mathbf{Y}) q(\theta^{(t)} \mid \theta^*)}{f(\theta^{(t)} \mid \mathbf{Y}) q(\theta^* \mid \theta^{(t)})}\right),$$

and rejected otherwise. If the candidate is rejected, the current value is retained for the next iteration. This algorithm is particularly useful in complex models where Gibbs sampling is not feasible [86, 60].

2.2.3 Efficient Posterior Sampling in Bayesian Inference

In Bayesian inference, the goal is often to approximate the posterior distribution of model parameters, particularly when dealing with complex models such as those involving Gaussian Markov Random Fields (GMRFs). Markov Chain Monte Carlo (MCMC) methods are typically employed for this purpose due to their flexibility and general applicability. However, MCMC can be challenging to implement efficiently for GMRFs, given the large number of parameters and the correlations among them.

2.2.3.1 Challenges with MCMC for GMRFs

When applying MCMC to GMRFs, one of the primary challenges is the potential for poor mixing and slow convergence due to the high-dimensional parameter space and the inherent correlations among the parameters. To address these issues, several strategies have been developed. One effective approach is block updating, where groups of parameters are updated jointly rather than individually. This method has been shown to improve both the mixing and convergence rates of the MCMC algorithm [71, 100]. In block updating, joint proposals are made for parameter blocks, and the acceptance or rejection of these proposals is done simultaneously, which can lead to more efficient exploration of the parameter space.

Another method that has proven useful for latent Gaussian models is elliptical slice sampling [89]. This technique also involves jointly proposing updates for the field parameters, allowing for more effective navigation of the posterior distribution. However, a more general and highly efficient MCMC method is Hamiltonian Monte Carlo (HMC). First proposed by Neal (2011), Hamiltonian Monte Carlo (HMC) is an advanced MCMC technique that leverages concepts from Hamiltonian dynamics to efficiently explore high-dimensional probability distributions. Originally developed for use in physics, HMC has been adapted for Bayesian inference and is particularly well-suited to sampling from complex posterior distributions [90].

2.2.3.2 Hamiltonian Dynamics

Hamiltonian dynamics describe the evolution of a physical system in a state space defined by position and momentum variables. For a system in a d-dimensional space, the state of the system at any given time is described by the position vector $\mathbf{q}$ and the momentum vector $\mathbf{p}$. The total energy of the system is represented by the Hamiltonian function $H(\mathbf{q}, \mathbf{p})$, which is the sum of the potential energy $U(\mathbf{q})$ and the kinetic energy

$K(\mathbf{p})$:

$$H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + K(\mathbf{p}).$$

The potential energy $U(\mathbf{q})$ is often associated with the height of a surface at position $\mathbf{q}$, while the kinetic energy $K(\mathbf{p})$ is given by:

$$K(\mathbf{p}) = \frac{\|\mathbf{p}\|^2}{2m},$$

where m represents the mass of the object, and $\|\mathbf{p}\|$ is the norm of the momentum vector.

The evolution of the system over time is governed by Hamilton's equations:

$$\frac{d\mathbf{q}_i}{dt} = \frac{\partial H}{\partial \mathbf{p}_i} = \frac{\partial K}{\partial \mathbf{p}_i},$$

$$\frac{d\mathbf{p}_i}{dt} = -\frac{\partial H}{\partial \mathbf{q}_i} = -\frac{\partial U}{\partial \mathbf{q}_i},$$

for $i = 1, \ldots, d$. These equations describe how the position and momentum variables change over time, and they preserve the total energy of the system.

2.2.3.3 HMC in Bayesian Inference

In the context of Bayesian inference, the position variables $\mathbf{q}$ correspond to the parameters of the model, and the potential energy function $U(\mathbf{q})$ is defined as the negative log-posterior distribution:

$$U(\mathbf{q}) = -\log P(\mathbf{q} \mid \mathbf{D}) = -\log P(\mathbf{q}) - \log L(\mathbf{q} \mid \mathbf{D}),$$

where $P(\mathbf{q})$ is the prior distribution, and $L(\mathbf{q} \mid \mathbf{D})$ is the likelihood function given the data $\mathbf{D}$.

The momentum variables $\mathbf{p}$ are auxiliary variables introduced to facilitate sampling, and their distribution is typically Gaussian, governed by the kinetic energy function. The joint distribution of $\mathbf{q}$ and $\mathbf{p}$ is given by:

$$P(\mathbf{q}, \mathbf{p}) \propto \exp(-H(\mathbf{q}, \mathbf{p})) = \exp(-U(\mathbf{q})) \exp(-K(\mathbf{p})).$$

HMC uses the dynamics described by Hamilton's equations to propose new states $(\mathbf{q}^*, \mathbf{p}^*)$ in the Markov chain. The advantage of HMC is that it can make large moves in the parameter space while maintaining a high acceptance probability, which is determined by the Metropolis criterion:

$$\text{Accept}\,(\mathbf{q}^*, \mathbf{p}^*)\,\text{with probability}\,\min\left[1, \exp(-\Delta H)\right],$$

where $\Delta H = H(\mathbf{q}^*, \mathbf{p}^*) - H(\mathbf{q}, \mathbf{p})$ is the change in Hamiltonian.

2.2.3.4 Challenges and Advances in HMC

While HMC offers significant advantages in exploring complex posterior distributions, it also has some challenges. One challenge is the need to compute gradients of the log-posterior, which can be analytically difficult or computationally expensive. Techniques such as automatic differentiation or numerical differentiation can be employed to overcome this issue [84].

Another challenge is the sensitivity of HMC to the choice of step size ϵ and the number of steps L. Poor choices can lead to inefficient sampling or even divergence of the algorithm. To address this, Hoffman (2014) proposed the No-U-Turn Sampler (NUTS), an adaptive variant of HMC that automatically adjusts the step size and number of steps, greatly improving the algorithm's robustness.

2.2.3.5 The No-U-Turn Sampler (NUTS)

The No-U-Turn Sampler (NUTS) is an advanced variant of HMC designed to overcome some of the practical challenges associated with the standard HMC algorithm, particularly the sensitivity to the choice of step size and the number of leapfrog steps. NUTS dynamically adjusts the trajectory length during the sampling process, thereby eliminating the need to manually tune the number of steps [66].

Dynamic Step Adjustment in NUTS

In standard HMC, the trajectory length, which is determined by the number of leapfrog steps L and the step size ϵ, must be carefully chosen to ensure efficient exploration of the posterior distribution. Too few steps result in poor exploration, while too many steps can lead to excessive computation without significant gains in sampling efficiency. NUTS addresses this issue by automatically deciding when to stop the trajectory based on the path it follows in the parameter space.

NUTS works by simulating multiple potential trajectories from the current position $\mathbf{q}$ and determining on-the-fly how far to extend each trajectory. The key innovation of NUTS is its ability to detect when a trajectory is about to double back on itself, or "make a U-turn," which would indicate that further steps are unlikely to contribute to efficient exploration. When such a U-turn is detected, NUTS terminates the trajectory and uses the points along the path to generate a proposal for the next state in the Markov chain.

The NUTS Algorithm

The NUTS algorithm can be summarized as follows:

1. **Initialization:** Start at the current state $(\mathbf{q}, \mathbf{p})$, where $\mathbf{q}$ represents the model

parameters, and $\mathbf{p}$ is the auxiliary momentum variable drawn from its Gaussian distribution.

2. Recursive Tree Building: NUTS builds a binary tree of possible states by repeatedly applying the leapfrog integrator. The tree is constructed such that each node represents a candidate state $(\mathbf{q}^*, \mathbf{p}^*)$, and the tree expands until a U-turn is detected. This is done by checking if the momentum $\mathbf{p}$ has changed direction relative to the trajectory's initial momentum, which would indicate that further exploration along this path is unnecessary.

3. Termination and Proposal Selection: Once a U-turn is detected, NUTS stops the trajectory and selects a proposal state $(\mathbf{q}', \mathbf{p}')$ from the set of candidate states generated during the tree-building process. The proposal is selected with a probability proportional to the target distribution, ensuring that the sampling process remains unbiased.

4. Metropolis Acceptance Criterion: The selected proposal $\mathbf{q}'$ is accepted or rejected based on the standard Metropolis criterion, which compares the Hamiltonian at the current and proposed states. If the proposal is accepted, $\mathbf{q}'$ becomes the new state; otherwise, the algorithm retains the current state.

5. Adaptation: NUTS includes an adaptation phase where the step size ϵ is adjusted during the warm-up (or burn-in) period to optimize the acceptance rate. This adaptive mechanism helps to balance the trade-off between exploration efficiency and computational cost.

2.2.3.6 Applications of HMC in Hierarchical Models

Hamiltonian Monte Carlo (HMC) is particularly powerful when applied to hierarchical models, which are common in many Bayesian statistical applications. Hierarchical models often involve multiple levels of parameters, where parameters at one level depend on those at another. This layered structure introduces dependencies among parameters, which can lead to slow mixing and poor convergence in standard MCMC methods, such as Gibbs sampling.

2.2.3.6.1 The Challenge of Centered Parameterization In hierarchical models, a common issue arises with what is known as the centered parameterization. Consider a simple hierarchical model where the observed data y_i are normally distributed with mean θ_i and known variance σ^2, and the θ_i themselves are drawn from a normal distribution with mean 0 and variance τ^2:

$$y_i \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma^2), \quad \theta_i \sim \mathcal{N}(0, \tau^2).$$

In this centered parameterization, there can be strong correlations between the parameters θ_i and the hyperparameter τ, particularly when τ is small. This correlation can cause inefficient sampling because small changes in τ can lead to large changes in the θ_i, making the posterior distribution difficult to explore.

2.2.3.6.2 Non-Centered Parameterization To address this issue, a reparameterization strategy known as non-centered parameterization is often employed. This technique involves transforming the model in such a way that the dependencies among parameters are reduced, leading to better mixing in the HMC algorithm. Specifically, the model can be reparameterized by introducing a new variable z_i defined as:

$$\theta_i = \tau z_i, \quad z_i \sim \mathcal{N}(0, 1).$$

The model now becomes:

$$y_i \mid z_i, \tau \sim \mathcal{N}(\tau z_i, \sigma^2), \quad z_i \sim \mathcal{N}(0, 1).$$

In this non-centered parameterization, the z_i are independent of τ a priori, which reduces the correlation between the parameters and the hyperparameters. This change often leads to more efficient sampling by HMC, as the sampler can move more freely in the parameter space without being constrained by the strong dependencies present in the centered parameterization.

2.2.3.6.3 Benefits of HMC in Hierarchical Models The use of HMC in hierarchical models, particularly when combined with non-centered parameterization, offers several advantages:

- Efficient Exploration: HMC is well-suited to high-dimensional spaces, and its ability to make large, informed jumps in the parameter space allows it to explore the posterior distribution more efficiently than traditional MCMC methods. This is especially important in hierarchical models, where the parameter space can be complex and multi-modal.

- Reduction of Autocorrelation: By reducing the correlation between parameters through non-centered parameterization, HMC can reduce the autocorrelation in the Markov chain, leading to faster convergence and more reliable estimates of the posterior distribution.

- Improved Convergence: HMC's reliance on gradient information helps it navigate the posterior landscape more effectively, which can lead to improved convergence properties, particularly in models with challenging posterior geometries.

2.2.3.6.4 Practical Implementation in Stan The Stan software package (Stan Development Team, 2015) has popularized the use of HMC and its variants, such as the No-U-Turn Sampler (NUTS), in hierarchical models. Stan automates the implementation of non-centered parameterizations where appropriate, making it easier for users to apply these techniques without requiring deep expertise in MCMC methods.

Moreover, Stan leverages automatic differentiation to efficiently compute the gradients required by HMC, further enhancing its utility in complex models. This capability is particularly valuable in hierarchical models, where the posterior distribution's complexity can make analytical gradient calculations infeasible.

2.2.3.6.5 Case Studies and Applications HMC has been successfully applied in various hierarchical modeling scenarios, such as in the analysis of multi-level data, random effects models, and spatio-temporal models. For instance, in disease mapping, hierarchical models are used to account for spatial correlations between regions, and HMC has been employed to estimate the posterior distribution of disease risk across different geographic areas. By using non-centered parameterizations, researchers have been able to obtain more accurate and reliable estimates of spatial patterns in disease risk.

Overall, the application of HMC in hierarchical models represents a significant advancement in Bayesian computation, allowing for the effective handling of complex models that were previously challenging to estimate. The combination of HMC with non-centered parameterization techniques provides a robust framework for addressing the unique challenges posed by hierarchical structures in Bayesian modeling.

Chapter 3

Hierarchical Bayesian Disease Mapping Model Performance in the Presence of Spatially Concentrated Prevalence

3.1 Introduction

In Chapter 1, we provided a detailed review of tuberculosis in the United States as well as the concept of disease mapping. The focal point of this chapter is the application and comparison of hierarchical Bayesian models, specifically the Besag, York, and Mollié (BYM), BYM2, and Intrinsic Conditional Autoregressive (ICAR) models, in disease mapping [120, 77, 97] and the introduction of a horseshoe prior [24] to the disease mapping literature. These models are adept at depicting spatial variations in disease risk and are essential for identifying areas with heightened risk, which is critical for formulating effective public health interventions. Traditionally, these models employ Gaussian priors, which are well-suited for shrinking estimates towards spatial neighbors, a process that assumes spatial continuity and reduces local noise by making neighboring areas appear more similar.

While shrinking to spatial neighbors can be beneficial, it may not always be appropriate, particularly in scenarios where disease rates are not spatially continuous or where there are sharp boundaries or isolated hotspots. For instance, in metropolitan and densely populated areas, the spatial dynamics of TB spread may involve abrupt changes in disease prevalence between neighboring regions. In such cases, models that overly shrink data towards spatial neighbors might obscure important spatial patterns and lead to misleading conclusions [81].

To address these challenges, this study not only compares traditional disease mapping models that use Gaussian priors with equivalent models incorporating Laplace priors but also introduces the Horseshoe prior as a promising alternative. The Laplace priors, which induce sparsity in the model, could provide a more refined and localized understanding of disease distribution, enabling more accurate identification of high-risk zones and sharp transitions in disease prevalence [14].

In addition, the Horseshoe prior is particularly suited for handling sparse signals and managing spatial dependencies, which are often characteristic of TB distribution. The Horseshoe prior applies global-local shrinkage to the model parameters, allowing for significant effects to be preserved while shrinking less significant effects toward zero. This approach is especially beneficial in settings where disease rates show discontinuities or isolated spikes, as it can prevent overfitting in sparse data while maintaining the ability to detect sharp contrasts in disease prevalence [24].

The Horseshoe model integrates the spatial structure of the data into the shrinkage process, making it ideal for cases where the underlying process is sparse, with only a few regions exhibiting significant effects. By comparing the BYM, BYM2, and ICAR models with both Laplace and Gaussian random effects to a horseshoe model, this study seeks to evaluate which approach most effectively captures the complex spatial patterns of TB incidence in Georgia.

Through this analysis, we contribute to the field of disease mapping by providing a nuanced understanding of TB distribution, particularly in complex urban settings. The outcomes of this study have the potential to significantly influence public health strategies, ensuring that interventions are more closely aligned with the localized realities of TB spread, thereby contributing to the global effort to reduce TB incidence and mortality.

3.2 Methods

3.2.1 Study Area

Our study encompasses the 125 ZIP codes that make up the 5 core counties of the Atlanta Metropolitan Statistical Area (MSA). We provide a detailed map illustrating the population distribution by ZIP code across these counties (see Figure 3.1). The ZIP code specific population size is relatively evenly distributed across the area, with the densest populations located in the northeastern suburbs. All population and socioeconomic data used in this study are sourced from the U.S. Census [112].

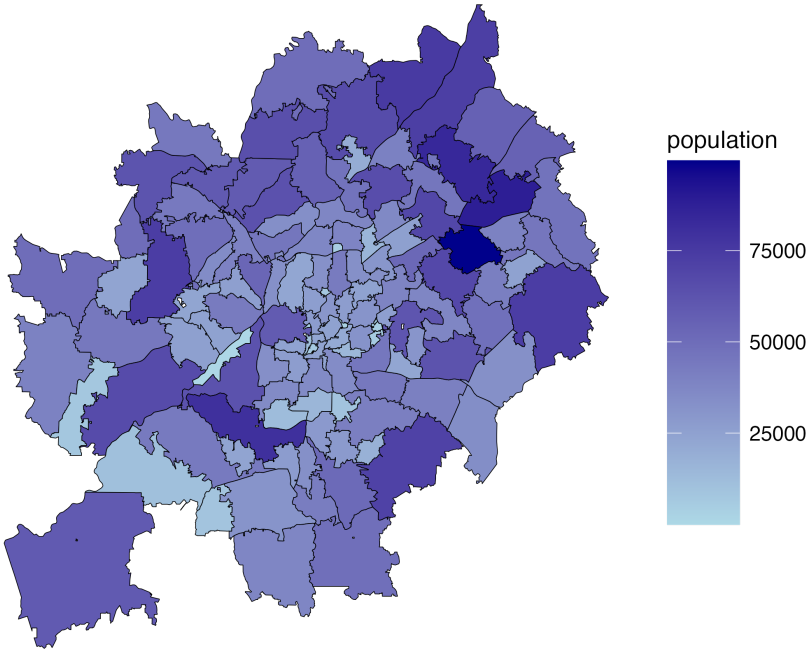


Figure 3.1: Distribution of population by ZIP code in metro Atlanta, Decennial Census (2020).

The CDC highlights that non-US born individuals significantly contribute to the annual incidence of TB, despite representing only 14% of the total U.S. population [26]. Limited access to healthcare is a known factor exacerbating TB prevalence in various communities [23, 22]. The American Community Survey provides data on the percentage of non-US born individuals by ZIP code, which is crucial for analyzing TB count variations [113]. Figure 3.2a demonstrates that, across the 125 ZIP codes in metro Atlanta, GA, there was an average of 15.33% non-US born residents. The lowest and highest percentages of foreign-born populations in individual ZIP codes were 0.88% (Moreland, GA) and 52.55% (Clarkston, GA), respectively, with no missing data at the ZIP code level.

Furthermore, our analysis incorporates critical factors commonly associated with higher TB rates, such as the incidence of HIV and socioeconomic inequality, as measured by the Gini index. We specifically examine the percentage of the population living with HIV (Figure 3.2b) and the Gini index [56], which reflects income inequality within these communities (Figure 3.2c), as these conditions are known to exacerbate the spread of TB [38, 68]. Our data reveal that the average prevalence of individuals living with HIV across these ZIP codes is approximately 0.197 cases per 1,000 people, with certain areas showing higher rates, highlighting the significant overlap between TB and HIV epidemics [3].

The Gini index ranges from 0, representing perfect equality where everyone has the same income or wealth, to 1, representing perfect inequality where one individual possesses all the income or wealth and others have none. The Gini index for ZIP codes in the study area ranges from 0.3443 to 0.5772, indicating a considerable variation in income inequality. Lower Gini index values suggest more equal income distribution, while higher values indicate greater income disparity. In the context of TB, higher income inequality, as reflected in the upper end of this range, may contribute to the concentration of TB cases in specific areas by exacerbating conditions of poverty, overcrowding, and limited access to healthcare. This variation in income inequality across the studied ZIP codes provides crucial insight into the social determinants of TB, suggesting that areas with greater income disparity might experience higher TB rates due to these compounded socioeconomic vulnerabilities. The data for these covariates, sourced from the U.S. Census and the American Community Survey, will be detailed in subsequent analyses to highlight their relationships with TB cases in the studied area.

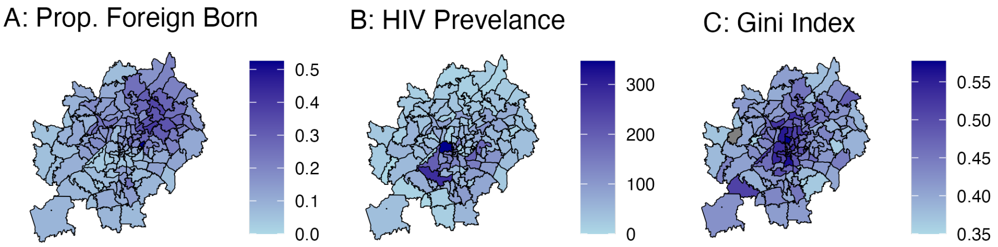


Figure 3.2: Proportion Foreign-Born, HIV prevalence, and Gini index by ZIP code in metro Atlanta, American Community Survey 1-year estimates (2022).

3.2.2 Bayesian Methodology for Disease Mapping

As detailed in Chapter 2, the Bayesian methodology is a widely used approach for modeling disease incidence data. The Bayesian methodology allows for the incorporation of prior knowledge and uncertainty into the modeling process, providing a flexible and coherent framework for estimating disease risk and generating disease maps. Bayesian methods have been successfully applied to a wide range of disease mapping problems, including tuberculosis incidence [27], cancer incidence [78], and infectious diseases [4].

3.2.3 Hierarchical Bayesian Disease Mapping Models

Disease mapping models aim to identify areas of elevated disease risk. In contrast to frequentist approaches that rely solely on data fit, these models also integrate subjective prior information and typically result in estimates with higher bias but less variance. A hierarchical Bayesian framework is particularly well-suited for disease mapping models as it stabilizes the risk estimated by shrinking (smoothing) unstable estimates to a weighted average of neighboring risks. It allows for borrowing of information across small areas and multiple populations to produce more stabilized rates of disease [80, 77].

In hierarchical Bayesian disease mapping models, the parameter of interest is often the *Standardized Incidence Ratio* (SIR), which measures the risk of disease in a specific area relative to a reference area [12]. The SIR for a small area i is denoted by θ_i and is defined as the ratio of the observed number of cases in the small area to the expected number of cases in that area, based on the overall or background rate of disease in the study population.

Mathematically, the SIR can be expressed as:

$$\theta_i = \frac{Y_i}{E_i} \tag{3.1}$$

where Y_i and E_i denote the observed and expected number of cases in the ith small area, respectively.

The SIR provides a measure of the relative risk of disease in a small area compared to the overall or background risk. If the SIR for a particular small area is greater than 1, it suggests that the area has a higher risk of disease compared to the reference population. Conversely, if the SIR is less than 1, it suggests a lower risk of disease in the area compared to the reference population. An SIR of 1 indicates that the risk of disease in the area is the same as the overall or background risk [12].

Hierarchical Bayesian models can estimate SIRs for each small area in the study region, allowing for the identification of areas with elevated disease risk. These models can also incorporate spatial dependence and other relevant covariates to improve the accuracy of risk estimates and identify potential risk factors for the disease of interest.

3.2.4 Three-Stage Bayesian Hierarchical Model

When modeling the risk of a rare disease across small areas, particularly when the population of an area is relatively large compared to the disease counts, the resulting risk estimates are subject to increased random variation [12]. This can make it difficult to determine whether extreme risk estimates are indicative of a true increased (or decreased) risk in a particular area or simply the result of extra variation due to small numbers.

To address this issue, a three-stage parameter hierarchy can be used to reduce the effect of variation on risk estimates by borrowing information from other regions. The first stage of the model defines the likelihood function of parameters representing spatial underlying disease risk to the data. In the second stage, random effects are incorporated to allow for similarities in disease risk based on spatial proximity in the form of a spatial prior distribution. Finally, in the third level, the parameters of the prior distribution are related to a hyper-prior distribution [71].

This three-stage hierarchical Bayesian model has been shown to be effective in estimating disease risk and identifying areas with elevated risk [12, 71]. By incorporating prior information and spatial dependence, the model can produce more stable and accurate risk estimates, making it a valuable tool in disease mapping and public health research.

3.2.4.1 Stage 1: Likelihood

The first stage of the hierarchy assumes that the count of the disease in a specific area follows a Poisson distribution given by:

$$Y_i \sim Pois(\mu_i) \tag{3.2}$$

where:

$$\mu_i = E_i\theta_i \tag{3.3}$$

The Poisson likelihood is expressed as:

$$L(\theta_i) = \prod_{i=1}^{m} \frac{e^{-E_i\theta_i}(E_i\theta_i)^{Y_i}}{Y_i!} \tag{3.4}$$

and the corresponding log-likelihood is given by:

$$\ell(\theta_i) = \sum_{i=1}^{m} E_i\theta_i + \sum_{i=1}^{m} y_i ln(E_i\theta_i) - \prod_{i=1}^{m} Y_i! \tag{3.5}$$

From here, we obtain the maximum likelihood estimator by setting the first derivative of the log-likelihood equal to zero, yielding:

$$\hat{\theta}_i = \frac{Y_i}{E_i}, \text{ SIR associated with region } i \tag{3.6}$$

where $\hat{\theta}_i$ is the estimate of the SIR in area i.

In a conjugate Bayesian formulation we may assign a gamma prior distribution to θ_i. In this case, $\hat{\theta}_i \sim \Gamma(a, b)$ where a and b are the shape and scale parameters given by:

$$P(\theta \mid a, b) = \frac{(\theta)^{a-1}}{\Gamma(a)b^a} e^{-\theta b} \tag{3.7}$$

When combined with the Poisson likelihood, this yields a gamma posterior distribution, where $\hat{\theta}_i$ has the following posterior distribution:

$$\hat{\theta}_i \sim Gamma(a + Y_i, b + E_i) \tag{3.8}$$

This Gamma-Poisson likelihood formulation is ideal for measuring risk in large areas such as a state, but for smaller areas such as counties or ZIP codes it is not as reliable since the mean and variance are based on E_i and can yield overly large estimates of risk in areas where the expected numbers of cases are low. Further, this likelihood formulation is very restrictive in that it does not easily account for covariate effects or allowing spatial correlation between risk in nearby areas [119]. In order to achieve such goals, more sophisticated models that can smooth the extreme values when estimating the risk and account for covariates and spatial correlation, these models are provided in Stage 2.

3.2.4.2 Stage 2: Random Effects Models

The second stage of the Bayesian hierarchical disease mapping model incorporates random effects to capture the unmeasured or unobserved risk factors in the study population. These random effects can be further divided into spatially structured and unstructured components, which account for the dependent and heterogeneous variations of risks in space [77].

One of the simplest models used in stage 2 is the Poisson Log-Normal model, which assumes that the count of the disease in a specific area follows a Poisson distribution. In this model, a non-spatial random effect term is added to explain the unstructured heterogeneity in SIRs for a particular disease in a specific region compared to the overall study area. The log-normal model is given by:

$$\theta_i = e^{\beta_0 + \beta X_i + V_i} \tag{3.9}$$

where $\beta_0 + \beta_{X_i}$ represents the overall log SIR of the disease in the study region compared to the reference rate, and β's are the weights describing the degree of change in the log SIR for every unit change in the covariates X_i. V_i is a normally distributed area-specific random effects term that captures the residual or unexplained (log) relative risk of the event in area i [12]. τ_v is the precision parameter of V_i and its density is given by:

$$P(\mathbf{V} \mid \tau_v) \propto \tau_v^{\frac{n}{2}} \left(\frac{1}{\sqrt{2\pi\tau_v}} \right)^n e^{-\frac{1}{2}\mathbf{V'Q_vV}} \sim MVN(\mathbf{0}, \tau_v \mathbf{I_n}) \tag{3.10}$$

where $\mathbf{V} = (V_1, ..., V_n)'$ is the vector of random effects, $\mathbf{Q_v} = \tau_v \mathbf{I_n}$ is the precision matrix and $\mathbf{I_n}$ is the $n \times n$ identity matrix [100].

The Poisson Log-Normal model is useful for capturing the unobserved heterogeneity in SIRs but it does not account for spatial correlation between areas, which may result in biased estimates. Addressing this required the addition of spatial random effects.

Conditional Autoregressive Models for Areal Data

Areal data, comprising measurements from defined regions, often exhibit variability in small populations, which obscures underlying patterns. To clarify these patterns, Conditional Autoregressive (CAR) models smooth this variability by sharing information across neighboring regions. In a network of N regions, the adjacency relation, denoted as $i \geq j$ (for $i \neq j$), indicates neighboring regions [15].

Spatial dependencies between units i and j are modeled with a normal random variable U, expressed as an N-dimensional vector $U = (U_1, \ldots, U_n) \mid \cdot$. Each U_i in its conditional distribution is influenced by the aggregated weighted values of its neighboring regions $(w_{ij}U_j)$, with conditional variance given by $\sum w_{ij}\tau$:

$$U_i \mid U_j, j \neq i \sim N(a, b), \quad a = \sum_{j=1}^{N} w_{ij}U_j \tag{3.11}$$

A Gaussian Markov random field (GMRF) is defined by translating these local conditional distributions into a global distribution. As established by Besag (1974), under some regularity conditions, the joint distribution for U conforms to a multivariate normal distribution centered at zero. The variance is given by a precision matrix Q, which is the inverse of the covariance matrix Σ ($\Sigma = Q^{-1}$) [11]:

$$U \sim N(0, Q^{-1}) \tag{3.12}$$

The construction of Q for U involves the adjacency matrix A and the diagonal matrix D. Matrix A indicates neighboring relationships, while D is the nxn diagonal matrix where entries i, i are the number of neighbors of region i and the off-diagonal entries are 0. In addition, $D_\tau = \tau D$. For a map with N regions, the connectivity is defined by A, and the number of connected components varies from 1 (fully connected) to N (completely disconnected). Consider a map with regions n_1, n_2, n_3, n_4, and adjacency relations ($1 \geq 2$, $2 \geq 3$, $3 \geq 4$)[120]. The matrices are defined as:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.13}$$

The precision matrix Q is symmetric and positive definite, formulated as $Q = D(I - \lambda A)$ with I the identity matrix and $0 < \lambda < 1$:

$$Q = \begin{bmatrix} 1-0.5 & -0.5 & 0 & 0 \\ -0.5 & 1.5 & -0.5 & 0 \\ 0 & -0.5 & 1.5 & -0.5 \\ 0 & 0 & -0.5 & 1-0.5 \end{bmatrix} \tag{3.14}$$

The log probability density of U is:

$$\log(\det(Q)) - \frac{1}{2}U^T Q U \tag{3.15}$$

Computing $\det(Q)$ is computationally intensive, particularly for a large N, impacting the efficiency of MCMC samplers that require recalculating U's density with each new proposal[6].

Intrinsic Conditional Autoregressive Models

The Intrinsic Conditional Autoregressive (ICAR) model simplifies the Conditional Autoregressive (CAR) model by setting a key parameter, denoted as α, to 1, effectively removing it from the model. This adjustment simplifies the precision matrix Q from $D(I - A)$ to just $D - A$. For instance, in the given example, this results in $D - A$ taking the form:

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

In this modified ICAR model, the determinant of Q becomes zero. It's important to note that the ICAR prior is considered improper, but it becomes proper when incorporating data [12].

When employing Markov Chain Monte Carlo (MCMC) samplers to calculate log probabilities, a notable advantage of the ICAR model is that the term $n^2 \log(\det(Q))$ remains constant, allowing it to be omitted during calculations. This reduces the computational complexity from N^3 to N^2, enabling efficient fitting of large areal map datasets using an MCMC sampler on standard laptops within hours rather than days [88].

In the ICAR model, each U_i follows a normal distribution with a mean equal to the average of its neighboring regions. The variance decreases as the number of neighbors, denoted as d_i, increases. The conditional specification of the ICAR model is expressed as:

$$p(U_i \mid U_{i \geq j}) = N\left(\frac{1}{d_i}\sum_{i \geq j} U_j, \frac{\tau^2}{d_i}\right)$$

Here, τ^2 represents the unknown variance.

The joint specification of the ICAR random vector U, centered at 0 with a common variance of 1, can be rephrased in terms of pairwise differences:

$$p(U) \propto \exp\left(\frac{1}{2}\sum_{i \geq j}(U_i - U_j)^2\right) \tag{3.16}$$

Expressing the joint density in terms of pairwise differences facilitates understanding of the model's behavior: each $(U_i - U_j)^2$ term introduces a penalty based on the difference between neighboring region values, ultimately encouraging spatial smoothing. It's important to note that the pairwise difference is non-identifiable, meaning that adding a constant to all U_i values does not affect the term $(U_i - U_j)$. To address this, the constraint $\sum_i U_i = 0$ is imposed, which centers the model. With this constraint, the log probability density is well-defined because the parameter integration is limited to the set of parameters that sum to 1 [12].

The Intrinsic Conditional Autoregressive (ICAR) model first proposed in 1991 by Besag, York and Mollié and in the case of our study is given by:

$$log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i \tag{3.17}$$

Here, U_i represents the spatially structured random effects and are modeled conditional on their neighbors given the equations [12] given by:

$$U_i \mid \underbrace{\{U_j = u_j, i \sim j\}}_{\text{Neighbors of } i}, \tau_u, \omega_{ij} \sim N\left(\bar{u_i}, \frac{1}{\sum_{j=1}^{n} \tau_u \omega_{ij}}\right) = ICAR(\omega_{ij}, \frac{1}{\tau_u}) \tag{3.18}$$

where

$$\bar{u_i} = \frac{\sum_{j=1}^{n} \omega_{ij} u_j}{\sum_{j=1}^{n} \omega_{ij}} \tag{3.19}$$

$\bar{u_i}$ represents the mean disease count of the neighbors of area i and ω_{ij} is an indicator that equals 1 where i and j are adjacent, 0 otherwise. $i \sim j$ denotes that areas i and j are neighbors. τ_u is the precision parameter of U_i. The density of vector

$\mathbf{U} = (U_1, ..., U_n)'$is given by

$$P(\mathbf{U} \mid \tau_u) \propto \tau_u^{\frac{n-1}{2}} e^{-\frac{\tau_u}{2} \sum_{i \sim j} (\mathbf{U_i} - \mathbf{U_j})^2}$$

$$\propto \tau_u^{\frac{n-1}{2}} e^{-\frac{1}{2} \mathbf{U}' \mathbf{Q_u} \mathbf{U}} \tag{3.20}$$

here, $\mathbf{Q_u} = \tau_u \mathbf{R_u}$ being the precision matrix and $\mathbf{R_u}$ the $n \times n$ spatial structure matrix [12, 101] defined as:

$$\mathbf{R_u} = \begin{cases} n_i & \text{if } i = j \\ -1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases} \tag{3.21}$$

where n_i is the number of neighbors of area i.

The Besag-York-Mollie (BYM) Model

The ICAR model is spatially adaptive in that the variation in the model is largely dependent on the neighborhoods leading to conditions where adjacency is used as a proxy for dependence. However, assuming spatial correlation between regions can be problematic in practice. For this reason the ICAR model has been modified to combine the spatial correlation from a traditional conditional autocorrelation model with the unstructured heterogeneity component from the Poisson Log Normal model to form a BYM model [12, 30]. This model in given by:

$$\theta_i = e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i + V_i} \tag{3.22}$$

$$log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i + V_i \tag{3.23}$$

$$U_i \sim ICAR(\omega_{ij}, \frac{1}{\tau_U})$$

$$V_i \sim N(0, \frac{1}{\tau_V})$$

The BYM model employs both spatial (U_i) and non-spatial (V_i) error components to address the extra variance not explained by the Poisson distribution. This approach helps to manage the unaccounted variance that extends beyond the data's spatial structure. Nonetheless, the process of fitting the BYM model can be complex as either the spatial component U_i or the non-spatial component V_i might excessively influence the variance explanation. The lack of hyperpriors for these components can lead to the exploration of a wide range of extreme posterior distributions, which may slow down or even hinder successful model fitting [97].

To implement the BYM model, Besag, York, and Mollié (1991) utilized gamma hyperpriors on the precision parameters τ_U and τ_V, choosing specific parameters for each. Subsequent adaptations have introduced constraints to ensure a balanced prior that equally weights spatial and non-spatial variance. This balance is based on the formula by Bernardinelli, Clayton, and Montomoli (1995) [10]:

$$\mathrm{sd}(V_i) = \frac{1}{\sqrt{\tau_U}} \approx \frac{1}{0.7\sqrt{\bar{m} \times \tau_U)}} \approx sd(U_i) \tag{3.24}$$

where $\bar{m}$ is the average number of neighbors for all regions. The choice of gamma hyperpriors for τ_V depends on $\bar{m}$, necessitating dataset-specific hyperprior evaluation for each new analysis.

BYM-2 Model

The BYM model is robust but has several issues. Firstly, the structured and unstructured components cannot be seen independently from each other, making it difficult to determine the true value of the model parameters under certain conditions. Secondly, choosing the hyperpriors is difficult as τ_U and τ_V do not represent variability on the same level, as τ_u is a joint hyperprior and τ_V is a conditional hyperprior. Lastly, it is not guaranteed that hyperpriors used in one application will have the same interpretation in another application, as with the ICAR and Poisson Log-Normal Models [71, 77]. To address these issues, Riebler et al.(2016) developed a reparameterization of the BYM model that resolves the identifiability issues [97]. The BYM-2 model ensures that the hyperparameters of the Gaussian random field are near orthogonal, facilitating clearer interpretation and intuitive prior assignment. Additionally, akin to the approach by Leroux, Lei, and Breslow (2000), it employs a unified precision (scale) parameter ϕ for the aggregate components and a mixing parameter ρ to balance spatial and non-spatial variations. For ϕ to accurately represent the standard deviation of the combined components, it's essential that for each i, $\text{Var}(U_i) \times \text{Var}(V_i) \approx 1$. To achieve this, a scaling factor s is introduced in the model, adjusting the variance proportion ρ. Since s varies depending on the dataset, it is treated as data within the model. Riebler et al. (2016) suggest normalizing the model so that the geometric mean of these variances equals 1. For implementation using R, this scaling factor is typically calculated using the INLA::inla.scale.model function in R and then incorporated into the stan model as data. This model is given by:

$$\theta_i = e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_3 X_{3i}+\rho_i} \tag{3.25}$$

$$\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \rho_i \tag{3.26}$$

$$\rho_i = \frac{1}{\tau_\rho}(V_i\sqrt{1-\phi} + U_i^*\sqrt{\frac{\phi}{s}}) \tag{3.27}$$

$$U_i \sim ICAR(\omega_{ij}, \frac{1}{\tau_U})$$

$$V_i \sim N(0, \frac{1}{\tau_V})$$

Here, $\phi \in [0, 1]$ is a mixing parameter, τ_ρ is the precision parameter, and U_i are the spatially structured random effects. These parameters address the identifiability issues and facilitate interpretation and transition priors [97].

3.2.5 Laplace Priors in Bayesian Disease Mapping

To address the limitations of traditional models in capturing abrupt spatial transitions in TB incidence, we introduce models with spatial Laplace priors. These models are designed to provide a more nuanced understanding of disease distribution, particularly in areas where disease incidence exhibits stark geographical contrasts. By leveraging Laplace regularization, these models aim to produce more localized estimates of disease risk, potentially leading to more effective identification of high-risk areas [123, 118].

Laplace priors, are characterized by their ability to induce sparsity in the parameter estimates. This is particularly useful in situations where it is believed that only a small subset of covariates contributes to the model, a common scenario in disease mapping when identifying key risk factors. The Laplace prior is defined as:

$$p(U) \propto \exp\left(\frac{1}{2}\sum_{i \geq j} |U_i - U_j|\right) \tag{3.28}$$

Here, the spatial random effects takes of the values of the absolute difference instead of the squared difference as the ICAR model. The Laplacian distribution has a sharper peak and heavier tails compared to a Gaussian distribution, leading to more parameters being driven to zero and some having larger absolute values, thereby promoting sparsity [13].

The use of Laplace priors in a Bayesian framework can be implemented through various computational methods, such as the Lasso or Bayesian Lasso, which can effectively handle high-dimensional data by selecting a subset of relevant covariates while shrinking the others towards zero.

3.2.5.1 Choosing Between Laplace and Gaussian Priors

In the Bayesian hierarchical modeling of disease risk, the selection between Laplace and Gaussian priors is pivotal and should be closely aligned with the specific characteristics of the data and the objectives of the study. Laplace priors are particularly beneficial when the goal is to pinpoint specific risk factors (via fixed effects) or areas (via random intercepts) that significantly contribute to the TB incidence, given their capability to identify localized patterns and sharp variations within the data. This feature is crucial in a metropolitan setting like Atlanta, where disease incidence may vary sharply between different neighborhoods or ZIP codes.

On the other hand, Gaussian priors are more appropriate for generalizing and smoothing disease incidence rates across a metro area. They are particularly useful when the data exhibits gradual trends or when mitigating overfitting is a concern, a common scenario in disease mapping studies. Gaussian priors help in obtaining more stable and continuous risk estimates across the geographic landscape, which may be crucial for understanding broader spatial trends in disease incidence.

3.2.6 Smoothing vs. Shrinkage in Disease Mapping

In the context of disease mapping, spatial smoothing techniques are commonly used to reduce local noise and highlight broader spatial patterns by enforcing spatial coherence. The fundamental assumption of smoothing methods is that disease rates should vary smoothly over space, implying that neighboring regions are likely to exhibit similar disease rates [77].

While smoothing can be valuable in many settings, it is not universally appropriate, particularly in cases where the assumption of spatial continuity does not hold. In some disease mapping scenarios, disease rates may exhibit sharp discontinuities or be concentrated in distinct hotspots rather than varying smoothly across space. For example, environmental barriers, social determinants, or behavioral factors might lead to stark differences in disease rates between neighboring regions. In such cases, smoothing can blur these sharp boundaries, leading to misleading conclusions by artificially merging distinct areas and obscuring critical epidemiological features [49].

Moreover, in situations where the disease is sparse—meaning that only a few regions exhibit significant deviations from the baseline rate—smoothing can dilute these significant signals by averaging them with neighboring areas where the disease is less prevalent or absent. This can result in an over-smoothed map that fails to capture the true extent and intensity of disease clusters, potentially undermining public health interventions that rely on accurate identification of high-risk areas.

3.2.6.1 The Horseshoe Model

To address the limitations of traditional methods to smooth or shrink to spatial neighbors, shrinkage techniques such as the horseshoe prior offer a compelling alternative, particularly in scenarios where disease rates exhibit sparsity or discontinuities. Unlike smoothing, which shrinks towards the values of spatial neighbors, the horseshoe prior employs a global-local shrinkage mechanism that allows for selective shrinkage of the data. This means that small or insignificant effects are shrunk towards zero, while larger, more significant effects are preserved, allowing the model to capture sharp contrasts and isolated hotspots that smoothing might obscure [24].

The horseshoe prior is a type of Bayesian hierarchical model that is particularly well-suited for handling sparse signals in high-dimensional settings, such as those encountered in disease mapping. The model is specified for spatial random effects U_i at location i as follows:

$$U_i \mid \lambda_i, \tau \sim N\left(0, \lambda_i^2 \tau^2\right)$$

where:

- U_i represents the spatial random effect at location i, capturing the deviation of the disease rate at that location from the global mean.
- λ_i is the local shrinkage parameter for location i, which controls the degree of shrinkage applied to U_i. The local shrinkage is governed by a half-Cauchy distribution: $\lambda_i \sim C^+(0,1)$. This distribution is heavy-tailed, allowing the horseshoe prior to retain large effects while shrinking smaller, less significant effects more aggressively towards zero.
- τ is the global shrinkage parameter, also modeled as $\tau \sim C^+(0,1)$. The global parameter τ influences the overall level of shrinkage across all locations, determining the general tendency of the model to shrink effects towards zero, thereby promoting sparsity [24, 39]

This hierarchical structure ensures that the horseshoe prior can adapt to different spatial patterns, applying minimal shrinkage in areas with strong signals (such as disease hotspots) while shrinking weaker signals towards zero to avoid overfitting.

3.2.7 Implementation in Disease Mapping

The conditional distribution of the spatial random effects U_i, given the neighboring effects and shrinkage parameters, can be expressed as:

$$p(U_i \mid U_{j \neq i}, \lambda_i) = N\left(\sum_{i \sim j} U_j d_{i,i}^{-1}, \frac{1}{d_{i,i}\lambda_i^2 \tau^2}\right)$$

Here, $d_{i,i}$ represents the diagonal elements of the spatial precision matrix D, which encodes the spatial relationships between different locations. This formulation integrates the spatial structure of the data directly into the shrinkage process, allowing the model to account for spatial dependencies while applying appropriate levels of shrinkage [120].

3.2.8 Comparison of Smoothing Techniques

The choice between smoothing (shrinking to neighbors values) and shrinkage to zero methods in disease mapping should be guided by the nature of the spatial variation in the disease rates:

- **Shrinking to neighbors** is most appropriate when the disease rates are expected to vary smoothly across space. Smoothing is effective in highlighting broad spatial trends and reducing noise, particularly in settings where spatial continuity is a reasonable assumption [77, 119].
- **Shrinkage to zero** is ideal when the disease rates are sparse, with only a few regions showing significant effects, or when the data exhibits sharp discontinuities or isolated spikes. The horseshoe prior's global-local shrinkage mechanism allows it to shrink small effects towards zero, preserving significant spatial heterogeneities without imposing unnecessary continuity [24, 16].

For example, in a scenario where tuberculosis (TB) prevalence is concentrated in specific neighborhoods due to localized factors (such as housing conditions, access to healthcare, or community transmission dynamics), the horseshoe prior might be more effective than smoothing in capturing these distinct clusters. Smoothing might inadvertently blend high-prevalence areas with surrounding lower-prevalence areas, leading to an underestimation of the true risk in the hotspots.

While spatial smoothing focuses on creating coherence by shrinking towards spatial neighbors, it may not be suitable in all disease mapping scenarios, particularly when dealing with sharp spatial contrasts or sparse data. The horseshoe prior offers a flexible and robust alternative, leveraging its global-local shrinkage mechanism to handle sparse and discontinuous data effectively. This makes it an invaluable tool for capturing complex spatial patterns in disease mapping, ensuring that significant effects are preserved while minimizing the risk of overfitting [120, 77, 24, 7].

3.2.8.1 Stage 3: Hyperpriors

In Bayesian disease mapping models, particularly those employing hierarchical structures, the selection of appropriate hyperpriors for spatial and non-spatial random effects is crucial. The choice of hyperpriors significantly influences the model's stability and the interpretability of its results. The following subsections detail typical hyperprior settings for both spatial and non-spatial random effects.

Non-Spatial Random Effects

For non-spatial random effects, which capture unstructured variability in the data, Gaussian distributions are commonly used as priors. A typical choice is a zero-mean normal distribution with a precision parameter (the inverse of variance). Specifically, a common hyperprior for the precision parameter of the non-spatial random effects is a Gamma distribution, often with small shape and rate parameters to ensure a non-informative or weakly informative prior. This approach allows the data to primarily inform the estimates.

Spatial Random Effects

For spatial random effects, which account for structured variability due to spatial relationships, Intrinsic Conditional Autoregressive (ICAR) models or Besag-York-Mollié (BYM) models are often utilized. The hyperprior for the precision parameter of the spatial random effect in ICAR or BYM models is typically a Gamma distribution, akin to the non-spatial case. These parameters control the degree of spatial smoothing the model is able to achieve. A parameter like $\text{Gamma}(5, 0.2)$ is relatively informative and indicates a lower level of smoothing, as it concentrates the prior distribution around smaller values. Conversely, a parameter like $\text{Gamma}(0.5, 0.01)$ represents a weakly informative prior, allowing for greater flexibility and resulting in a higher degree of smoothing. The choice of these parameters directly impacts the model's ability to balance local variation and global trends in the spatial data.

Hyperpriors in BYM2 Models

In the BYM2 model, a reparameterization of the BYM model, hyperpriors are set for the precision of the combined random effect and a mixing parameter (often denoted as ϕ). The precision parameter typically has a Gamma hyperprior. The mixing parameter, which controls the balance between structured and unstructured variation, is usually given a Beta distribution hyperprior, like $\text{Beta}(1, 1)$.

The choice of these hyperpriors should reflect the level of prior knowledge and the specific characteristics of the data. For instance, more informative priors can be employed if there is strong prior knowledge about the parameter values. It is crucial to conduct sensitivity analyses to assess the impact of different hyperprior choices on the model's results. These hyperprior settings serve as guidelines and can vary based on specific modeling needs and data characteristics. The goal is to provide sufficient flexibility for the model to learn from the data while incorporating any available prior knowledge.

3.3 Simulation Study

3.3.1 Generating Simulated Data

To compare the performance of the different disease mapping models, a simulation study was conducted. Simulated data was generated to establish a realistic ground truth for disease risk variation. We assume a total of 1016 cases in the study area with 125 ZIP codes and a population of 4,781,945 with an overall rate of 0.000165 cases per person. To generate the simulated data, we first calculated the expected number of cases E_i in each ZIP code by multiplying the overall rate by the population of each ZIP code. We then generated 1000 samples of observed case counts Y_i for each ZIP code from a Poisson distribution with the expectation μ_i as $E_i \times e^{\theta_i}$, where θ_i represents the underlying disease risk for ZIP code i. The log-linear predictor θ_i was specified as $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$, where X_{1i} represents the proportion of the population in ZIP code i that is foreign born, where X_{2i} represents the HIV prevalence in ZIP code i and where X_{3i} represents the Gini index for ZIP code i. The fixed effects were set to $\beta_0 = -0.521$, $\beta_1 = 2.64$, $\beta_2 = -.00399$, and $\beta_3 = 0.762$, to model the observed prevalence of tuberculosis under normal conditions in metro Atlanta. To illustrate, Figure 3.3a shows the geographical distribution of one realization of simulated case counts based on a total count of 1016 cases in Atlanta. Figure 3.3b depicts the expected case counts based on the total number of simulated cases.

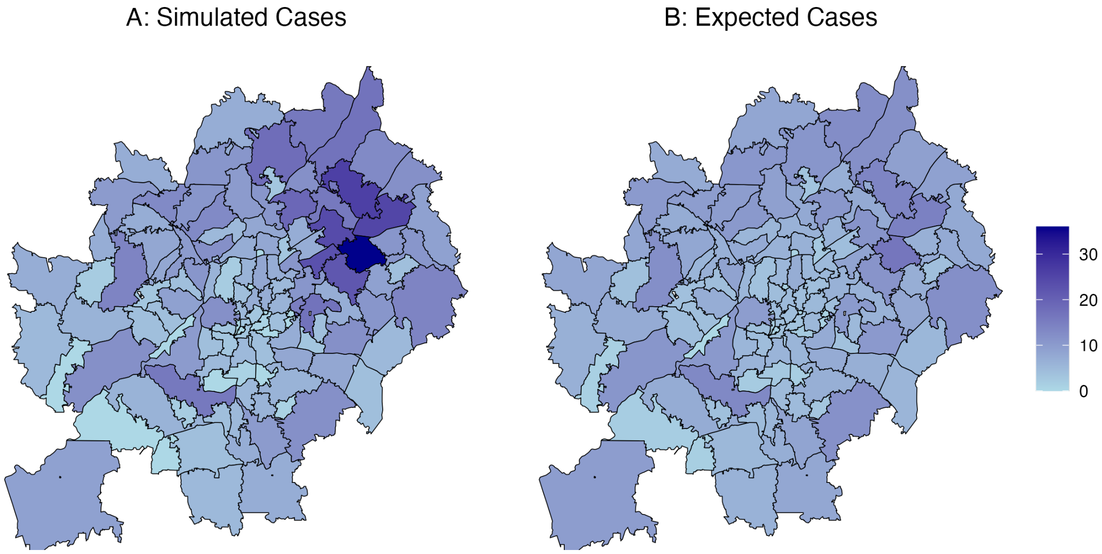


Figure 3.3: Simulated and expected case counts for 1016 total cases in Atlanta. (A) Displays the geographical distribution of a single realization of the simulated case counts. (B) Shows the corresponding expected case counts, offering a comparison between simulated variability and the underlying expected pattern across the region.

3.3.2 Hamiltonian Monte Carlo Settings

For a single replication, we use `cmdstanr` version 0.8.1 [46] and R software version 4.4.1 [96] to obtain 4 chains of length 30,000. Convergence is assessed using the Gelman-Rubin diagnostic ($\hat{R}$). The Gelman-Rubin diagnostic evaluates MCMC convergence by analyzing the difference between multiple Markov chains [51]. Specifically, convergence is assessed by comparing the estimated between-chains and within-chain variances for each model parameter, where large differences between these variances indicate nonconvergence. After discarding the first 29,000 iterations for each replication, we are left with a posterior sample of 1,000 SIR estimates per ZIP code.

To ensure the models run appropriately, several key Hamiltonian Monte Carlo (HMC) parameters are fine-tuned based on the complexity of the model. For the ICAR model, we set the `adapt_delta` to 0.8, the `max_treedepth` to 10, and use 5,000 warmup iterations. The BYM and BYM2 models require a `adapt_delta` of 0.9, a `max_treedepth` of 12, and 10,000 warmup iterations to ensure proper convergence and sampling. For the more complex Horseshoe model, we employ a `adapt_delta` of 0.95, a `max_treedepth` of 15, and extend the warmup period to 30,000 iterations.

The selected `adapt_delta` values help control the target acceptance rate of the HMC sampler, with higher values mitigating the risk of divergent transitions, particularly in more complex models. The `max_treedepth` parameter is adjusted according to model complexity to allow the sampler to explore the posterior distribution comprehensively. The number of warmup iterations is tailored to each model type to provide sufficient time for the sampler to adapt and stabilize before collecting samples for the posterior distribution.

These settings are critical for obtaining reliable and accurate posterior estimates while minimizing computational challenges such as divergences, non-convergence, or inefficient exploration of the posterior space. By applying these configurations, we ensure the acquisition of 1,000 high-quality samples per ZIP code, resulting in robust and representative posterior distributions across all model types.

3.3.3 Model Comparison

For model selection in disease mapping, we use two primary metrics: the Deviance Information Criterion (DIC) and the Watanabe-Akaike Information Criterion (WAIC).

DIC is widely used to evaluate hierarchical models by balancing model fit and complexity, defined as the sum of the expectation of the deviance and the effective number of parameters. It provides a trade-off between goodness-of-fit and model simplicity.

WAIC, on the other hand, is a fully Bayesian criterion that approximates cross-validation using the posterior distribution. It addresses some of the limitations of DIC, particularly regarding overfitting and the estimation of effective parameters.

A more detailed review of both criteria will be provided in Chapter 4.

An additional approach to model comparison in Bayesian hierarchical modeling is Leave-One-Out Cross-Validation (LOO-CV). LOO-CV is particularly relevant in hierarchical models because it assesses the model's predictive accuracy by systematically leaving out one observation at a time and predicting its value based on the remaining data. This method provides a robust estimate of the model's predictive performance, as it directly evaluates how well the model generalizes to unseen data.

In the context of Bayesian hierarchical modeling, LOO-CV is typically implemented using the Pareto Smoothed Importance Sampling (PSIS) method [114], which efficiently approximates the leave-one-out distribution without the need for extensive recomputation. The PSIS-LOO approach is advantageous because it scales well with complex models, where direct computation of LOO-CV would be computationally expensive.

The LOO-CV estimate of the predictive accuracy is given by the expected log pointwise predictive density (elpd), which sums the log-likelihoods of the observed data points given the posterior distribution, with each data point left out in turn. This metric is closely related to WAIC, but LOO-CV is often preferred in practice because it is less sensitive to the choice of prior and provides a more direct measure of out-of-sample predictive performance.

$$\text{elpd}_{\text{LOO-CV}} = \sum_{i=1}^{n} \log p(y_i \mid y_{-i}, \theta) \tag{3.29}$$

where $p(y_i \mid y_{-i}, \theta)$ represents the predictive distribution for y_i given all other data points y_{-i}. A model with a higher elpd is considered to have better predictive accuracy [115, 128].

In practice, both WAIC and LOO-CV are useful tools for model comparison, and they often yield similar results. However, LOO-CV is generally regarded as more robust, particularly in hierarchical models with complex structure and varying levels of spatial smoothing [83]. By directly evaluating the model's ability to predict new data points, LOO-CV provides an intuitive and reliable measure of model performance that complements the information provided by DIC and WAIC.

3.3.4 Simulation Results

3.3.5 Exploration of Random Effects Posterior Density

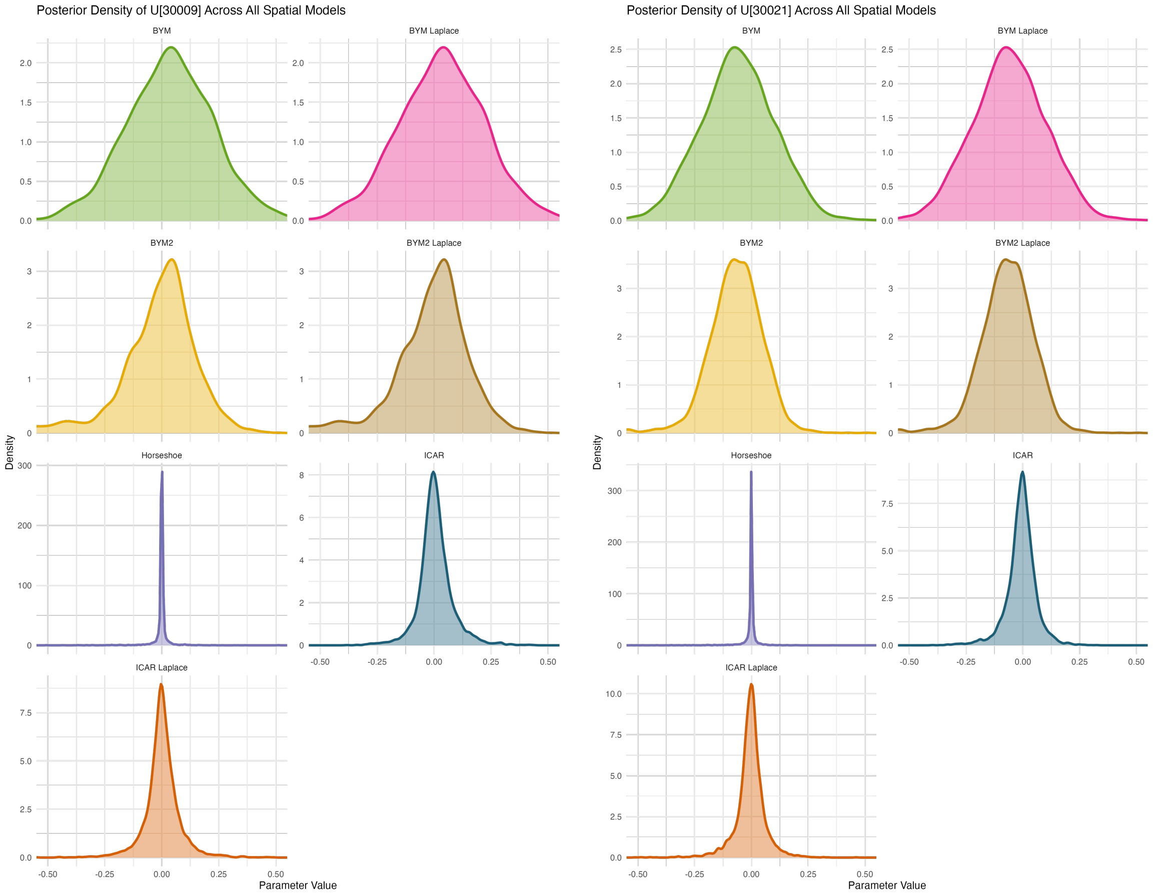
In addition to analyzing the spatial distribution of Standardized Incidence Ratios (SIR) estimates, we explored the posterior densities of the spatial random effects across different models. This exploration is crucial for understanding how each model captures the underlying spatial structure of TB incidence that is not explained by the covariates. The random effects are essential as they reveal spatial autocorrelation and local variability in disease risk.

Figure 3.4a presents the posterior density plots of the spatial random effects for ZIP codes 30009 (median number of TB cases) and 30021, an area identified with an extremely high number of TB cases compared to its neighboring ZIP codes. In the traditional spatial models (ICAR, BYM, BYM2), the density plots are relatively smooth and symmetric, resembling a normal distribution. This indicates that these models attribute the high incidence in ZIP code 30021 to both the fixed effects and the spatial random effects, with the random effects capturing additional local risk not explained by the covariates. The density plots for the L_1-regularized models (ICAR Laplace, BYM Laplace, BYM2 Laplace). L_1 regularization introduces sparsity by shrinking many of the smaller random effects towards zero, while allowing for larger deviations in areas with significant risk. In ZIP code 30021, the density plots remain relatively spread out, similar to the traditional models, indicating that the L_1-regularization does not overly shrink the random effect in this high-risk area. This suggests that the model effectively identifies and preserves significant spatial signals corresponding to TB hotspots.

Finally, we analyze the posterior density of the horseshoe model. This model employs a horseshoe prior, which is known for its ability to handle sparse signals by strongly shrinking small coefficients towards zero while allowing large coefficients to remain relatively unaffected. In ZIP codes with a median number of cases (such as 30009) the Horseshoe model produces posterior densities that are highly peaked around zero. This sharp peak reflects the model's strong shrinkage of small random effects towards zero in areas where the TB incidence does not significantly deviate from the overall mean or is adequately explained by the covariates. The horseshoe prior effectively filters out negligible spatial signals, leading to sparser and more interpretable random effects.

Comparing the posterior densities between ZIP code 30021 (high TB incidence) and ZIP code 30009 (low TB incidence) under the Horseshoe model highlights how the horseshoe prior prevents areas with high numbers of cases from pulling their neighbors' estimates upward. In ZIP code 30009, where the number of TB cases is close to the median and there is no substantial deviation from neighboring areas, the horseshoe prior strongly shrinks the random effects toward zero. This results in a posterior density that is sharply peaked, indicating minimal unexplained spatial variability. The model suggests that the observed cases in this area are adequately explained by the fixed effects and overall spatial trend.

In contrast, ZIP code 30021 exhibits a sharp increase in TB cases compared to its neighbors. The horseshoe prior allows the large random effect in this high-incidence area to escape shrinkage, maintaining its magnitude to capture the substantial local risk. The posterior density for this ZIP code is more spread out and resembles a normal distribution centered away from zero. Importantly, because the horseshoe prior heavily shrinks the random effects of neighboring areas toward zero, it prevents the high incidence in ZIP code 30021 from artificially elevating the random effects of adjacent ZIP codes like 30009. This property preserves distinct boundaries between high-risk and average-risk areas, ensuring that the model accurately reflects areas with true elevated risk without spreading this effect to neighboring regions.



(a) Posterior density plots of the random effect U for the ZIP code 30009 across all spatial models.

(b) Posterior density plots of the random effect U for ZIP code 30021 (Clarkston) across all spatial models.

3.3.6 Simulated Data Analysis

In this section, we evaluate the performance of seven spatial models—ICAR, ICAR Laplace, BYM, BYM Laplace, BYM2, BYM2 Laplace, and Horseshoe—using simulated case count data. Although these models are designed with different underlying assumptions and regularization techniques, the resulting Standardized Incidence Ratios (SIR) are surprisingly similar across all models. This consistency suggests that the spatial patterns in the simulated data are relatively stable and well-captured by each model, despite their methodological differences.

Figures 3.5 and 3.6 illustrate the geographical distribution of posterior SIR estimates across Metro Atlanta. While each model is intended to address spatial variability in different ways—whether through strong smoothing (as with traditional ICAR and BYM) or by introducing sparsity (as with the Laplace regularized models and the Horseshoe model)—the SIR estimates do not vary significantly between models.

One possible explanation for this similarity is that the underlying simulated data may not contain strong local variations or anomalies that would highlight the differences in model behavior. In such scenarios, where spatial signals are relatively uniform across the region, the benefits of advanced regularization techniques like Laplace regularization or global-local shrinkage might be less pronounced. As a result, all models converge to similar SIR estimates, focusing more on capturing the overall spatial trend rather than identifying distinct local deviations.

Despite expectations that models incorporating Laplace regularization or the Horseshoe model would reveal sharper or more localized differences, the SIR estimates remain consistent with those produced by traditional models. This consistency across models underscores the robustness of SIR estimates in the face of varying modeling approaches, especially when the data itself lacks strong local heterogeneity.

Table 3.1 provides a summary of the goodness-of-fit metrics and DIC scores. These metrics, while showing some variation, generally support the observation that the models perform similarly in terms of SIR estimation. Particularly of note in the table is the ICAR, ICAR Laplace, and horseshoe models perform best across all three indicators. The modest differences in DIC scores across models suggest that while the models employ different regularization techniques, the overall fit to the data remains comparable.

This analysis indicates that, in the context of the simulated data, the choice of spatial model may have a limited impact on the estimated SIRs, particularly when the data lacks significant local variability. However, this consistency among models also highlights their reliability in capturing broad spatial trends, setting the stage for their application to more complex real-world data in the subsequent section.

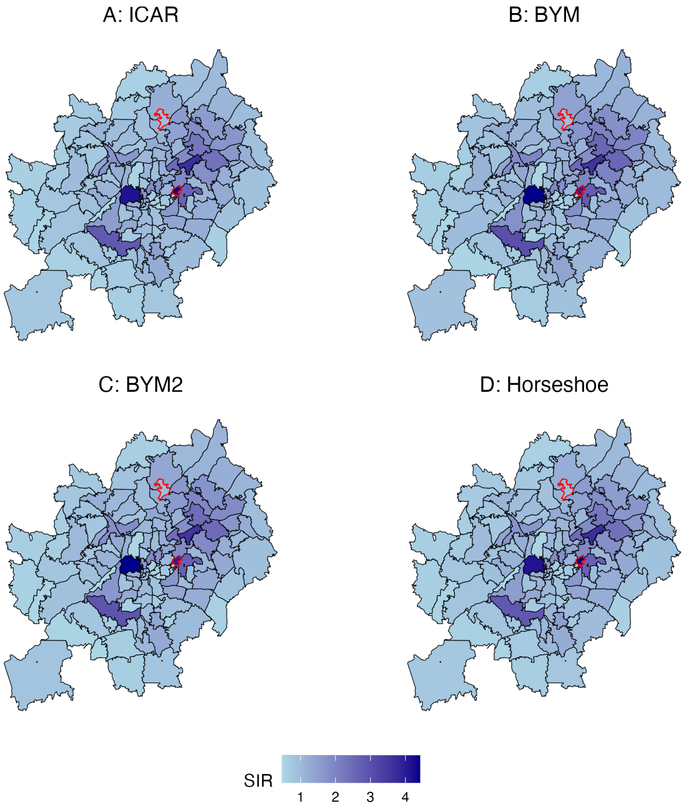


Figure 3.5: Geographical distribution of posterior SIR estimates for Metro Atlanta using the ICAR, BYM, BYM2, and Horseshoe models applied to simulated data.

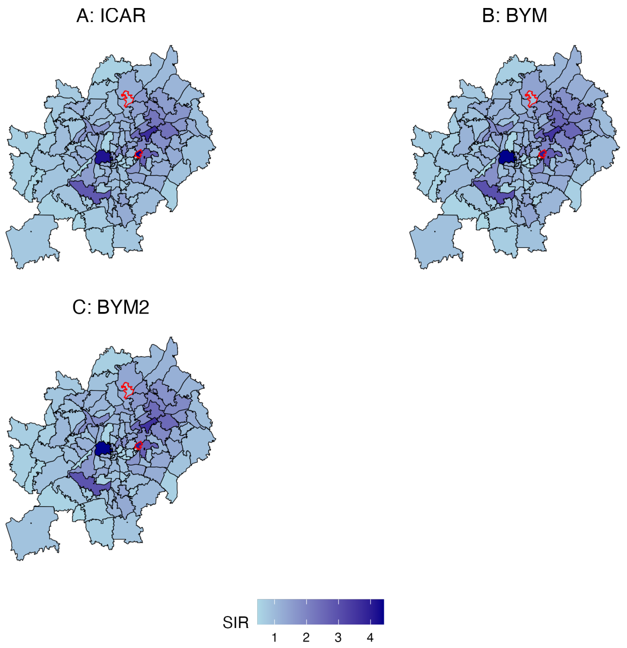


Figure 3.6: Geographical distribution of posterior SIR estimates for Metro Atlanta using the ICAR Laplace, BYM Laplace, and BYM2 Laplace models applied to simulated data.

Table 3.1: Goodness-of-fit metrics for the seven spatial models based on simulated data across Metro Atlanta.

Model	DIC	WAIC	LOO-CV
ICAR	471.492	472.970	473.056
$\text{ICAR}_{Laplace}$	471.636	472.866	472.927
BYM	501.246	524.893	531.363
$\text{BYM}_{Laplace}$	502.087	527.475	535.298
BYM2	485.761	497.389	499.330
$\text{BYM2}_{Laplace}$	487.430	500.888	503.390
Horseshoe	474.334	473.760	473.795

3.3.7 Sensitivity Analysis

This sensitivity analysis investigates the impact of increasing the Standardized Incidence Ratio (SIR) for ZIP code 30021, the most affected area in our study, on the model fit as measured by the Deviance Information Criterion (DIC) across seven spatial models: ICAR, ICAR Laplace, BYM, BYM Laplace, BYM2, BYM2 Laplace, and Horseshoe. The plot shown in Figure 3.7 illustrates the relationship between the increase in SIR and the corresponding changes in DIC for each model.

To assess the impact, we conducted 20 simulations where the case count in ZIP code 30021 was incrementally increased by 25% in each simulation, starting from the baseline case count. For each instance, 1,000 samples were generated, the SIR recalculated, and the DIC recomputed for each model. DIC serves as a metric that balances goodness-of-fit and model complexity, with lower DIC values indicating a better-fitting model. Tracking changes in DIC as the SIR increases provides insight into how well each model adapts to localized increases in disease incidence.

The plot in Figure 3.7 reveals distinct patterns in how DIC values responded to the incremental increases in SIR across the different models. For the ICAR and BYM models, which emphasize global spatial smoothing, the DIC values showed a steady increase as the SIR rose. These models initially maintained relatively stable DIC values, but as the SIR continued to rise, there was a noticeable upward trend in DIC. This suggests that while ICAR and BYM are robust in capturing overall spatial trends, they may struggle to adapt to significant local changes, leading to a less optimal fit as reflected in the rising DIC.

The ICAR Laplace and BYM Laplace models showed a more complex response. Initially, there was a sharp increase in DIC as the SIR began to rise, reflecting these models' sensitivity to the localized changes in data. However, as the SIR continued to increase, the DIC values stabilized, albeit at higher levels compared to the traditional ICAR and BYM models. This behavior suggests that while Laplace regularization enhances sensitivity to local variations, it also introduces additional complexity, which is penalized in the DIC calculation.

The **BYM2 Laplace** model exhibited a similar pattern to the other Laplace-regularized models, with an initial sharp increase in DIC followed by stabilization. The sharp rise indicates the model's quick response to the increasing SIR, but the subsequent leveling off suggests that the model may be adjusting to the increased local incidence rates, though at the cost of a higher overall DIC due to the complexity.

Interestingly, the Horseshoe model displayed a unique behavior compared to the other models. As the SIR for ZIP code 30021 increased, the DIC for the Horseshoe model initially rose but then decreased, indicating an improvement in model fit in certain instances as the SIR increased. This pattern suggests that the Horseshoe model, known for its ability to handle sparse signals and adapt to varying data, effectively balanced model complexity and fit, particularly when dealing with the significant localized increases in SIR. The Horseshoe model's flexibility allowed it to adjust to the changing data in a way that improved fit in some scenarios, resulting in a lower DIC compared to other models.

The sensitivity analysis shows that traditional models like ICAR and BYM are less responsive to localized increases in SIR, leading to a gradual increase in DIC as local incidence rates rise. In contrast, the Laplace-regularized models, while initially more sensitive to local changes, tend to stabilize at higher DIC values, reflecting their added complexity. However, the Horseshoe model stands out for its ability to improve fit in response to increased SIR in certain cases, as evidenced by the reduction in DIC after the initial increase. This suggests that the Horseshoe model may be particularly well-suited for scenarios involving significant localized variations in disease incidence.

This analysis highlights the importance of selecting the appropriate model based on the specific characteristics of the data. For regions with substantial localized variations, the Horseshoe model may offer a better balance between fit and complexity, as reflected by its ability to lower DIC in response to increasing SIR. In contrast, traditional models may provide more stable performance when broader spatial trends are of primary concern.

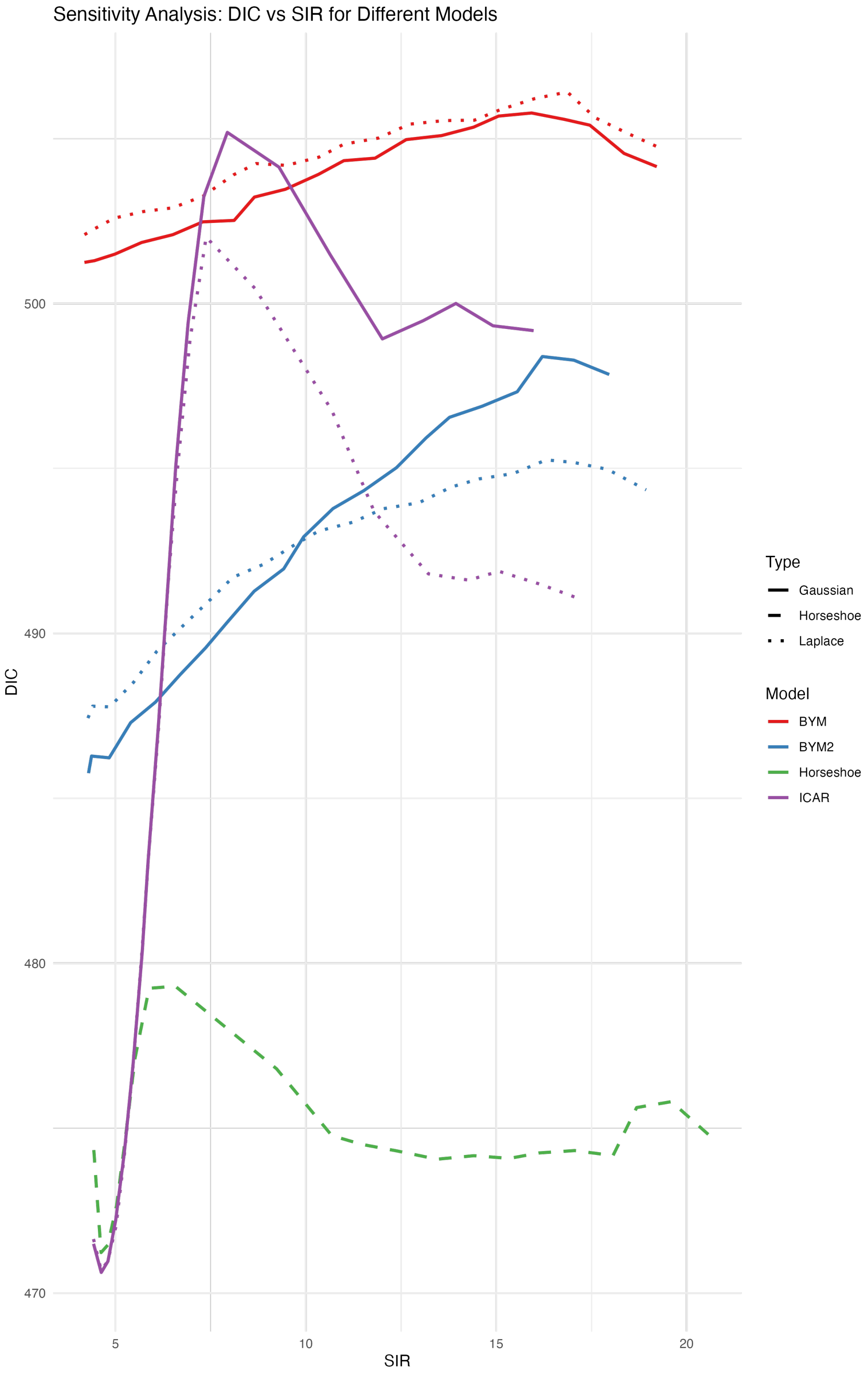


Figure 3.7: Change in DIC across the seven spatial models as the SIR for ZIP code 30021 increases. Each line represents the DIC trend as the case count is increased by 25% increments across 20 simulations.

This sensitivity analysis provides valuable insights into the responsiveness of different spatial models to localized increases in disease incidence, particularly highlighting the Horseshoe model's potential to improve fit in response to significant changes. These findings are critical for guiding model selection in disease mapping applications, depending on whether the focus is on broad spatial patterns or detailed local variations.

3.4 Case Study: Analysis of TB Incidence in Metro Atlanta Using Real Data

Building on the methodology established in the simulation study, we applied the same suite of spatial models—ICAR, ICAR Laplace, BYM, BYM Laplace, BYM2, BYM2 Laplace, and Horseshoe—to analyze real-world tuberculosis (TB) incidence data from Metro Atlanta. This section presents the results of applying these models to the real data, with a focus on identifying areas of elevated TB risk and evaluating the models' performance by examining the density of the spatial random effects.

3.4.1 Model Application and Results

Using the previously described data on TB cases and population estimates for Metro Atlanta ZIP codes, we calculated the Standardized Incidence Ratio (SIR) for each area. The same spatial models used in the simulation study were fitted to this data, incorporating the relevant socioeconomic and demographic covariates.

The results from the model fitting reveal how each model performs when applied to real data. Figures 3.8 and 3.9 show the spatial distribution of posterior SIR estimates for the traditional and L_1-regularized models, respectively, highlighting areas of elevated TB risk. Notably, the Horseshoe model continues to exhibit a strong ability to adapt to varying data, often yielding a lower DIC, especially in areas with higher TB incidence.

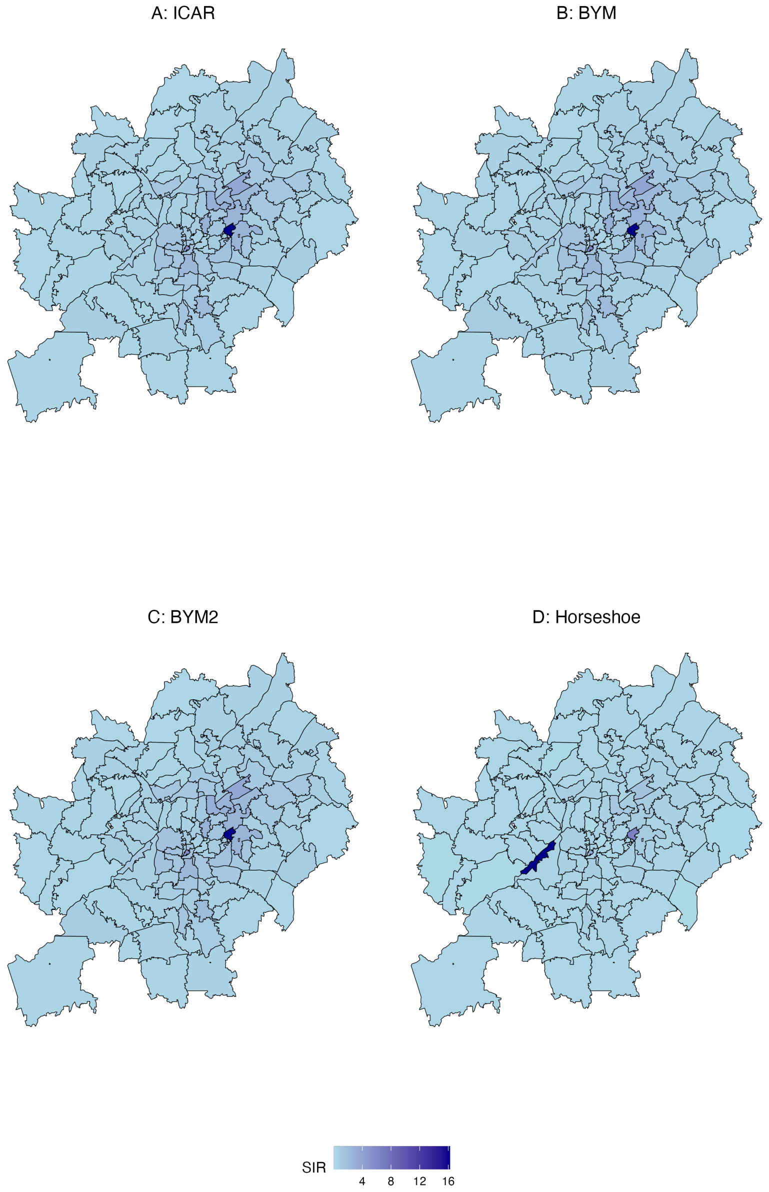


Figure 3.8: Spatial distribution of posterior SIR estimates for TB incidence in Metro Atlanta, comparing results from the ICAR, BYM, BYM2, and Horseshoe models.

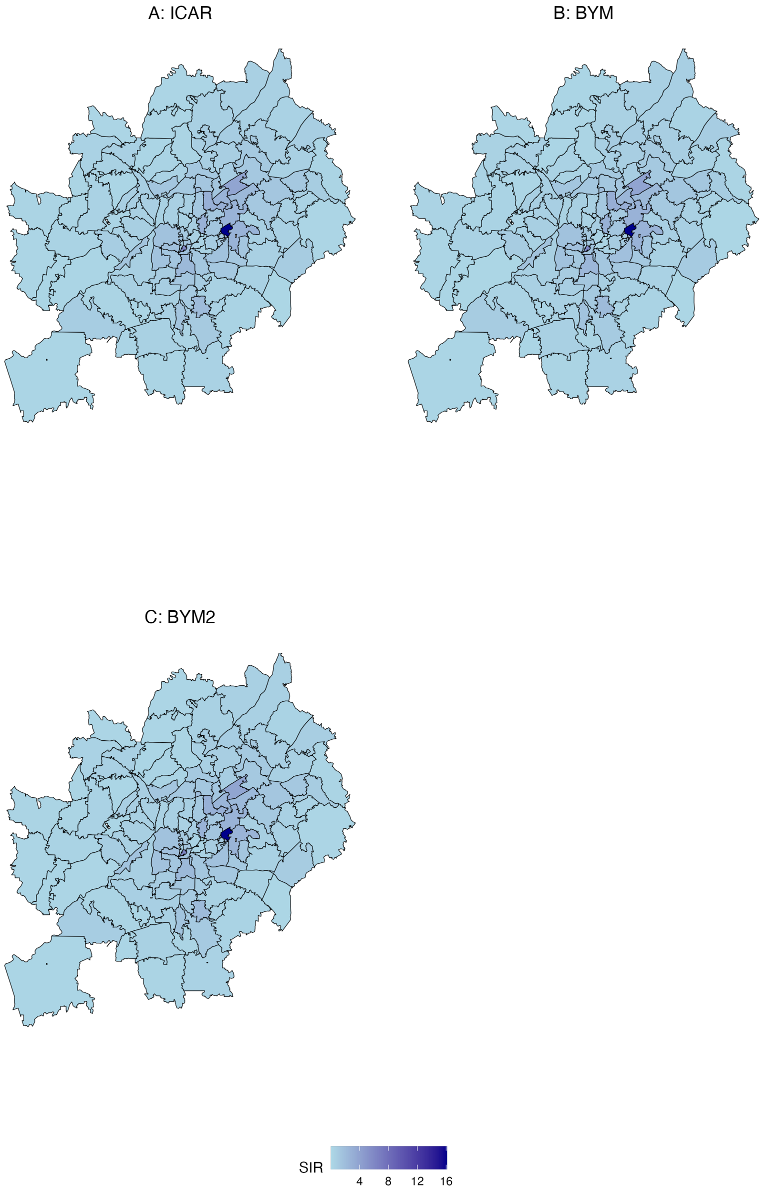


Figure 3.9: Spatial distribution of posterior SIR estimates for TB incidence in Metro Atlanta, comparing results from the ICAR Laplace, BYM Laplace, and BYM2 Laplace models.

Table 3.2 summarizes the goodness-of-fit metrics for each model. The DIC, WAIC, and Leave One Out Cross Validation (LOO-CV) values provide insights into the trade-offs between model complexity and fit. Lower values for these metrics indicate better model performance. Using the actual TB data from metro Atlanta yields a different model as the best fit for each metric.

Table 3.2: Goodness-of-Fit Metrics for Spatial Models Applied to TB Incidence in Metro Atlanta

Model	**DIC**	**WAIC**	**LOO-CV**
ICAR	545.311	604.488	626.551
ICAR Laplace	540.584	606.322	628.607
BYM	527.331	601.387	631.575
BYM Laplace	535.779	600.942	631.595
BYM2	542.119	603.223	629.334
BYM2 Laplace	537.422	602.763	627.127
Horseshoe	532.117	602.967	628.541

3.5 Discussion

The application of the seven spatial models—ICAR, ICAR Laplace, BYM, BYM Laplace, BYM2, BYM2 Laplace, and Horseshoe—to real TB incidence data from Metro Atlanta provided valuable insights into the performance and behavior of these models in a practical setting. Despite the varying methodologies and regularization techniques employed by each model, the Standardized Incidence Ratios (SIR) produced across the majority of ZIP codes were remarkably consistent. This consistency suggests that the spatial distribution of TB incidence in Metro Atlanta is relatively stable and effectively captured by all models, regardless of the specific approach used.

However, one notable exception emerged in ZIP code 30036, which has a very small population of 327 people and recorded 3 TB cases during the study period. In this ZIP code, the Horseshoe model produced a significantly higher SIR of 41, which sharply contrasts with the estimates from the other models. This discrepancy is likely due to the Horseshoe model's sensitivity to sparse signals and extreme values. The small population and relatively high number of cases in ZIP code 30036 created conditions where the Horseshoe model amplified the impact of this data point, leading to an unusually high SIR.

The divergence in the Standardized Incidence Ratio (SIR) for ZIP code 30036 highlights both the strengths and limitations of the Horseshoe model. The model's ability to detect and amplify localized anomalies is evident, as it maintains a high SIR for 30036 while the posterior SIRs in its neighboring ZIP codes remain lower than elsewhere. This indicates that the Horseshoe model does not automatically smooth excess risk into adjacent areas, preserving sharp contrasts in the data. This characteristic is beneficial for pinpointing potential hotspots, especially in regions with low population density where extreme values may signify a true area of concern. However, it also underscores the need for careful interpretation, as the model may overestimate risk in areas with small populations due to statistical fluctuations. Therefore, while the Horseshoe model is adept at highlighting significant local variations, analysts must consider the broader context to avoid misinterpretation of the results.

The overall consistency in SIR estimates across the models, apart from ZIP code 30036, indicates that for most areas in Metro Atlanta, the choice of spatial model may have a limited impact on the estimated TB risks. This uniformity suggests that the broad spatial patterns of TB incidence are robust and can be captured effectively by any of the traditional or L_1-regularized models. This finding is particularly relevant for public health applications where the primary concern is understanding and addressing broad spatial trends rather than pinpointing highly localized anomalies.

Despite this consistency, the behavior of the Horseshoe model in ZIP code 30036 serves as a reminder of the importance of model selection based on the specific characteristics of the data. In areas where the population is small and case counts are sparse, models like the Horseshoe, which are more sensitive to outliers and extreme values, may produce significantly different results from traditional models. This makes the Horseshoe model a powerful tool for identifying potential risk areas that might be overlooked by more conservative approaches, but also highlights the need for cautious interpretation, particularly in public health contexts where resource allocation decisions are made based on model outputs.

The slight differences in goodness-of-fit metrics, such as the Deviance Information Criterion (DIC), Watanabe-Akaike Information Criterion (WAIC), and Leave-One-Out Cross-Validation (LOO-CV), reflect the varying complexity and assumptions inherent in each model. While these differences are generally modest, they provide additional context for understanding how each model balances fit and complexity. The Horseshoe model, for example, showed a nuanced ability to balance these aspects, which may explain its distinctive performance in regions like ZIP code 30036.

In conclusion, the analysis suggests that while all the models examined provide reliable estimates of TB risk in most areas, the choice of model can become critical in regions with small populations or anomalous data. The Horseshoe model, with its sensitivity to local variations, may offer advantages in detecting potential hotspots but requires careful consideration to avoid overinterpretation of statistical outliers. This case study underscores the importance of tailoring model selection to the specific needs and characteristics of the data, particularly in public health applications where the stakes of misidentifying risk areas can be high. Future research could explore the application of these models to other diseases or geographic regions, further refining the understanding of how model choice impacts the identification and interpretation of spatial risk patterns.

Future work in this area could benefit from the incorporation of local measures of fit to better assess how well the models capture spatial variability across different regions. While global metrics like DIC, WAIC, and LOO-CV provide valuable insights into overall model performance, they can sometimes obscure regional differences in fit. By utilizing geographically weighted diagnostics or locally estimated posterior predictive checks, researchers could identify areas where the model performs particularly well or poorly. This would allow for a more nuanced understanding of spatial heterogeneity in model performance, potentially leading to more targeted model improvements and refinements.

Additionally, exploring spatially varying covariates could enhance the models' ability to capture localized effects and improve overall fit. Traditional models often assume that the relationship between covariates and the outcome is constant across space, which may not hold true in practice. Incorporating spatially varying coefficients would allow the model to account for differences in how covariates influence TB risk in different regions, providing a more accurate and context-sensitive analysis. This approach could be particularly valuable in public health applications, where the goal is to understand not only where high-risk areas are located but also why those areas are at increased risk.

By combining local measures of fit with spatially varying covariates, future models could offer more precise and actionable insights into the spatial dynamics of disease risk. This would not only improve the models' predictive power but also enhance their utility in informing public health interventions and policy decisions. As public health challenges become increasingly complex, the ability to accurately model and understand spatial risk patterns at a local level will be essential for effective disease prevention and control.

Chapter 4

A Localized Approach to Model Adequacy: Zooming in on Sharp Spatial Transitions in Disease Surveillance Models

4.1 Introduction

In the field of spatial statistics, assessing model adequacy is a critical concern, especially in the context of public health and disease surveillance. Public health models must be robust and reliable to ensure that the interventions they inform are both effective and efficient. Traditional global measures, such as the Deviance Information Criterion (DIC), have been extensively used to evaluate the overall fit of Bayesian hierarchical models. The DIC, as described by Spiegelhalter et al. (2002), balances the goodness of fit with model complexity by incorporating a penalty term for the effective number of parameters, making it a valuable tool in model selection and adequacy assessment [107]. However, while these global measures are essential, they often fail to capture the local variations in data, which are particularly crucial in spatial models dealing with geographically diverse regions.

This issue is particularly relevant in the context of tuberculosis (TB) in Metro Atlanta, where the spatial distribution of the disease is influenced by a complex interplay of socio-economic, demographic, and environmental factors. TB is a public health concern in this region, with prevalence rates varying significantly across different neighborhoods. These variations are driven by factors such as population density, access to healthcare, and the percentage of foreign-born residents, all of which contribute to the localized nature of TB outbreaks. Effective monitoring and intervention strategies require models that can accurately capture these spatial heterogeneities, making the assessment of local model adequacy crucial.

This chapter focuses on exploring the local pattern of TB in Clarkston, GA, is particularly significant due to the unique demographic and socio-economic characteristics of the area. Clarkston is known as one of the most ethnically diverse cities in the United States, often referred to as "the Ellis Island of the South." It has become a primary resettlement site for refugees from around the world, leading to a population with a high percentage of foreign-born residents. This demographic factor is critical when studying TB, as foreign-born individuals are often at higher risk for TB due to various factors, including limited access to healthcare, previous exposure in countries with higher TB prevalence, and the challenges associated with integrating into a new health system. Moreover, the socio-economic challenges faced by many residents, such as overcrowded living conditions and limited access to healthcare, further exacerbate the risk of TB transmission. Therefore, understanding the local pattern of TB in Clarkston not only provides insights into the effectiveness of current public health interventions but also highlights areas where targeted efforts are needed to address the unique vulnerabilities of this population. This focus on Clarkston underscores the importance of localized model adequacy in accurately capturing and responding to the health needs of diverse and at-risk communities[102].

Spatial models, such as those used in disease mapping, often rely on the assumption of spatial correlation among study units. This assumption, while simplifying the modeling process, can obscure local patterns that are critical for understanding disease dynamics in specific areas. For instance, Wheeler et al. (2010) emphasize that traditional global diagnostics might not adequately reflect the local model fit, potentially leading to erroneous conclusions about model adequacy [126]. This shortcoming is particularly concerning in the context of public health, where localized patterns of disease incidence, such as those observed with TB in Metro Atlanta, can vary significantly across regions due to differences in socio-economic factors, access to healthcare, and other local determinants.

To address this limitation, recent advancements have focused on partitioning the DIC into local components, enabling a more granular assessment of model adequacy. The concept of a localized DIC allows researchers to evaluate the model fit at the level of individual observations or specific geographic units, rather than relying solely on a global assessment. This approach, detailed in Wheeler et al. (2010), provides a more nuanced understanding of how well the model captures local variations in the data [126]. Additionally, by visualizing local DIC and comparing it across different model specifications, researchers can identify regions where the model may be underperforming or where additional covariates might improve the fit.

Moreover, the complexity of Bayesian hierarchical models, particularly those used in spatial analysis, requires careful consideration of model parsimony to avoid overfitting. Spiegelhalter et al. (2002) discuss the importance of balancing model fit with complexity, noting that an overly complex model may perform well on the training data but fail to generalize to new datasets [107]. This issue is particularly relevant in spatial models where the number of parameters can be large and difficult to estimate accurately. The partitioned DIC addresses this by incorporating a complexity penalty, ensuring that models remain interpretable and generalizable.

The relevance of these local diagnostic tools is further emphasized in the work by Jun et al. (2014), which provides a comprehensive Bayesian perspective on model complexity and the trade-offs involved in model selection. They discuss the importance of penalizing model complexity to prevent overfitting, particularly in hierarchical models where the number of parameters can be difficult to estimate. The authors argue that while global measures of fit, such as the DIC, are useful for overall model comparison, they may not fully capture the intricacies of local data structures, especially in complex models [67]. By decomposing the DIC into its component parts, including the penalty for model complexity and the fit of individual data points, researchers can gain deeper insights into the adequacy of their models. This partitioning allows for the identification of local inadequacies in model fit, which is crucial for accurately modeling spatial variations in disease prevalence, such as those observed in TB surveillance in Metro Atlanta.

In the context of TB in Metro Atlanta, where disease mapping models must account for sharp spatial transitions in disease prevalence, the use of localized diagnostics becomes even more critical. Disease mapping models, such as the Bayesian hierarchical models used in this study, are designed to capture the spatial distribution of TB while accounting for various covariates, including socio-economic factors and demographic variables. However, the accuracy of these models depends on their ability to adequately capture local variations in disease incidence, which can only be assessed through localized diagnostic tools.

In this chapter, we build on these foundations to explore a localized approach to model adequacy in the context of TB surveillance models in Metro Atlanta. By employing localized DIC and other diagnostic tools, we aim to identify and address sharp spatial transitions in TB incidence, ultimately leading to more accurate and reliable models. This approach is not only academically significant but also has practical implications for public health, particularly in regions like Metro Atlanta where TB burden is unevenly distributed. Through the use of localized diagnostics, public health officials can develop more targeted interventions tailored to the specific needs of different communities, thereby improving health outcomes at the population level.

The following sections will delve into the methods employed, present the results of a simulation study, and discuss the broader implications of this localized approach to model adequacy for future research and public health practice.

4.2 Methods

4.2.1 Overview of Global Measures of Model Fit

Global measures of model fit are essential tools in Bayesian modeling, providing a comprehensive assessment of a model's performance by balancing goodness of fit with model complexity. The Deviance Information Criterion (DIC) is one of the most widely used global measures, particularly in the context of complex hierarchical models. Introduced by Spiegelhalter et al. (2002), DIC has become a standard for comparing Bayesian models [107].

4.2.1.1 Deviance Information Criterion (DIC)

The DIC is defined as:

$$DIC = D(\bar{\theta}) + 2p_D \tag{4.1}$$

Where:

- $D(\bar{\theta})$ represents the deviance at the posterior mean of the model parameters.
- p_D is the effective number of parameters, providing a penalty for model complexity.

Deviance, in this context, measures how well the model's predictions match the observed data. The complexity term, p_D, penalizes models with more parameters, aiming to prevent overfitting. The DIC is particularly useful in model comparison, where it helps identify models that provide a balance between fit and complexity.

4.2.1.2 Limitations of DIC

While DIC is a valuable tool for assessing global model fit, it has several limitations, especially when applied to spatial models. One of the main drawbacks is its reliance on the posterior mean of the model parameters, which may not adequately capture the full uncertainty in complex models. Additionally, DIC assumes that the posterior distribution is approximately multivariate normal, which may not hold in cases of highly non-linear models or models with multimodal posterior distributions.

Moreover, DIC does not account for local variations in the data, which can be particularly problematic in spatial models where the disease distribution may vary significantly across different regions. This limitation necessitates the use of alternative measures that can provide a more nuanced understanding of model adequacy.

4.2.2 Watanabe–Akaike information criterion (WAIC)

To address some of the limitations of DIC, the Watanabe–Akaike Information Criterion (WAIC) has been proposed as a more robust alternative. WAIC, introduced by Watanabe (2010), is fully Bayesian and does not rely on approximations like the posterior mean or assumptions of normality [124]. WAIC is considered more suitable for complex models, particularly in the context of hierarchical and spatial models.

4.2.2.1 Calculation of WAIC

WAIC is calculated using the log-likelihood of the data given the model, averaged over the posterior distribution of the parameters. The WAIC for a model is defined as:

$$WAIC = -2\left(\sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i \mid \theta_s)\right) - V_{\text{WAIC}}\right) \tag{4.2}$$

Where:

- y_i represents the observed data for the i-th observation.
- θ_s represents the s-th sample from the posterior distribution of the parameters.
- $p(y_i \mid \theta_s)$ is the likelihood of y_i given the parameter sample θ_s.
- S is the total number of posterior samples.
- V_{WAIC} is the variance of the log-likelihood, defined as:

$$V_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_{\theta\sim p(\theta|y)}[\log p(y_i \mid \theta)] \tag{4.3}$$

WAIC provides a measure of model fit that is averaged over the entire posterior distribution, making it a more accurate reflection of the model's predictive performance. Additionally, WAIC accounts for the effective number of parameters, similar to DIC, but does so in a fully Bayesian manner[124].

4.2.2.2 Advantages of WAIC

WAIC has several advantages over DIC, particularly in the context of complex and hierarchical models. Since WAIC is fully Bayesian, it better captures the uncertainty in the model parameters by averaging over the entire posterior distribution rather than relying on point estimates. This makes WAIC more robust in cases where the posterior distribution is highly skewed, multimodal, or otherwise non-normal[124].

Another advantage of WAIC is its applicability to a wide range of models, including those with non-standard likelihoods or non-conjugate priors. WAIC is also asymptotically equivalent to leave-one-out cross-validation (LOO-CV), providing an interpretable measure of out-of-sample predictive accuracy [115].

4.2.2.3 Limitations of WAIC

Despite its advantages, WAIC is not without limitations. The calculation of WAIC involves computing the variance of the log-likelihood, which can be computationally intensive, especially in large datasets or models with many parameters. Additionally, while WAIC is generally more robust than DIC, it may still struggle in cases where the model is misspecified or where the data exhibits extreme heterogeneity.

4.2.3 Localized Measures of Fit: Local DIC and Local WAIC

Given the limitations of global measures like DIC and WAIC, there is a growing need for localized measures that can assess model adequacy at finer spatial resolutions. This is particularly important in spatial models used for disease mapping, where local variations in disease prevalence can significantly impact the effectiveness of public health interventions [126].

4.2.3.1 Local DIC for Administrative Units

Local DIC extends the concept of DIC to smaller administrative units, such as ZIP codes or census tracts, allowing for a more detailed assessment of model performance across different geographic areas. The Local DIC for a specific unit is computed as:

$$DIC_i = \bar{D}_i(\theta) + pD_i \tag{4.4}$$

Where:

- $\bar{D}_i(\theta)$ is the mean deviance for individual observation i, indicating how well the model predicts each specific data point.
- pD_i represents the complexity contribution from observation i, analogous to leverage in classical regression.

Local DIC provides a more granular understanding of model performance, particularly in areas with sharp spatial transitions. This can help identify regions where the model may require refinement or where additional covariates might improve the model's accuracy [126]. Notably, the sum of the local DICs across all observations equals the global DIC, ensuring that these localized assessments collectively reflect the overall model performance.

4.2.3.2 Local WAIC for Administrative Units

Local WAIC extends the concept of WAIC to individual observations or geographic units, providing a localized measure of model adequacy. Local WAIC is calculated similarly to global WAIC but focuses on the contribution of individual data points or small groups of data points.

The Local WAIC for the i-th observation or unit is defined as:

$$WAIC_i = -2\left(\log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i \mid \theta_s)\right) - \mathrm{Var}_{\theta\sim p(\theta|y)}[\log p(y_i \mid \theta)]\right) \tag{4.5}$$

Where the terms are defined as before, but focused on the specific observation or unit. Local WAIC, like Local DIC, allows for the identification of areas where the model may be underperforming, enabling more targeted model refinement and public health interventions. Importantly, the sum of the local WAIC values across all observations is equal to the global WAIC, ensuring that these localized assessments collectively reflect the overall model performance.

4.3 Case Study: Local Measures of fit for TB in Clarkston, GA

The objective of this simulation study is to evaluate the effectiveness of model adequacy tools in identifying and addressing local inadequacies within spatial models of tuberculosis (TB) Standardized Incidence Ratios. Specifically, the study assesses the performance of the Deviance Information Criterion (DIC) and the Watanabe–Akaike Information Criterion (WAIC) in distinguishing between models that vary in their levels of spatial smoothing and regularization.

As detailed in Chapter 3, this study employs seven disease mapping models to simulate TB incidence in Clarkston, GA and its immediate surroundings 38 ZIP codes. The models include the Intrinsic Conditional Autoregressive (ICAR), Besag-York-Mollié (BYM), and the extended Besag-York-Mollié (BYM2) models, each of which is analyzed in both its traditional form and with L1 Laplace regularization. Additionally, the horseshoe model is incorporated due to its capability to handle sparse signals and mitigate overfitting in regions with low data support. The primary aim is to compare these models' ability to detect local variations in disease incidence and to evaluate the trade-offs between global and local smoothing.

4.3.1 Approach

The study utilizes existing TB data from Clarkston, GA, known for its diverse population and significant TB burden[91]. The approach begins with TB case counts obtained between 2017 and 2021 obtained from the Georgia Department of Public Health (GDPH), allowing for a controlled comparison of the seven disease mapping models outlined in Chapter 3.

Each model is fitted to the GDPH data, and their performance is evaluated using both DIC and WAIC. These criteria are computed at local levels (ZIP codes). The goal is to identify regions where the models may underperform, thus requiring further refinement or the incorporation of additional covariates. The analysis focuses on the models' ability to capture overall spatial patterns and their sensitivity to local variations in TB incidence.

4.3.2 Identifying Areas of Concern

The analysis begins with the computation of global DIC and WAIC for each model, followed by the calculation of Local DIC and Local WAIC for ZIP codes surrounding Clarkston, GA. These localized measures are crucial for detecting regions where the models fail to adequately capture the underlying TB incidence patterns.

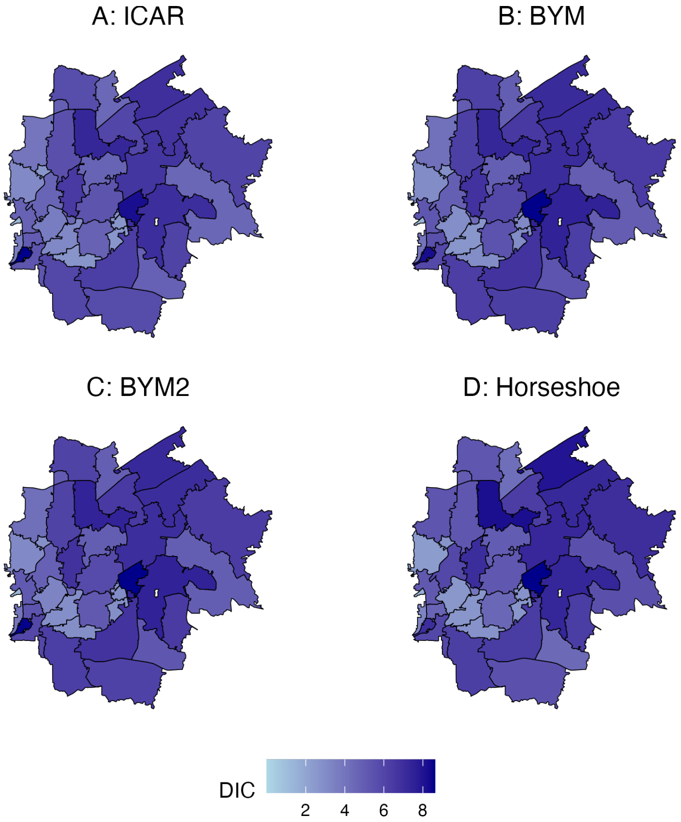


Figure 4.1: Geographical distribution of Deviance Information Criterion (DIC) Values around Clarkston, GA using the ICAR, BYM, BYM2, and Horseshoe models. Higher DIC values indicate regions with poor model fit.

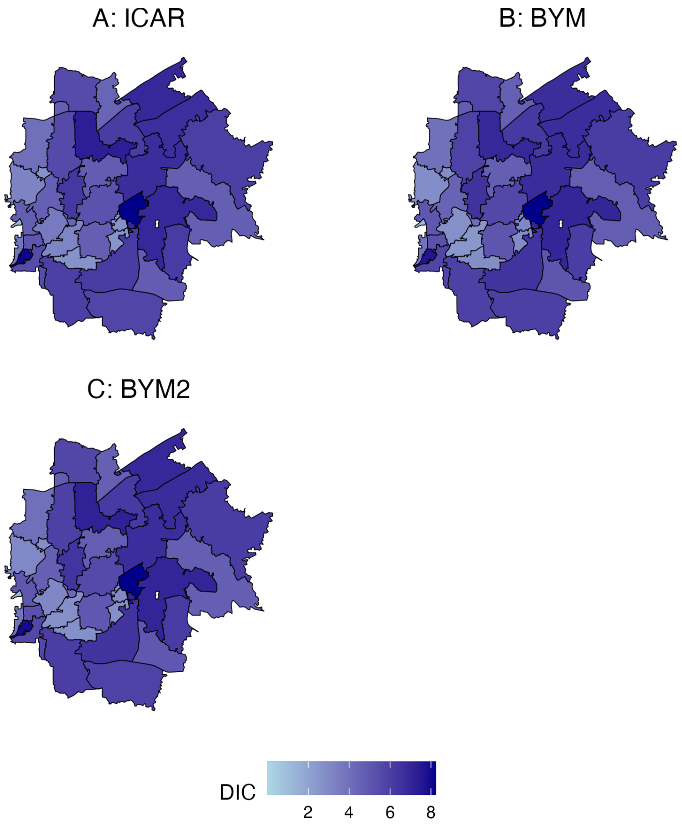


Figure 4.2: Geographical distribution of DIC Values around Clarkston, GA using the L1 Regularized Models. Higher DIC values indicate regions with poor model fit.

Figures 4.1 and 4.2 presents the spatial distribution of DIC values across Clarkston, highlighting areas with poorer model fit. These regions are identified as areas of concern, where the models may require additional refinement or the integration of further covariates to improve predictive accuracy.

Similarly, WAIC values are computed for the same regions to offer a comparison with DIC. Figures 4.3 and 4.4(models with L1 Laplace Regularization) presents the spatial distribution of DIC values across Clarkston, highlighting areas with poor model fit. Since WAIC is a fully Bayesian measure that accounts for the entire posterior distribution, it often provides a more robust evaluation of model fit, particularly in regions with complex spatial structures. The results indicate that both criteria are effective in detecting local inadequacies, with WAIC sometimes offering additional insights due to its comprehensive nature.

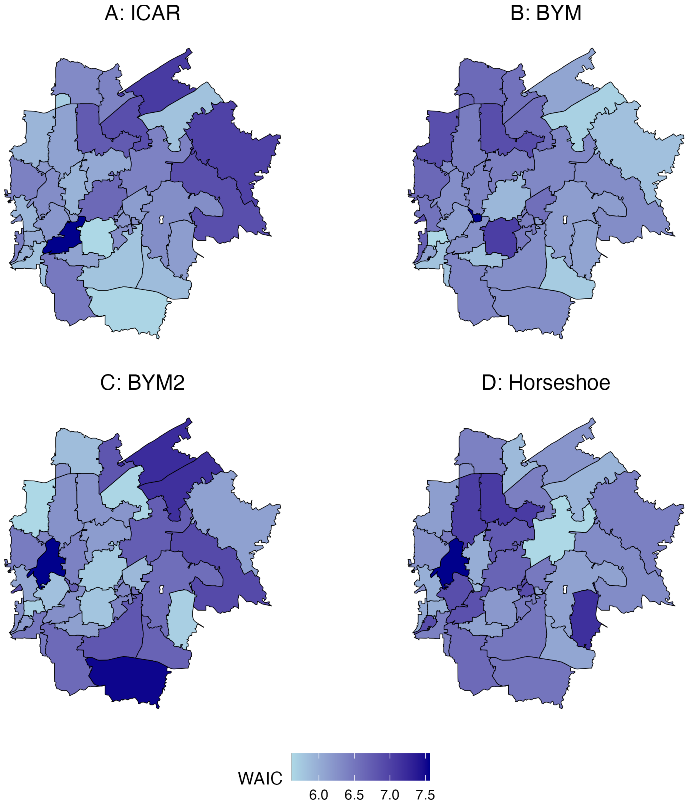


Figure 4.3: Geographical distribution of Watanabe–Akaike Information Criterion (WAIC)Values around Clarkston, GA using the ICAR, BYM, BYM2, and Horseshoe models. Higher WAIC values indicate ZIP codes with poor model fit.

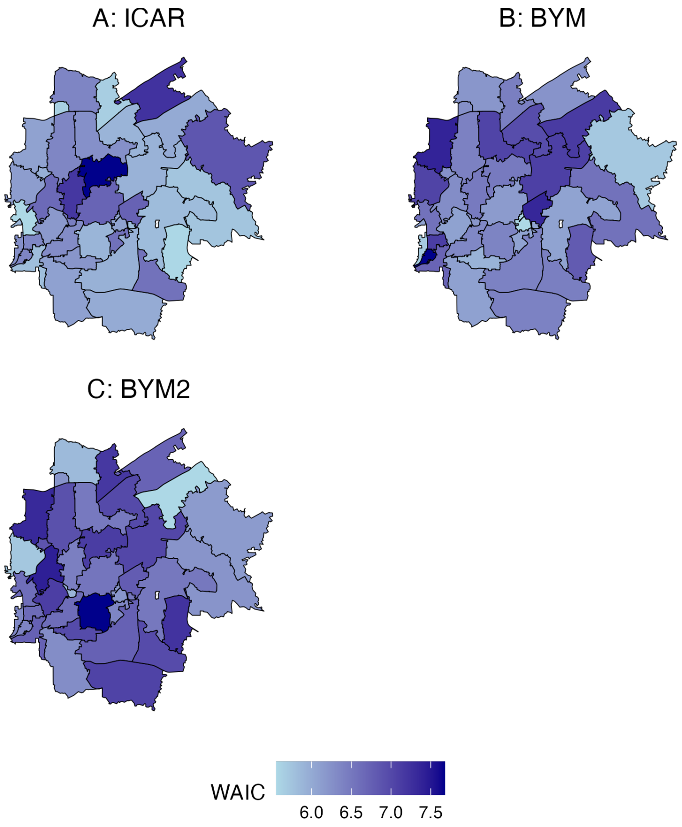


Figure 4.4: Geographical distribution of WAIC Values around Clarkston, GA using the L1 Regularized Models.

4.3.3 Leverage versus Model Adequacy Criteria

To further explore the relationship between model influence and adequacy, we examine the leverage of individual data points or spatial units against their corresponding DIC and WAIC values. Leverage, in this context, refers to the degree to which certain observations influence the overall model fit. High leverage points can skew model results, leading to potential misinterpretations if not properly accounted for.

Figure 4.5 illustrates the relationship between leverage and DIC. Regions with both high leverage and high DIC values are of particular concern, as they may indicate areas where the model is overly sensitive to specific local data points.

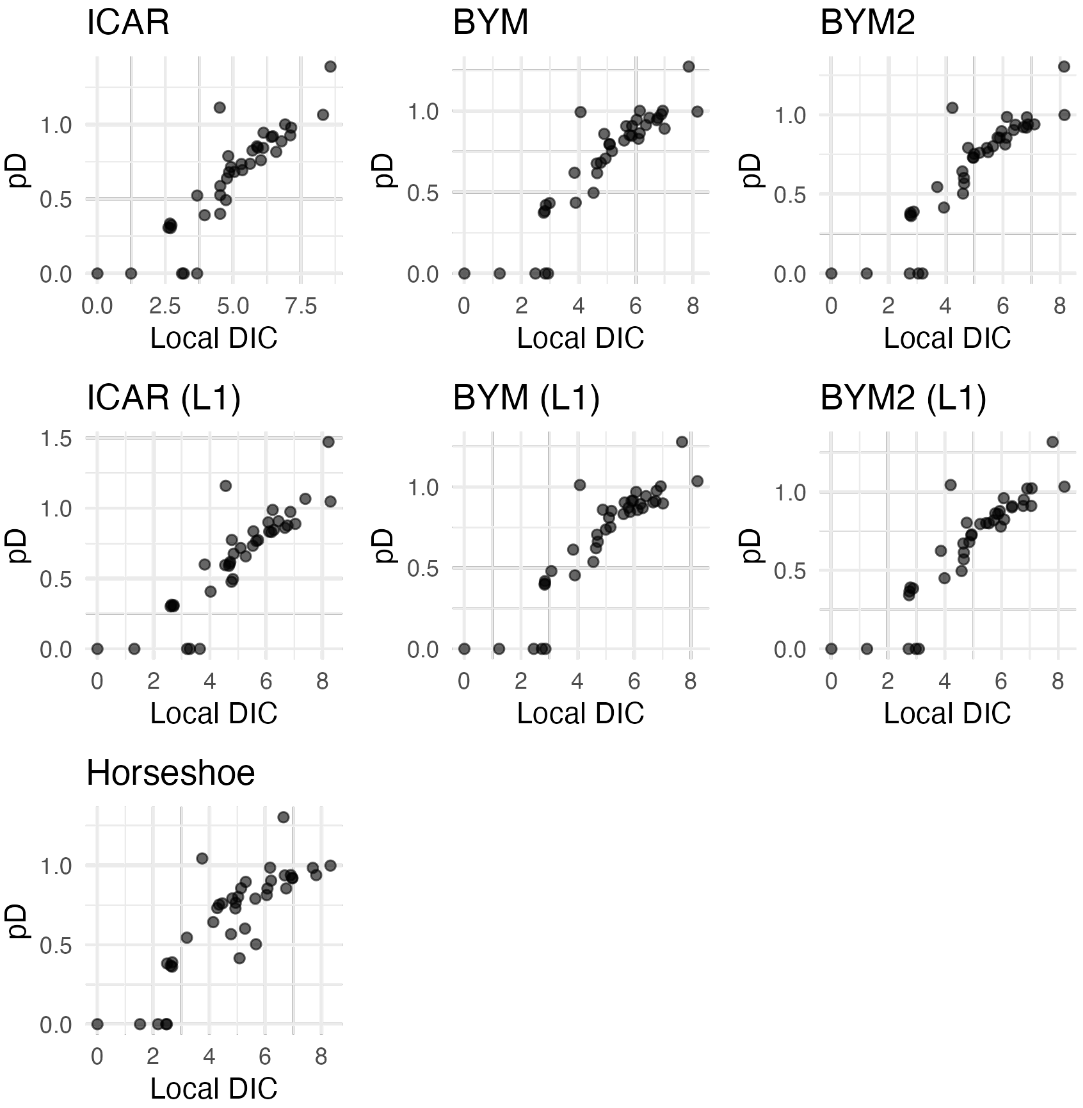


Figure 4.5: Leverage versus DIC: Analysis of the relationship between leverage and DIC in Clarkston, GA. Regions with high leverage and high DIC values are of particular concern.

4.4 Discussion

This study aimed to assess the local adequacy of various Bayesian hierarchical models in capturing the spatial distribution of tuberculosis (TB) in Metro Atlanta, with a particular emphasis on Clarkston, GA. Clarkston's unique demographic characteristics—especially its high percentage of foreign-born residents and socio-economic diversity—make it a critical area for TB surveillance and intervention. The results underscore the importance of selecting appropriate models that can account for the socio-economic and demographic heterogeneity in accurately reflecting the complexities of TB prevalence in this region.

4.4.1 Model Adequacy and Local Patterns of TB

The evaluation of model adequacy through localized diagnostic tools such as the Deviance Information Criterion (DIC) and the Watanabe-Akaike Information Criterion (WAIC) provided nuanced insights into the performance of different models across Metro Atlanta. Clarkston, with its significant percentage of foreign-born residents (over 50% in some areas), exemplifies the necessity of models that accurately account for socio-economic and demographic disparities.

The DIC and WAIC values for the Horseshoe, ICAR, BYM, and their respective L1-regularized versions were calculated and compared across different geographic units. The results indicate substantial variation in model performance, reflecting the spatial heterogeneity of TB incidence in the region. Notably, the ICAR and BYM models, particularly in their L1-regularized forms, demonstrated superior fit and predictive accuracy in areas like Clarkston, where we observe sharp transitions between this particular ZIP code and its neighbors. The lower WAIC values for these models suggest that they are more effective in capturing the local nuances of TB spread, likely due to their ability to flexibly incorporate spatial correlations.

4.4.2 Implications for TB Surveillance in Clarkston

The findings from this study have significant implications for public health practice, particularly in regions like Clarkston, where the TB burden is disproportionately high. The high percentage of foreign-born residents, many of whom may have been exposed to TB in countries with higher prevalence, presents a unique challenge for TB control efforts. Traditional global models may not adequately capture these localized risks, potentially leading to underestimation of TB incidence in such communities.

The superior performance of localized diagnostics in this study suggests that more granular models, which consider the specific socio-economic and demographic context of areas like Clarkston, are essential for accurate TB surveillance. Public health officials can leverage these findings to develop more targeted interventions that address the unique needs of high-risk communities. For instance, increasing access to TB screening and treatment in refugee communities, enhancing public health education, and tailoring interventions to the socio-economic conditions in Clarkston could significantly reduce the TB burden in this area.

4.4.3 Future Work: Spatially Varying Coefficient Models

While this study has provided valuable insights into the spatial dynamics of TB in Metro Atlanta, particularly in Clarkston, there is a need for further research that explores the use of spatially varying coefficient (SVC) models. These models offer the potential to capture the complex interactions between covariates and spatial location, allowing for the effects of socio-economic and demographic factors on TB risk to vary across space. This would enable a more refined understanding of how different factors contribute to TB prevalence in specific areas, providing an even more localized assessment of disease risk.

SVC models could improve upon the limitations observed in the current models by allowing coefficients to vary spatially, thereby accommodating the heterogeneous impact of variables like foreign-born population percentage and income inequality across different regions. This approach would be particularly valuable in areas with high socio-economic diversity, such as Clarkston, where the relationships between risk factors and TB incidence may not be uniform across space.

Moreover, the integration of SVC models with dynamic data sources, such as real-time health surveillance data, could further enhance the accuracy and timeliness of TB risk predictions. Future work should focus on the development and application of SVC models in TB surveillance, with the goal of providing public health officials with more precise tools for targeting interventions and resources in the most vulnerable communities.

In conclusion, this study highlights the critical importance of using localized diagnostics and advanced spatial models to accurately capture the spatial heterogeneity of TB in areas like Clarkston, GA. By advancing research into spatially varying coefficient models, future work can contribute to the development of more effective public health strategies tailored to the specific needs of diverse and high-risk populations. This approach has the potential to significantly improve TB surveillance and control efforts, ultimately reducing the burden of this disease in vulnerable communities.

Chapter 5

Novel Insights in Spatial Epidemiology: Comparing Bayesian Spatially Varying Coefficient Models

5.1 Introduction

In spatial epidemiology, accurately capturing the distribution of infectious diseases such as tuberculosis (TB) is essential for effective public health interventions, particularly in heterogeneous urban settings. Traditional Bayesian hierarchical frameworks, have been valuable in identifying factors influencing disease distribution. However, these models often assume stationarity, meaning that they treat the relationship between covariates and disease outcomes as constant across space. This assumption ignores local variations that are essential for precise disease mapping and effective intervention strategies. These models, including the widely used intrinsic conditional autoregressive (ICAR) models [13], have provided valuable insights into disease risk patterns. However, ICAR models, despite their ability to account for spatial correlation, suffer from several limitations when applied to regions with sharp socioeconomic transitions between small areas, such as Metro Atlanta.

The most significant challenge with ICAR models is their tendency to over-smooth across areas with local heterogeneity. By applying global smoothing parameters, ICAR models often fail to capture abrupt changes in disease risk between neighboring regions, leading to oversimplified and sometimes misleading results. This lack of adaptability to local discontinuities, coupled with sensitivity to outliers even when using heavy-tailed priors, calls for more flexible models that can better accommodate complex spatial patterns.

To address these limitations, Bayesian Spatially Varying Coefficient Models (BSVCMs) have emerged as a powerful alternative. BSVCMs introduce location-specific effects, allowing for more local adaptability in the relationships between covariates and disease outcomes. By allowing coefficients to vary spatially, BSVCMs are able to capture regional differences in covariate effects that traditional ICAR models smooth over, making them better suited for heterogeneous environments like Metro Atlanta. These models like one proposed by [121] offer significant advantages in reducing oversmoothing, providing a more direct assessment of spatial risk factors, and offering greater flexibility in capturing the local variability of disease drivers.

However, despite their advantages, BSVCMs present their own challenges. Estimating a large number of location-specific coefficients increases computational complexity, especially when dealing with high-dimensional data. Additionally, the flexibility of these models introduces a risk of overfitting, which can diminish predictive accuracy when applied to new data. The complexity of these models also complicates interpretation, making it more difficult to communicate results to stakeholders.

To address these challenges, this chapter introduces a novel approach—*Locally Adaptive Smoothing* (LAS). LAS modifies spatial weights dynamically based on differences in expected disease counts between neighboring regions, enhancing the model's ability to detect local discontinuities and prevent over-smoothing. By incorporating LAS, our proposed Bayesian Spatially Varying Coefficient Model (BSVCM) retains the flexibility needed to capture complex spatial patterns while addressing key limitations related to over-smoothing and adaptability.

In this chapter, we will outline the theoretical foundations of Bayesian Spatially Varying Coefficient Models and detail the enhancements introduced through LAS. A case study of TB incidence in Metro Atlanta will demonstrate the practical benefits of the proposed model, focusing on how it improves upon traditional methods by capturing sharp transitions in TB risk across neighborhoods. Special attention will be given to areas such as Clarkston, GA, which have unique demographic profiles that present challenges for traditional spatial models.

The proposed method offers a significant contribution to spatial epidemiology, particularly in urban environments with high levels of socioeconomic diversity. By providing a more nuanced understanding of the spatial distribution of TB risk, the Locally Adaptive Smoothing-enhanced BSVCM enables public health professionals to design more targeted and effective intervention strategies.

5.2 Methods

5.2.1 Bayesian Hierarchical Disease Mapping: Overview

Spatial epidemiology plays a crucial role in understanding and mitigating the spread of infectious diseases. Accurate disease mapping enables public health officials to identify high-risk areas, allocate resources efficiently, and implement targeted interventions. Bayesian hierarchical models have emerged as a powerful tool for disease mapping, offering robust frameworks to estimate disease risk across small geographical areas. These models account for spatial correlation by leveraging information from neighboring regions, a process often referred to as “borrowing strength.” Among these, the Intrinsic Conditional Autoregressive (ICAR) model is widely utilized due to its ability to model spatial dependence effectively.

5.2.2 Extensions to the ICAR Model

To enhance the flexibility and robustness of Bayesian hierarchical models in spatial disease mapping, various extensions to the traditional Intrinsic Conditional Autoregressive (ICAR) model have been developed. These extensions incorporate different prior distributions, each addressing specific limitations inherent in the ICAR framework. The Laplace prior [13, 14], for example, introduces sparsity in spatial differences, enabling the model to capture sharp spatial transitions between neighboring areas more effectively. This is particularly advantageous in regions where abrupt changes in disease risk are expected, ensuring that the model does not overly smooth these critical boundaries.

Similarly, the Student-t prior [111] accommodates outliers by employing heavy-tailed distributions, thereby enhancing the model’s robustness to extreme values in spatial risk estimates. This feature is essential in scenarios where certain regions may exhibit unusually high or low disease counts that deviate significantly from the surrounding areas. Additionally, the Horseshoe prior [24] promotes sparsity by shrinking most spatial effects towards zero while allowing significant effects to remain unaltered. This selective shrinkage focuses the model on areas with meaningful spatial variation, effectively highlighting regions of interest without diluting the influence of substantial spatial effects.

These extensions collectively improve the ICAR model's ability to accurately represent complex spatial patterns by addressing issues related to over-smoothing, outlier sensitivity, and the identification of significant spatial effects. By incorporating these advanced priors, Bayesian hierarchical models become more adept at capturing the nuanced spatial dynamics that characterize heterogeneous environments, thereby providing more reliable and interpretable disease risk maps.

5.2.3 Limitations of Traditional ICAR Models and Their Extensions

Despite the advancements introduced by these extensions, traditional ICAR models continue to exhibit limitations, particularly when applied to regions with complex spatial dynamics such as Metro Atlanta. One of the foremost challenges is the tendency of ICAR models to over-smooth in heterogeneous regions, which results in the failure to capture abrupt changes in disease risk between neighboring areas. This over-smoothing undermines the model's ability to identify and represent critical local variations, leading to oversimplified disease maps that do not accurately reflect underlying spatial patterns.

Furthermore, the use of global smoothing parameters in ICAR models impedes their adaptability to local discontinuities. These parameters are applied uniformly across all regions, disregarding region-specific variations that may be crucial for accurately modeling disease dynamics. Consequently, ICAR models are less sensitive to local changes in disease risk, reducing their effectiveness in diverse and heterogeneous urban settings where socio-economic factors can vary markedly over short distances.

Additionally, ICAR models remain sensitive to outliers and extreme values, even when incorporating heavy-tailed priors such as the Student-t distribution. This sensitivity can lead to poor estimation of spatial risk, particularly in areas where disease counts significantly deviate from the norm. The presence of outliers can disproportionately influence spatial risk estimates, compromising the model's reliability and interpretability.

These limitations underscore the necessity for enhanced models that offer greater local adaptability and flexibility, enabling more accurate capture of complex spatial patterns. Bayesian Spatially Varying Coefficient Models (BSVCMs) represent the next step in addressing these challenges. By allowing for location-specific effects, BSVCMs improve the model's ability to reflect true spatial heterogeneity, thereby providing a more nuanced and accurate representation of disease risk across diverse and dynamic urban landscapes.

5.2.4 Bayesian Spatially Varying Coefficient Models

To overcome the limitations of traditional ICAR models, Bayesian Spatially Varying Coefficient Models (BSVCMs) have been developed [125, 48, 121]. BSVCMs introduce location-specific effects, thereby enhancing the model's ability to adapt to local spatial variations and capturing complex spatial dependencies more accurately.

5.2.4.1 Model Structure

The BSVCM extends the traditional Bayesian hierarchical framework by allowing the coefficients of covariates to vary spatially. The general form of the model is expressed as:

$$\log(\theta_i) = \beta_0 + \sum_{k=1}^{K} \beta_{k,i} X_{k,i} + U_i \tag{5.1}$$

where:

- θ_i represents the log-relative risk of disease at location i.
- β_0 is the global intercept term.
- $\beta_{k,i}$ denotes the spatially varying coefficient for covariate k at location i.
- $X_{k,i}$ is the value of covariate k at location i.
- U_i captures the spatially structured random effect, accounting for unobserved heterogeneity.

Each spatially varying coefficient β_{ki} is modeled as:

$$\beta_{k,i} = \beta_k + s_{k,i} \tag{5.2}$$

where β_k is the global effect of covariate k, and $s_{k,i}$ is the spatially varying component modeled using a multivariate ICAR prior.

5.2.4.2 Key Advantages

Bayesian Spatially Varying Coefficient Models (BSVCMs) offer several key advantages over traditional spatial models, primarily due to their inherent flexibility and adaptability to local spatial variations. A significant benefit of BSVCMs is their ability to provide local adaptability. By allowing the coefficients of covariates to adjust for each specific location, BSVCMs effectively reduce over-smoothing—a common issue in traditional models that tend to average effects across regions. This localized adjustment ensures that the unique characteristics of each area are accurately captured, leading to more precise and meaningful estimates of disease risk.

Moreover, BSVCMs exhibit a high degree of flexibility in modeling regional differences in covariate effects. This flexibility is crucial in heterogeneous environments where the influence of socioeconomic and demographic factors on disease incidence can vary markedly from one neighborhood to another. By accommodating these regional disparities, BSVCMs provide a more nuanced understanding of the underlying determinants of disease distribution, facilitating the identification of area-specific risk factors.

Additionally, BSVCMs enable direct assessment of spatial risk factors by explicitly modeling the spatially varying relationships between covariates and disease outcomes. This capability allows researchers to evaluate how the impact of each covariate differs across locations, offering deeper insights into the spatial dynamics of disease transmission. Such detailed assessments are invaluable for informing targeted public health interventions and policy decisions, as they highlight the specific factors driving disease risk in different regions.

5.2.5 Challenges Associated with Bayesian Spatially Varying Coefficient Models

Despite their enhanced flexibility and numerous advantages, Bayesian Spatially Varying Coefficient Models (BSVCMs) present several challenges that must be addressed to ensure their effective application. One primary challenge is computational complexity. Estimating a large number of location-specific coefficients significantly increases the computational burden, especially when dealing with extensive datasets or a multitude of covariates. This heightened computational demand can limit the scalability of BSVCMs and may require the use of advanced computational techniques or high-performance computing resources to achieve feasible run times.

Another notable challenge is the risk of overfitting. The high degree of flexibility inherent in BSVCMs allows the model to fit the training data very closely, potentially capturing noise rather than the underlying signal. This overfitting can lead to reduced predictive accuracy when the model is applied to new, unseen data, undermining its utility for generalization and practical application in public health surveillance.

Moreover, BSVCMs often suffer from interpretability issues. The introduction of numerous spatially varying coefficients can complicate the model structure, making the results more difficult to interpret and communicate effectively to stakeholders. This complexity can pose challenges in translating statistical findings into actionable public health strategies, as stakeholders may find it challenging to grasp the nuanced spatial relationships modeled by BSVCMs.

These challenges highlight the necessity for models that strike a balance between flexibility and interpretability. It is imperative to develop methodologies that maintain the ability to capture complex spatial patterns while ensuring that the results remain comprehensible and actionable for decision-makers. Addressing these challenges is essential for the broader adoption and practical implementation of BSVCMs in spatial epidemiology and public health research.

5.2.6 Proposed Method: Locally Adaptive Smoothing (LAS)

To address the limitations inherent in both traditional ICAR models and BSVCMs, this study introduces a novel approach termed **Locally Adaptive Smoothing** (LAS). LAS enhances the adaptability and flexibility of spatial models by dynamically adjusting spatial weights based on local differences in expected disease counts. This method aims to preserve local spatial details while preventing over-smoothing, thereby providing a more accurate representation of disease risk patterns.

5.2.6.1 Dynamic Spatial Weights

LAS modifies the spatial weights between neighboring regions by incorporating a function that accounts for differences in expected disease counts. This dynamic adjustment ensures that regions with large differences in covariate values are treated distinctly, allowing the model to adapt to local variations more effectively.

5.2.6.2 Adjusted Spatial Weights

The LAS-adjusted spatial weights are defined as:

$$\omega_{ij}^{LAS} = \omega_{ij} \times \exp\left(-\frac{|E_i - E_j|}{\zeta}\right) \tag{5.3}$$

where:

- ω_{ij} represents the original spatial weight between regions i and j, typically based on adjacency or distance. If ω_{ij} is zero, the adaptive weights remain zero (i.e., the weights assigned to neighbors are adaptive, not the set of neighbors themselves).
- E_i and E_j are the estimated disease counts at locations i and j, respectively.
- ζ is a smoothing parameter that controls the sensitivity of the weight adjustment to differences in expected disease counts.

The introduction of Locally Adaptive Smoothing (LAS) brings several significant advancements to Bayesian Spatially Varying Coefficient Models (BSVCMs). One of the foremost contributions of LAS is its ability to reduce over-smoothing. By dynamically reducing the influence of dissimilar neighboring regions, LAS preserves intricate local spatial details that are often lost in traditional models. This preservation ensures that distinct disease risk patterns remain clearly identifiable, thereby maintaining the integrity of sharp spatial transitions without them being inadvertently blended with surrounding areas.

Furthermore, LAS enhances the adaptability of the model by allowing it to respond effectively to local discontinuities and heterogeneity. This adaptability is crucial in accurately representing regions with unique spatial characteristics, ensuring that areas exhibiting unusual or contrasting disease dynamics are distinctly captured. By accommodating such regional variations, LAS ensures that the model remains sensitive to the specific socioeconomic factors influencing disease risk in different locales. Collectively, these contributions enable BSVCMs to provide a more precise and nuanced understanding of spatial disease distribution, facilitating the development of targeted and effective public health interventions.

5.2.7 Restructured Spatially Varying Coefficient Model with LAS

Building upon the LAS approach, this study introduces a restructured BSVCM that integrates LAS-adjusted spatial weights to further enhance model performance. This restructured model addresses multiple limitations of existing models, offering improved accuracy and flexibility in capturing complex spatial patterns.

5.2.7.1 Model Specification

The restructured model is specified as follows:

$$\theta_i = \beta_0 + \sum_{k=1}^{K} \beta_k(x_{k,i}) + U_i^{LAS} \tag{5.4}$$

where:

- θ_i is the log-relative risk of disease at location i.
- $\beta_0 + U_i^{LAS}$ in the spatially varying intercept term.
- $\beta_{k,i}(x_{k,i})$ represents the effect of covariate k at location i, allowing for spatial variation in covariate effects.

- U_i^{LAS} denotes the spatially structured random effect incorporating LAS-adjusted weights, capturing residual spatial variation not explained by covariates.

The spatial random effects U_i^{LAS} are modeled as:

$$[U_i^{LAS}|\{U_j = u_j, i \sim j\}, \tau_u, \omega_{ij}^{LAS}] \sim \mathscr{N}\left(\bar{u_i}, \frac{1}{\sum_{j=1}^{n} \tau_u \omega_{ij}^{LAS}}\right) \tag{5.5}$$

where:

- $\bar{u_i}$ is the mean disease count of neighboring regions, serving as the conditional mean in the ICAR prior.
- τ_u is the precision parameter controlling the variance of the spatial random effects.
- $\omega_{ij}^{LAS} = \omega_{ij} \cdot \exp\left(-\frac{|E_i - E_j|}{\zeta}\right)$ represents the LAS-adjusted spatial weight between regions i and j.

Integrating Locally Adaptive Smoothing (LAS) into the Bayesian Spatially Varying Coefficient Model (BSVCM) offers several significant advancements. Primarily, the incorporation of LAS-adjusted weights effectively mitigates the over-smoothing and adaptability issues inherent in traditional Intrinsic Conditional Autoregressive (ICAR) models. This enhancement allows for a more nuanced and accurate representation of spatial disease patterns by ensuring that significant local variations are preserved and accurately depicted in the disease risk estimates.

Furthermore, the dynamic adjustment of spatial weights through LAS substantially enhances the model's accuracy and flexibility. By responding to complex spatial dependencies and local variations, the model becomes adept at capturing intricate spatial dynamics that are often present in heterogeneous environments such as Metro

Atlanta. This capability results in more precise estimates of disease risk, as the model can adapt to the unique spatial characteristics of different regions without being constrained by global smoothing parameters.

Additionally, the restructured model with LAS maintains an optimal balance between flexibility and interpretability. This balance facilitates clearer communication of spatial risk factors to stakeholders, as the model retains the ability to model intricate spatial patterns while ensuring that the results remain comprehensible and actionable. The enhanced interpretability of the model aids in the effective dissemination of findings and supports informed decision-making in public health interventions.

5.2.8 Case Study: Application to Tuberculosis Mapping in Metro Atlanta

To demonstrate the efficacy of the proposed LAS-enhanced Bayesian Spatially Varying Coefficient Model (BSVCM), we applied the model to a case study of tuberculosis (TB) incidence in Metro Atlanta. This region presents a challenging environment for spatial disease mapping due to its significant socio-economic diversity, varying population densities, and disparate access to healthcare services. These factors contribute to a complex spatial landscape, where disease risk can vary markedly over short distances, necessitating advanced modeling techniques to capture the underlying spatial dynamics accurately.

5.2.8.1 Model Implementation

The LAS-enhanced BSVCM was implemented using the `cmdstanr` package in R [47], which interfaces with the `CmdStan` backend for Bayesian inference. The model was configured to run 4000 warmup iterations and 1000 sampling iterations per chain, ensuring adequate exploration of the posterior distribution while maintaining computational efficiency. Multiple Markov Chain Monte Carlo (MCMC) chains were employed to verify convergence, with diagnostics including the potential scale reduction factor ($\hat{R}$) and effective sample size (ESS) being meticulously evaluated. These diagnostics confirmed the reliability of the posterior estimates, with $\hat{R}$ values close to 1.00 and high ESS values indicating effective sampling and model convergence.

The implementation process involved specifying appropriate priors for all model parameters to ensure proper regularization and facilitate efficient sampling. The smoothing parameter τ is pivotal in determining the influence of neighboring regions based on the differences in their expected disease counts. To appropriately model τ, a Gamma distribution was selected as the prior, specifically $tau \sim \text{Gamma}(2, 1)$.

The Gamma distribution is a preferred choice for positive continuous parameters within Bayesian hierarchical models due to its flexibility and conjugacy properties, which facilitate analytical tractability and efficient computation. The selection of the Gamma distribution with shape parameter $\alpha = 2$ and rate parameter $\beta = 1$ reflects a balance between informativeness and flexibility. This configuration yields a mean of $\alpha/\beta = 2$ and a variance of $\alpha/\beta^2 = 2$, providing a moderately informative prior that incorporates prior knowledge without imposing overly restrictive constraints on the parameter.

By setting τ to follow a $\text{Gamma}(2, 1)$ distribution, the model allows for sufficient variability in the smoothing parameter, enabling the LAS mechanism to adapt dynamically to the spatial heterogeneity present in the data. This prior facilitates the reduction of the influence of dissimilar neighbors, thereby preserving local spatial details and preventing the over-smoothing of distinct disease risk patterns. Furthermore, the chosen prior supports the model's capacity to respond to localized discontinuities and heterogeneity, ensuring that regions with unique spatial characteristics are accurately represented in the disease risk estimates.

The developed model estimates the log-relative risk θ_i for each ZIP code. The incorporation of LAS-adjusted spatial random effects U_i^{LAS} captures residual spatial variation not explained by the covariates, thereby accounting for unobserved heterogeneity and localized disease dynamics. This comprehensive modeling approach facilitates the generation of detailed and precise maps of TB incidence, effectively highlighting areas with elevated risk and supporting the implementation of targeted public health interventions.

5.2.9 Model Validation and Evaluation

To ensure the robustness and validity of the proposed Locally Adaptive Smoothing (LAS)-enhanced Bayesian Spatially Varying Coefficient Model (BSVCM), the model underwent rigorous validation and evaluation procedures. A fundamental component of this evaluation was the implementation of posterior predictive checks. These checks involved comparing the posterior predictive distributions generated by the model with the observed TB incidence data. This comparison served as a critical assessment of the model's ability to accurately capture the underlying data-generating process, allowing for the identification of any discrepancies between the model's predictions and the actual observations. By examining the alignment between predicted and observed values, we were able to ascertain the model's effectiveness in representing the spatial distribution of TB cases.

In addition to posterior predictive checks, several performance metrics were employed to comprehensively evaluate the model's effectiveness and reliability. The Expected Log Predictive Density (ELPD) was calculated to assess the model's predictive accuracy. Higher ELPD values indicated superior performance in terms of the model's ability to predict new data points accurately. Furthermore, the Watanabe-Akaike Information Criterion (WAIC) was utilized to balance model fit and complexity. WAIC provided a means to evaluate how well the model explained the data while penalizing for unnecessary complexity, with lower WAIC values signifying a better trade-off between fit and complexity.

Convergence diagnostics were meticulously monitored to ensure the reliability of the posterior estimates. The potential scale reduction factor ($\hat{R}$) was used to assess whether the Markov Chain Monte Carlo (MCMC) chains had converged to the target distribution. Values of $\hat{R}$ approaching 1.00 indicated that the chains had sufficiently converged, thereby guaranteeing the reliability of the posterior estimates. Additionally, the Effective Sample Size (ESS) was evaluated to determine the efficiency of the sampling process. Higher ESS values reflected better exploration of the posterior distribution and more reliable parameter estimates, ensuring that the model's inferences were both accurate and stable.

These validation and evaluation steps collectively ensured that the LAS-enhanced BSVCM provided reliable and accurate estimates of TB incidence across the study area. By rigorously assessing the model's predictive capabilities and convergence, we confirmed its suitability for detailed and precise spatial disease mapping, thereby supporting informed public health interventions and policy decisions.

5.3 Results

5.3.1 Posterior SVC Estimates

The application of the proposed Locally Adaptive Smoothing (LAS)-enhanced Bayesian Spatially Varying Coefficient Model (BSVCM) to tuberculosis (TB) incidence data in Metro Atlanta provided robust insights into the spatial variability of disease incidence. The model produced spatially varying coefficients (SVCs) for the key covariates—percentage of foreign-born residents, HIV prevalence, and the Gini index—allowing for more nuanced interpretations of how these socio-economic factors influence TB risk in different areas. The random intercepts and slopes for each covariate confirmed significant spatial heterogeneity across the study region, supporting the need for location-specific adjustments in the model.

The posterior estimates for the spatially varying intercept (β_0) revealed clear regional clusters where the baseline risk of TB was either elevated or diminished. The coefficients for foreign-born residents (β_1) showed a strong positive association with TB incidence in areas with high immigrant populations, while HIV prevalence (β_2) had a similarly strong positive impact on TB risk in urban centers. The Gini index (β_3), representing income inequality, exhibited more localized effects, with high values contributing to elevated TB risk in certain high-poverty neighborhoods.

These findings underscore the importance of modeling spatial heterogeneity in TB incidence, as the impact of each covariate varies substantially across the region. The inclusion of LAS-adjusted spatial random effects (U_i^{LAS}) allowed for even finer adjustments to local spatial trends, further improving the model's predictive accuracy.

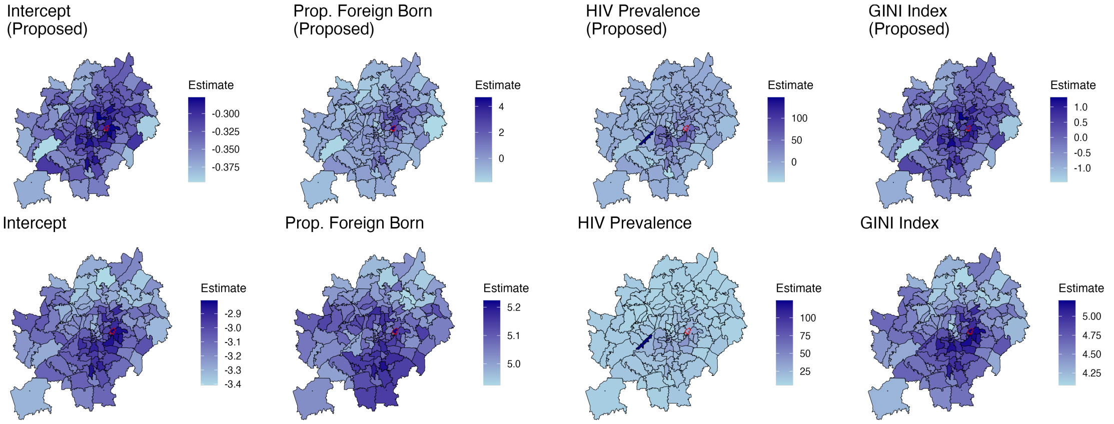


Figure 5.1: Posterior estimates of spatially varying coefficients (SVCs) for TB incidence across Metro Atlanta comparing the proposed versus existing models.

5.3.2 Posterior Predictive Checks

Posterior predictive checks (PPCs) were conducted to assess the model's ability to capture the underlying data-generating process. By comparing the predicted TB case counts to the observed values, we evaluated how well the LAS-enhanced BSVCM aligned with the actual incidence data. The PPC plots (Figure 5.2) demonstrated that the proposed model provided a close fit to the observed data, particularly in areas with higher TB counts.

Notably, the LAS-enhanced model outperformed the existing BSVCM in regions with sharp spatial transitions, such as areas experiencing rapid demographic changes or significant socio-economic disparities. These transitions were often under- or overestimated by traditional models, but the adaptive smoothing provided by the LAS mechanism captured these nuances effectively.

5.3.3 Predictive Interval Coverage

To further assess the model's performance, we evaluated the coverage of observed TB counts within the 95% posterior predictive intervals. The proposed LAS-enhanced BSVCM achieved a predictive interval coverage of 88.71%, exceeding the 86.52% coverage achieved by the current BSVCM. This improvement in coverage reflects the model's enhanced ability to quantify uncertainty in its predictions, particularly in regions with high spatial variability. The higher coverage rate indicates that the LAS-enhanced model better captures the true variability in TB counts, making it a more reliable tool for public health decision-making.

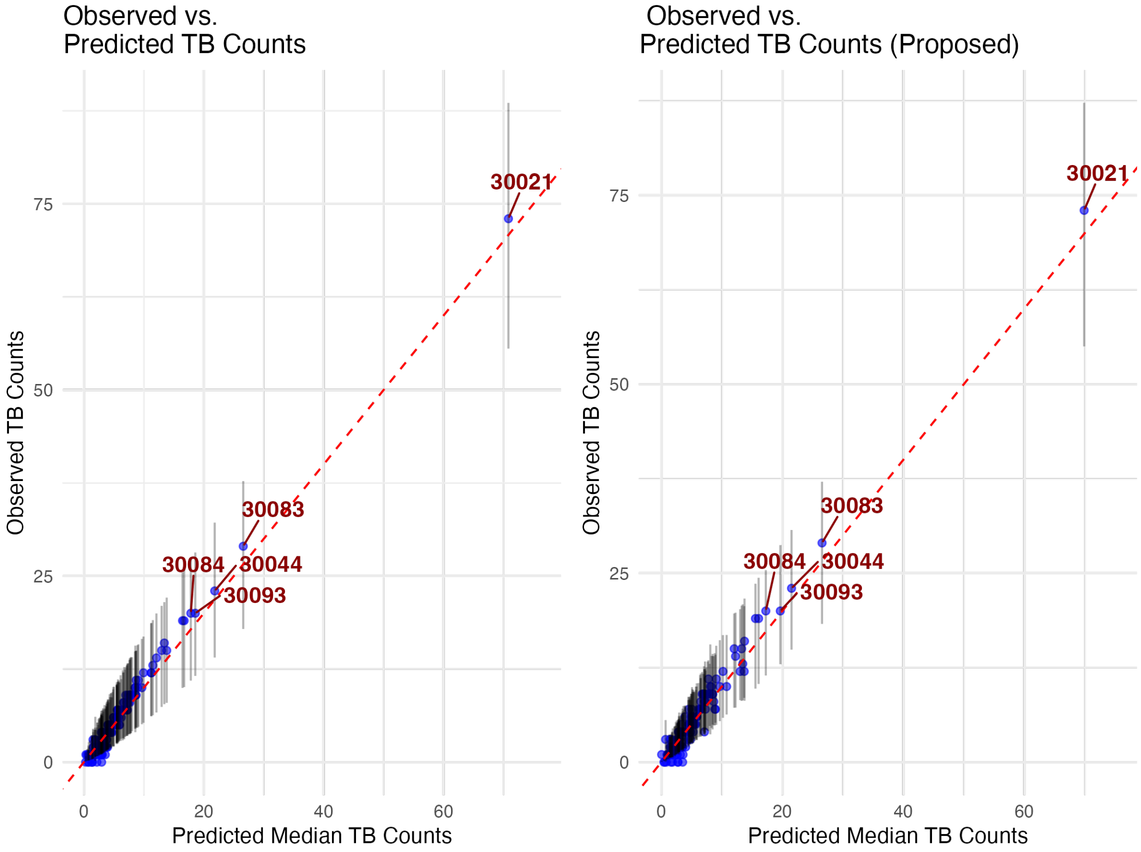


Figure 5.2: Posterior predictive checks comparing the proposed versus existing models, showing predicted versus observed TB counts across Metro Atlanta.

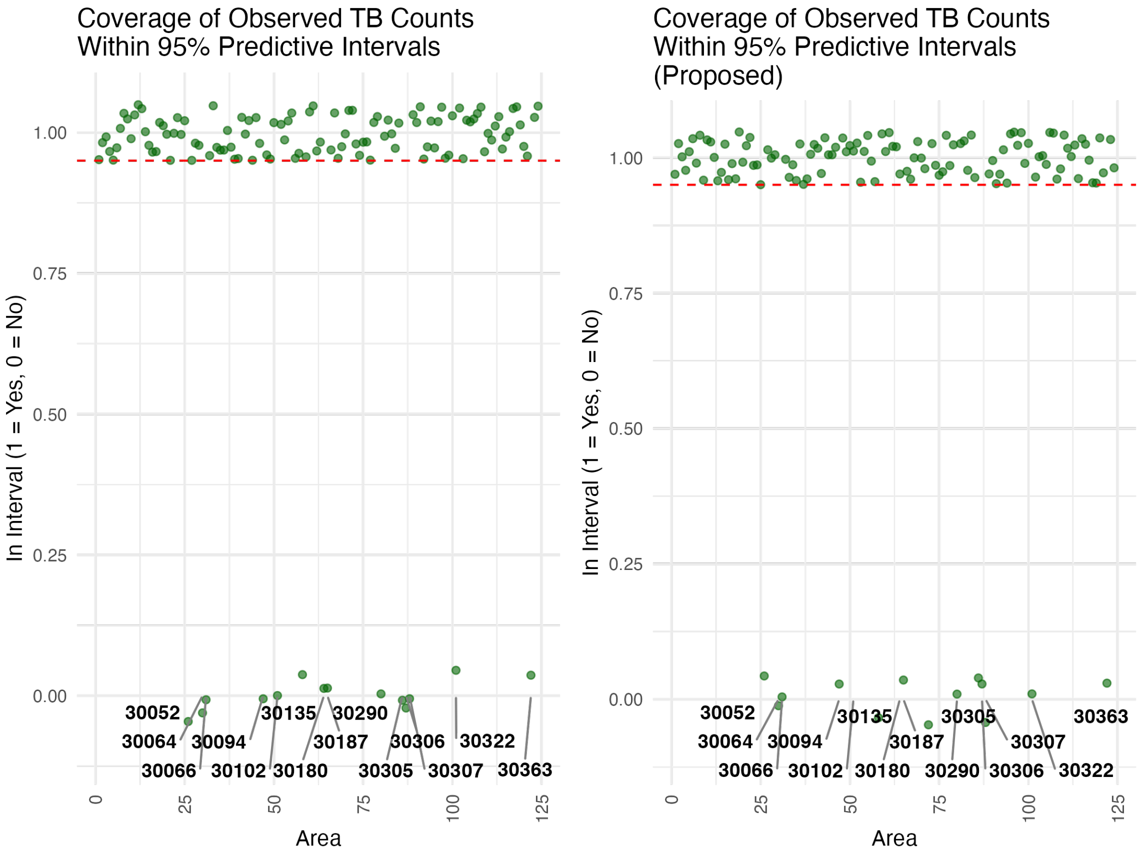


Figure 5.3: Coverage of observed TB counts within 95% posterior predictive intervals for both the LAS-enhanced BSVCM and the current BSVCM. The LAS-enhanced model achieved a coverage of 88.71%, while the current BSVCM achieved a coverage of 86.52%.

5.4 Discussion

5.4.1 Comparison to Existing Models

The LAS-enhanced BSVCM demonstrated significant improvements over the existing BSVCM in both predictive performance and model convergence. The median posterior estimate for the Standardized Incidence Ratio for the proposed model was 1.37, with a median absolute deviation (MAD) of 0.56, indicating a moderate level of variability that reflects the true spatial heterogeneity in TB incidence across Metro Atlanta. By contrast, the existing model had a lower median of 1.22 and a smaller MAD of 0.46, suggesting that it failed to capture some of the spatial variability present in the data.

Convergence diagnostics were particularly robust for the proposed model, with all chains exhibiting $\hat{R} = 1.00$, signaling near-perfect convergence. The existing BSVCM, on the other hand, had minor convergence issues, with $\hat{R} = 1.07$, indicating some instability in the posterior estimates. The effective sample size (ESS) for the proposed model was also considerably higher (5117 for bulk ESS, 3106 for tail ESS) compared to the existing model, which had much lower ESS values (39 for bulk, 59 for tail). These metrics indicate that the proposed model explored the parameter space more efficiently, leading to more reliable posterior estimates.

While the runtime for the LAS-enhanced model was longer (3552 seconds compared to 252 seconds for the existing model), the additional computational cost is justified by the improvements in model performance and predictive accuracy.

Model	**Median**	**MAD**
LAS-enhanced BSVCM	1.37	0.56
Current BSVCM	1.22	0.46

Table 5.1: Comparison of posterior medians and median absolute deviations (MAD) between the proposed and existing BSVCM models.

5.4.2 Predictive Accuracy and Model Complexity

The proposed model exhibited superior predictive accuracy as evidenced by improvements in Expected Log Predictive Density (ELPD) and Watanabe-Akaike Information Criterion (WAIC). The ELPD difference for the proposed model was -6.5, which marginally outperformed the current model's ELPD of -7.6. Similarly, the WAIC for the proposed model was 606.5, lower than the 608.8 WAIC of the existing model, indicating a better balance between model complexity and predictive accuracy.

Moreover, the Mean Absolute Error (MAE) for the proposed model was 0.097, a substantial improvement over the 1.14 MAE of the current model. This reduction in error demonstrates the effectiveness of the LAS mechanism in improving prediction accuracy, particularly in areas with sharp spatial boundaries or localized clusters of high TB incidence.

Model	**ELPD**	**WAIC**
LAS-enhanced BSVCM	-6.5	606.5
Current BSVCM	-7.6	608.8

Table 5.2: Comparison of ELPD and WAIC values between the proposed and existing BSVCM models.

5.4.3 Implications for Public Health Interventions

The LAS-enhanced BSVCM offers substantial improvements in the spatial mapping of TB incidence, which directly informs public health interventions. By capturing both the local and global effects of socio-economic and demographic factors on TB incidence, the model provides more accurate identification of high-risk areas. This allows public health officials to allocate resources more effectively, focusing interventions on the most vulnerable populations.

The model's improved uncertainty quantification also enhances its utility in decision-making. Public health strategies, especially in regions like Metro Atlanta where socio-economic disparities are prominent, benefit from the model's ability to capture complex spatial patterns and provide reliable predictions of TB incidence. This is particularly important when considering the resource limitations faced by public health agencies, as targeted interventions can now be based on more reliable data.

In summary, the proposed Locally Adaptive Smoothing (LAS) approach significantly enhances Bayesian Spatially Varying Coefficient Models (BSVCMs) by addressing their inherent limitations. By dynamically adjusting spatial weights based on local differences in expected disease counts, LAS prevents over-smoothing and improves the model's ability to capture complex spatial patterns. The application of this enhanced model to tuberculosis mapping in Metro Atlanta demonstrates its practical utility and effectiveness in uncovering nuanced spatial risk factors. These advancements not only contribute to the field of spatial epidemiology but also hold substantial promise for improving public health interventions and policy formulation.

Chapter 6

Conclusions and Future Directions

6.1 Summary of Contributions

This dissertation makes several key contributions to the field of spatial disease modeling through the development and application of advanced Bayesian methodologies.

In **Aim 1**, hierarchical Bayesian modeling was explored for the purpose of disease mapping, with a particular focus on spatially concentrated disease prevalence. The use of the Laplace Prior, commonly employed in genetic mapping, was adapted to capture sharp spatial transitions in disease incidence, demonstrating its effectiveness in improving model flexibility and accuracy in regions with concentrated health disparities.

In **Aim 2**, the utility of local Deviance Information Criterion (DIC) as a diagnostic tool was introduced to detect local lack of fit in regions with sharp changes in disease prevalence. This contribution provided a novel approach to evaluating model fit in spatial epidemiology, offering a more granular method for identifying areas where traditional models may fail to capture the nuances of disease distribution.

Finally, **Aim 3** focused on the development of an enhanced Bayesian Spatially Varying Coefficient Model (BSVCM) integrated with Locally Adaptive Smoothing (LAS). This model was specifically designed to address the limitations of over-smoothing in traditional models while also capturing spatial outliers. The LAS-enhanced BSVCM demonstrated significant improvements in modeling accuracy, particularly in regions with high spatial variability, making it a valuable tool for more accurate disease risk mapping and public health interventions.

These contributions collectively advance the methodological landscape of spatial disease modeling, providing more accurate, flexible, and diagnostic-rich tools for analyzing and predicting disease prevalence in heterogeneous environments.

6.2 Future Directions and Conclusion

While this research has improved our understanding of spatial disease modeling, there are several areas to explore further to enhance and apply these methods.

6.2.1 Application to Middle-Income Countries with Wealth Disparities

A key future direction is to apply these advanced Bayesian spatial models to study TB in middle-income countries outside the United States. Countries like India and South Africa have large wealth disparities and diverse socioeconomic conditions that greatly affect the spatial distribution of TB. In these countries, urban areas often have a sharp contrast between wealthy neighborhoods and informal settlements or townships, where overcrowding, poor sanitation, and limited access to healthcare make TB spread more easily.

In India, for example, the mix of rapidly developing urban centers and densely populated slums creates a complex environment for TB transmission. Applying our models here would involve including detailed socioeconomic data to capture the sharp changes in TB incidence due to wealth disparities. Similarly, in South Africa, the history of apartheid has led to noticeable spatial segregation, with under-resourced areas experiencing higher TB rates. Our models can help understand how these historical and socioeconomic factors affect current disease patterns.

Understanding the spatial distribution of socioeconomic groups offers valuable insights for effectively monitoring and surveying TB, especially in countries with significant wealth disparities. Socioeconomic factors influence many health determinants, including nutrition, living conditions, education, and access to medical care. Mapping socioeconomic data alongside TB incidence reveals which communities are most at risk. In India, this could highlight rural areas with limited healthcare access or urban slums with high population density. Socioeconomic conditions often determine living environments that can either hinder or facilitate disease spread; for example, overcrowded housing in Mumbai's Dharavi slum contributes to higher TB transmission rates. Spatial analysis can uncover systemic differences in healthcare access and disease burden, informing interventions focused on equity. In South Africa, disparities in healthcare infrastructure between urban and rural areas can be identified and addressed.

Utilizing this approach we will be able regions where TB rates are much higher due to socioeconomic factors. For instance, mapping TB incidence in relation to income levels in Mumbai's slums or Cape Town's townships can highlight areas needing urgent attention. Additionally, this approach will allow public health officials in those areas to direct limited healthcare resources to the communities that need them most, improving intervention strategies. This is especially critical in countries where healthcare funding is limited, so targeting resources effectively is crucial. By focusing on countries with significant wealth disparities like India and South Africa, we can contribute to global TB control efforts, especially in regions where the disease burden is high and resources are limited.

6.2.2 Adjusting Spatial Weighting Schemes

Another promising area for future work is adjusting the weighting scheme to use covariate values instead of expected disease counts. By basing spatial weights on differences in covariates—such as socioeconomic status, access to healthcare, or environmental factors—between neighboring regions, the models become more sensitive to underlying risk factors. This adjustment allows the models to capture spatial differences in covariate values, leading to more accurate risk estimates. For instance, considering air pollution levels in different parts of Delhi that may influence TB incidence. This is especially important in areas where population density does not directly correlate with TB risk due to varying living conditions. Lastly, adjusting the spatial weights to covariate similarities in neighboring areas will provide insights that inform more effective, targeted public health interventions based on specific local needs. Tailoring interventions to address factors like malnutrition or HIV co-infection rates in specific communities.

6.2.3 Broadening Applicability to Other Infectious Diseases

The flexibility of Locally Adaptive Smoothing Bayesian Spatially Varying Coefficient Models makes them suitable for mapping other infectious diseases with complex spatial dynamics, such as malaria, HIV/AIDS, or COVID-19. Comparative studies across different diseases can help validate and refine the models' applicability, ensuring they are robust and generalize well across various epidemiological contexts. For instance, applying the models to malaria incidence in sub-Saharan Africa could test their effectiveness in different transmission settings. These models can also provide cross-disease insights by identifying common spatial patterns and risk factors present in multiple diseases. Exploring how socioeconomic factors influence both TB and HIV/AIDS distribution in South Africa may reveal underlying connections. Additionally, enhancing public health preparedness is possible by offering tools that can be quickly adapted to new infectious diseases, improving response times. Modifying the models to map the spread of emerging diseases like COVID-19 could aid in rapid response efforts.

Conclusion

This dissertation has advanced the field of spatial disease modeling through Bayesian methods, addressing important challenges in accurately mapping and predicting TB incidence. By extending these models to new contexts—particularly middle-income countries with significant wealth disparities like India and South Africa—and including socioeconomic data, we can enhance disease monitoring, surveillance, and intervention strategies.

Future work building on these foundations holds significant potential for reducing the global burden of TB and improving public health outcomes. By continuing to refine these models and apply them to diverse settings, we contribute to a better understanding of disease dynamics and support efforts to achieve health equity worldwide.

Bibliography

[1] Constitution of the United States. `https://www.archives.gov/founding-docs/constitution`, 1787. Accessed: 2024-08-25.

[2] Deborah A Adams, Kathleen E Fullerton, Ruth A Jajosky, Pearl Sharp, Diana H Onweh, Alan W Schley, Willie J Anderson, Amanda Faulkner, and Kiersten J Kugeler. Summary of notifiable infectious diseases and conditions–united states, 2013. 2015.

[3] AIDSVu. Deeper look: Ending the hiv epidemic in atlanta, 2024. URL `https://aidsvu.org/resources/deeper-look-ehe-atlanta/`. Accessed: 2024-06-19.

[4] Aswi Aswi, SM Cramb, Paula Moraga, and Kerrie Mengersen. Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. *Epidemiology & Infection*, 147:e33, 2019.

[5] Robert D Baller, Luc Anselin, Steven F Messner, Glenn Deane, and Darnell F Hawkins. Structural covariates of us county homicide rates: Incorporating spatial effects. *Criminology*, 39(3):561–588, 2001.

[6] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. Hierarchical modeling and analysis for spatial data. *Chapman and Hall/CRC*, 2004.

[7] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. Hierarchical modeling and analysis for spatial data. *CRC Press*, 2014.

[8] Thomas Bayes. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, (53):370–418, 1763.

[9] Thomas Bayes and David Hume. Bayes's theorem. In *Proceedings of the British Academy*, volume 113, pages 91–109, 1763.

[10] Luisa Bernardinelli, D Clayton, and Cristina Montomoli. Bayesian estimates of disease maps: how important are priors? *Statistics in medicine*, 14(21-22): 2411–2431, 1995.

[11] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

[12] Julian Besag, Jeremy York, Ann Mollié, and Robert Gentlemen. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991.

[13] Julian Besag, Peter Green, David Higdon, and Kerrie Mengersen. Bayesian computation and stochastic systems. *Statistical science*, pages 3–41, 1995.

[14] Nicola G Best, Richard A Arnold, Andrew Thomas, Lance A Waller, and Erin M Conlon. Bayesian models for spatially correlated disease and exposure data. *Bayesian statistics*, 6:131–156, 1999.

[15] Nicola G Best, Richard A Arnold, Andrew Thomas, Lance A Waller, and Erin M

Conlon. Bayesian models for spatially correlated disease and exposure data. *Bayesian statistics*, 6:131–156, 1999.

[16] Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. The horseshoe+ estimator of ultra-sparse signals. 2017.

[17] Marta Blangiardo, Michela Cameletti, Gianluca Baio, and Håvard Rue. Spatial and spatio-temporal models with r-inla. *Spatial and spatio-temporal epidemiology*, 4:33–49, 2013.

[18] Henry M Blumberg, William J Burman, Richard E Chaisson, Charles L Daley, et al. American thoracic society/centers for disease control and prevention/infectious diseases society of america: treatment of tuberculosis. *American journal of respiratory and critical care medicine*, 167(4):603, 2003.

[19] Benjamin M Bolker. Linear and generalized linear mixed models. *Ecological statistics: contemporary theory and application*, pages 309–333, 2015.

[20] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[21] John S Brownstein, Clark C Freifeld, Ben Y Reis, and Kenneth D Mandl. Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Medicine*, 5(7):e151, July 2008. doi: 10.1371/journal.pmed.0050151.

[22] Kevin P Cain, Connie A Haley, Lori R Armstrong, Katie N Garman, Charles D Wells, Michael F Iademarco, Kenneth G Castro, and Kayla F Laserson. Tuberculosis among foreign-born persons in the united states: achieving tuberculosis elimination. *American journal of respiratory and critical care medicine*, 175(1): 75–79, 2007.

[23] Kevin P Cain, Stephen R Benoit, Carla A Winston, and William R Mac Kenzie. Tuberculosis among foreign-born persons in the united states. *Jama*, 300(4): 405–412, 2008.

[24] Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial intelligence and statistics*, pages 73–80. PMLR, 2009.

[25] CDC. Principles of epidemiology in public health practice: an introduction to applied epidemiology and biostatistics, 2006.

[26] CDC. Tuberculosis. 2020.

[27] Michael P Chen, Nong Shang, Carla A Winston, and Jose E Becerra. A bayesian analysis of the 2009 decline in tuberculosis morbidity in the united states. *Statistics in medicine*, 31(27):3278–3284, 2012.

[28] Ming-Hui Chen and Joseph G Ibrahim. Conjugate priors for generalized linear models. *Statistica Sinica*, pages 461–476, 2003.

[29] Kocfa Chung-Delgado, Sonia Guillen-Bravo, Alejandro Revilla-Montag, and Antonio Bernabe-Ortiz. Mortality among mdr-tb cases: comparison with drug-susceptible tuberculosis and associated factors. *PloS one*, 10(3):e0119332, 2015.

[30] David Clayton and John Kaldor. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, pages 671–681, 1987.

[31] Stephen R Cole, Haitao Chu, and Sander Greenland. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *American journal of epidemiology*, 179(2):252–260, 2014.

[32] Robert R Corbeil and Shayle R Searle. Restricted maximum likelihood (reml)

estimation of variance components in the mixed model. *Technometrics*, 18(1): 31–38, 1976.

[33] Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

[34] David GT Denison and Christofer C Holmes. Bayesian partitioning for estimating disease risk. *Biometrics*, 57(1):143–149, 2001.

[35] Molly Deutsch-Feldman, Robert H Pratt, Sandy F Price, Clarisse A Tsang, and Julie L Self. Tuberculosis—united states, 2020. *Morbidity and Mortality Weekly Report*, 70(12):409, 2021.

[36] Christopher Dye. Breaking a law: tuberculosis disobeys styblo's rule, 2008.

[37] Paul Elliot, Jon C Wakefield, Nicola G Best, and David John Briggs. *Spatial epidemiology: methods and applications*. 2000.

[38] Frank A Farris. The gini index and measures of inequality. *The American Mathematical Monthly*, 117(10):851–864, 2010.

[39] James R Faulkner and Vladimir N Minin. Locally adaptive smoothing with markov random fields and shrinkage priors. *Bayesian analysis*, 13(1):225, 2018.

[40] Centers for Disease Control and Prevention (CDC). Reported tuberculosis in the united states, 2021. Technical report, US Department of Health and Human Services, CDC, Atlanta, GA, 2022. URL `https://www.cdc.gov/tb/statistics/reports/2021/default.htm`.

[41] Centers for Disease Control, Prevention, et al. Report of verified case of tuberculosis (rvct), instruction manual. Technical report, 2021.

[42] Centers for Disease Control, Prevention (US), National Center for Prevention Services (US). Division of Tuberculosis Elimination, National Center for HIV,

STD, and TB Prevention (US). Division of Tuberculosis Elimination. *Reported tuberculosis in the United States, 2020.* US Department of Health and Human Services, Public Health Service, Centers . . . , 2021.

[43] Centers for Disease Control et al. Multidrug-resistant tuberculosis (mdr tb) fact sheet. Technical report, 2016.

[44] Centers for Disease Control et al. Extensively drug-resistant tuberculosis (xdr tb) fact sheet. Technical report, 2016.

[45] Anne Marie France, Juliana Grant, J Steve Kammerer, and Thomas R Navin. A field-validated approach using surveillance and genotyping data to estimate tuberculosis attributable to recent transmission in the united states. *American journal of epidemiology*, 182(9):799–807, 2015.

[46] Jonah Gabry, Rok Češnovar, Andrew Johnson, and Steve Bronder. *cmdstanr: R Interface to 'CmdStan'*, 2024. URL `https://mc-stan.org/cmdstanr/`. R package version 0.8.1, https://discourse.mc-stan.org.

[47] Jonah Gabry, Rok Češnovar, Andrew Johnson, and Steve Bronder. *cmdstanr: R Interface to 'CmdStan'*, 2024. URL `https://mc-stan.org/cmdstanr/`. R package version 0.8.1, https://discourse.mc-stan.org.

[48] Alan E. Gelfand, Hyon-Jung, Kim, C. F. Sirmans, Sudipto, and Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98:387 – 396, 2003. URL `https://api.semanticscholar.org/CorpusID:122987154`.

[49] A Gelman et al. Bayesian data analysis, 3rd: Boca raton. *Texts in Statistical Science*, 2013.

[50] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[51] Andrew Gelman and Donald B Rubin. Markov chain monte carlo methods in biostatistics. *Statistical methods in medical research*, 5(4):339–355, 1996.

[52] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[53] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[54] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[55] Malay Ghosh, Kannan Natarajan, Lance A Waller, and Dalho Kim. Hierarchical bayes glms for the analysis of spatial data: An application to disease mapping. *Journal of Statistical Planning and Inference*, 75(2):305–318, 1999.

[56] Corrado Gini. Measurement of inequality of incomes. *The economic journal*, 31(121):124–125, 1921.

[57] Marian Goble, Michael D Iseman, Lorie A Madsen, Dennis Waite, Lynn Ackerson, and C Robert Horsburgh Jr. Treatment of 171 patients with pulmonary tuberculosis resistant to isoniazid and rifampin. *New England journal of medicine*, 328(8):527–532, 1993.

[58] Peter J Green and Sylvia Richardson. Hidden markov models and disease mapping. *Journal of the American statistical association*, 97(460):1055–1070, 2002.

[59] David A Harville. [that blup is a good thing: The estimation of random effects]: Comment. *Statistical Science*, 6(1):35–39, 1991.

[60] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[61] Charles R Henderson. Estimation of genetic parameters. In *Biometrics*, volume 6, pages 186–187. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 1950.

[62] AN Hill, JE Becerra, and KG Castro. Modelling tuberculosis trends in the usa. *Epidemiology & Infection*, 140(10):1862–1872, 2012.

[63] N Thompson Hobbs and Mevin B Hooten. *Bayesian models: a statistical primer for ecologists*. Princeton University Press, 2015.

[64] James S Hodges. *Richly parameterized linear models: additive, time series, and spatial models using random effects*. CRC Press, 2013.

[65] James S Hodges and Murray K Clayton. Random effects old and new. *Stat. Sci*, 433, 2011.

[66] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1): 1593–1623, 2014.

[67] M Jun, M Katzfuss, J Hu, and VE Johnson. Assessing fit in bayesian models for spatial processes. *Environmetrics*, 25(8):584–595, 2014.

[68] Min Kyong Kim, Jayanta Bhattacharya, and Joydeep Bhattacharya. Is income inequality linked to infectious disease prevalence? a hypothesis-generating study using tuberculosis. *Social Science & Medicine*, 345:116639, 2024.

[69] Gary King. Geography, statistics, and ecological inference, 2000.

[70] Thomas Kistemann, Annette Munzinger, and Friederike Dangendorf. Spatial patterns of tuberculosis incidence in cologne (germany). *Social science & medicine*, 55(1):7–19, 2002.

[71] L. Knorr-Held and G. Rasser. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21, 2000.

[72] Karl-Rudolf Koch and Karl-Rudolf Koch. Bayes' theorem. *Bayesian Inference with Geodetic Applications*, pages 4–8, 1990.

[73] John Kruschke. Doing bayesian data analysis: A tutorial with r, jags, and stan. 2014.

[74] John K Kruschke and Wolf Vanpaemel. Bayesian estimation in hierarchical models. *The Oxford handbook of computational and mathematical psychology*, pages 279–299, 2015.

[75] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[76] Andrew Lawson and Duncan Lee. Bayesian disease mapping for public health. In *Handbook of statistics*, volume 36, pages 443–481. Elsevier, 2017.

[77] Andrew B Lawson. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press, 2018.

[78] Duncan Lee. A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and spatio-temporal epidemiology*, 2(2):79–89, 2011.

[79] Pei Li, Sudipto Banerjee, Timothy A Hanson, and Alexander M McBean. Bayesian models for detecting difference boundaries in areal data. *Statistica Sinica*, 25(1):385, 2015.

[80] Ying C MacNab. Hierarchical bayesian modeling of spatially correlated health service outcome and utilization rates. *Biometrics*, 59(2):305–315, 2003.

[81] Ying C MacNab. On gaussian markov random fields and bayesian disease mapping. *Statistical Methods in Medical Research*, 20(1):49–68, 2011.

[82] Elvin Magee, Cheryl Tryon, Alstead Forbes, Bruce Heath, and Lilia Manangan. The national tuberculosis surveillance system training program to ensure accuracy of tuberculosis data. *Journal of Public Health Management and Practice*, 17(5):427–430, 2011.

[83] Måns Magnusson, Aki Vehtari, Johan Jonasson, and Michael Andersen. Leave-one-out cross-validation for bayesian model comparison in large data. In *International conference on artificial intelligence and statistics*, pages 341–351. PMLR, 2020.

[84] Charles C Margossian. A review of automatic differentiation and its efficient implementation. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9(4):e1305, 2019.

[85] Alan L Melnick. *Introduction to geographic information systems in public health.* Jones & Bartlett Learning, 2002.

[86] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[87] DA Mitchison and AJ Nunn. Influence of initial drug resistance on the response to short-course chemotherapy of pulmonary tuberculosis. *American Review of Respiratory Disease*, 133(3):423–430, 1986.

[88] Mitzi Morris, Katherine Wheeler-Martin, Dan Simpson, Stephen J Mooney, Andrew Gelman, and Charles DiMaggio. Bayesian hierarchical spatial models: Implementing the besag york mollié model in stan. *Spatial and spatio-temporal epidemiology*, 31:100301, 2019.

[89] Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 541–548. JMLR Workshop and Conference Proceedings, 2010.

[90] Radford M Neal. Mcmc using hamiltonian dynamics. *arXiv preprint arXiv:1206.1901*, 2012.

[91] Georgia Department of Public Health. Division of Medical and Clinical Program Services. Tuberculosis Program. 2020 georgia tuberculosis surveillance report. Technical report, 2021.

[92] Karen L Olson, Shaun J Grannis, and Kenneth D Mandl. Privacy protection versus cluster detection in spatial epidemiology. *American Journal of Public Health*, 96(11):2002–2008, 2006.

[93] Stan Openshaw. An empirical study of some spatial interaction models. *Environment and Planning A*, 8(1):23–41, 1976.

[94] World Health Organization. *Global tuberculosis report 2021*. World Health Organization, 2021.

[95] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[96] R Core Team. *R: A Language and Environment for Statistical Computing*. R

Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

[97] Andrea Riebler, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistics in Medicine*, 35(13):2275–2291, 2016. doi: 10.1002/sim.6897. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6897`.

[98] Brian D Ripley. *Spatial statistics*. John Wiley & Sons, 2005.

[99] George K Robinson. That blup is a good thing: the estimation of random effects. *Statistical science*, pages 15–32, 1991.

[100] Håvard Rue and Leonhard Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, 2005.

[101] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[102] Sarah Ryniker. Resettled refugees in the american south: Discourses of victimization and transgression in clarkston, georgia. *The Aliens Within: Danger, Disease, and Displacement in Representations of the Racialized Poor*, 80:315, 2022.

[103] Henry Scheffe. A" mixed model" for the analysis of variance. *The Annals of Mathematical Statistics*, pages 23–36, 1956.

[104] John Snow. Map showing the distribution of cholera cases in soho, london. Included in *On the Mode of Communication of Cholera*, 1854. Illustrates the spatial clustering of cholera deaths during the 1854 outbreak.

[105] John Snow. On the mode of communication of cholera. *Edinburgh medical journal*, 1(7):668, 1856.

[106] Terry Speed. [that blup is a good thing: the estimation of random effects]: Comment. *Statistical science*, 6(1):42–44, 1991.

[107] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.

[108] Stephen B. Thacker. Principles and practice of public health surveillance. edited by: Teutsch sm, churchill re, 2000.

[109] Stephen B Thacker and Ruth L Berkelman. Public health surveillance in the united states. *Epidemiologic reviews*, 10(1):164–190, 1988. doi: 10.1093/oxfordjournals.epirev.a036021.

[110] Stephen B. Thacker, Ruth L. Berkelman, and Donna F. Stroup. The science of public health surveillance. *Journal of Public Health Policy*, 10(2):187, 1989. doi: 10.2307/3342679.

[111] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[112] United States Census Bureau. 2020 Census Redistricting Data (Public Law 94-171) Summary File. `https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files/2020.html`, 2021. Table 1.

[113] United States Census Bureau. American Community Survey 2022 1-Year Estimates, Table B05012. `https://data.census.gov/cedsci/table?tid=ACSDT1Y2022.B05012`, 2022.

[114] Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

[115] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432, 2017.

[116] Christopher Vinnard, E Paul Wileyto, Gregory P Bisson, and Carla A Winston. First use of multiple imputation with the national tuberculosis surveillance system. *Epidemiology Research International*, 2013, 2013.

[117] Jonathan C Wakefield, NG Best, and L Waller. Bayesian approaches to disease mapping. *Spatial epidemiology: methods and applications*, 59, 2000.

[118] Lance A Waller. Discussion: statistical cluster detection, epidemiologic interpretation, and public health policy. *Statistics and Public Policy*, 2(1):1–8, 2015.

[119] Lance A Waller and Bradley P Carlin. Disease mapping. *Chapman & Hall/CRC handbooks of modern statistical methods*, 2010:217, 2010.

[120] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*. John Wiley & Sons, 2004.

[121] Lance A Waller, Li Zhu, Carol A Gotway, Dennis M Gorman, and Paul J Gruenewald. Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. *Stochastic Environmental Research and Risk Assessment*, 21:573–588, 2007.

[122] Stephen D Walter. Disease mapping: a historical perspective. *Spatial epidemiology: methods and applications*, pages 223–239, 2000.

[123] Hao Wang and Abel Rodríguez. Identifying pediatric cancer clusters in florida using log-linear models and generalized lasso penalties. *Statistics and Public Policy*, 1(1):86–96, 2014.

[124] Sumio Watanabe and Manfred Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

[125] David C. Wheeler and Lance A. Waller. Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *Journal of Geographical Systems*, 11: 1–22, 2009. URL `https://api.semanticscholar.org/CorpusID:11919651`.

[126] David C Wheeler, DeMarc A Hickson, and Lance A Waller. Assessing local model adequacy in bayesian hierarchical models using the partitioned deviance information criterion. *Computational statistics & data analysis*, 54(6):1657–1671, 2010.

[127] World Health Organization. *Global Tuberculosis Report 2021*. World Health Organization, Geneva, 2021. ISBN 9789240018662. URL `https://www.who.int/publications/i/item/9789240018662`.

[128] Yao Yuling, Vehtari Aki, Simpson Daniel, and Gelman Andrew. Using stacking to average bayesian predictive distributions. *Retrieved August17*, 2018.