SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of
Economics, Management and Statistics

PhD program: **Economics and Statistics** Cycle: **XXXVI**
Curriculum: **Statistics**

ADVANCED BAYESIAN MODELING IN PUBLIC HEALTH AND ENVIRONMENTAL RISK ASSESSMENT

Surname: **AIELLO**
Name: **LUCA**
Registration number: **868623**

Supervisor: Prof. **LUCIA PACI**
Co-Supervisor: Prof. **RAFFAELE ARGIENTO**

Coordinator: Prof. **MATTEO MANERA**

Academic Year: 2023/2024

To my family.

Small minds discuss people,
average minds discuss events,
great minds discuss ideas.

Eleanor Roosevelt

Abstract

Public health and environmental challenges, including air pollution, climate change, and natural disasters, pose significant threats to global well-being. Addressing these issues requires a deep understanding of the underlying data to inform effective policy and mitigation strategies. This thesis applies advanced Bayesian modeling techniques to three key areas within public health and environmental risk assessment: Bayesian clustering methods for environmental data, spatial disease mapping, and earthquake parameter estimation using smartphone accelerometer data. Bayesian models are particularly suited for these domains due to their ability to manage uncertainty, incorporate prior knowledge, and handle complex relationships within high-dimensional, noisy, or sparse datasets. The first project in this thesis explores Bayesian clustering techniques for environmental data, providing insights into hidden structures and patterns in spatio-temporal datasets. The second project introduces a Bayesian nonparametric approach to spatial disease mapping, enhancing flexibility and precision in capturing disease dynamics. This method incorporates key covariates such as population demographics and socioeconomic factors and accounts for complex spatial dependencies. The third project focuses on earthquake detection. It develops a Bayesian survival model to analyze crowd-sourced data from smartphone accelerometers, enhancing the accuracy of earthquake parameter estimation. This thesis underscores the potential to improve decision-making and policy implementation in public health and environmental risk management by demonstrating the versatility of Bayesian methods across these diverse applications. The work highlights how Bayesian techniques can overcome the limitations of traditional statistical methods, offering more reliable estimates and insights in fields where uncertainty and data limitations are prevalent.

Acknowledgements

With immense pleasure and a deep sense of gratitude, I wish to express my heartfelt thanks to my Supervisor, Lucia Paci, and my co-Supervisor, Raffaele Argiento. So much of this work has been possible thanks to their unwavering support and trust in me. Thank you both, truly, from the bottom of my heart.

I am also profoundly grateful to Sudipto Banerjee, who provided me with the opportunity to work with him and offered incredible support throughout. My time at UCLA was enriched by his guidance, and I gained invaluable knowledge and insights.

A special thanks to Francesco Finazzi and Sirio Legramanti, two exceptional statisticians with whom I had the pleasure of collaborating.

I would also like to extend my heartfelt gratitude to Professors Arima and Mastrantonio, whose insightful comments and suggestions have greatly enhanced both the quality and readability of this thesis.

I would also like to extend my appreciation to the friends I've met along this incredible Ph.D. journey. To my Ph.D. colleagues, Francesco, Jiefeng, Riccardo, Ludovica, and Claudia – four years ago, the path seemed insurmountable, but together we've made it. To the remarkable people who made Bicocca feel like home: Federico, Alice, Roberto, Tommaso, Alessia, and Laura. And to my office mates, who shared countless days with me: Luca, for our shared love of football; Alessandro, for your candidness; Lorenzo, for not letting me get lost in Los Angeles; and Chiara, for the ever enlightening conversations and reflections.

A huge thanks goes to the y-SIS board, a true bolt from the blue in my life: Filippo, for your dedication; Marco, for the laughs; Veronica, for understanding my quirks; and Giorgia, for your immense heart.

To my friends who have always been there, my second family in Milan, thank you for your endless support and the energy you bring to my life. And to my friends from Verona, thank you for reminding me of my roots.

To my wonderful family, my deepest gratitude. Thank you for shaping me, for your faith in me, and for supporting me unconditionally. Thank you, Dad, Mum, Giulia, Antonio, Michele, and Valentina. And to my beloved nieces and nephews, Sofia, Giacomo, Anna and Aurora, thank you for bringing me joy and lighthearted moments.

Finally, my biggest thank-you goes to Nene, my life partner in crime. Your belief and support in me and the strength you share mean the world.

Each of you has contributed to this work. Thank you all.

Contents

List of Figures

List of Tables

Introduction

Public health and environmental issues are deeply interconnected. Factors such as air pollution, climate change, and natural disasters significantly contribute to the global burden of disease. These challenges affect millions of lives each year, highlighting the need to understand the underlying data for policymaking and mitigation strategies.

Air pollution is one of the most pressing environmental health hazards globally. According to the World Health Organization (WHO), approximately 7 million premature deaths occur annually due to exposure to indoor and outdoor air pollution (World Health Assembly, 2015). Key sources include vehicular emissions, industrial activities, agricultural practices, wildfires, and the burning of solid fuels. The European Environment Agency (EEA) estimates that over 238,000 premature deaths occur annually in the European Union due to air pollution, primarily from respiratory and cardiovascular diseases (European Environment Agency, 2022). In the United States, despite a long-term downward trend in emissions, around 6,343 million metric tons of carbon dioxide equivalents were emitted in 2022 (U.S. Environmental Protection Agency, 2024). Regions with high industrial activity and vehicular traffic continue to suffer from poor air quality, resulting in severe health outcomes such as lung cancer, asthma, chronic obstructive pulmonary disease, and cardiovascular conditions.

Natural disasters, particularly earthquakes, have become more frequent in recent decades. Between 1970 and 2000, medium- and large-scale disasters averaged around 90-100 reports per year, escalating to 350-500 reports annually between 2001 and 2020 (United Nations Office for Disaster Risk Reduction, 2024). While disaster peaks were more common from 2000 to 2009, the overall frequency remains high. This increase highlights the urgent need for real-time analysis and risk mitigation strategies to minimize the impact of such events.

In the face of these growing challenges, statistical modeling has become indispensable in public health and environmental risk assessment (Brookmeyer and Stroup, 2004; Waller and Gotway, 2004; Covello and Merkhoher, 1993; Suter II, 2016). Data-driven decision-making increasingly relies on models that can manage uncertainty, capture complex relationships, and handle heterogeneous data sources. Traditional statistical methods, however, often struggle with the high-dimensional, noisy, and sparse datasets typical in these fields. Bayesian modeling offers a powerful alternative by incorporating prior knowledge, accounting for uncertainty (Lindley, 2013), and providing robust predictive frameworks (for a comprehensive overview of Bayesian analysis, see Gelman et al., 2013). These methods offer valuable insights into the complex relationships between environmental factors and health outcomes, enhancing the effectiveness of policy and intervention strategies.

Bayesian approaches are particularly advantageous in public health (Diggle and Giorgi, 2019) and environmental applications (Piegorsch and Bailer, 2005) because they allow for the formal inclusion of expert knowledge, historical data, and other external information. This is especially important

in situations where data is limited or unreliable, as often occurs in early-stage disease outbreaks or real-time natural disaster monitoring. The flexibility of Bayesian models makes them well-suited for capturing complex dependencies between variables, such as those present in disease transmission, environmental pollution, and seismic activity.

This thesis comprises three interrelated research projects, each showcasing the strength and versatility of Bayesian methods in addressing critical challenges across public health and environmental risk management. These projects, i.e., Bayesian clustering methods for environmental data, spatial disease mapping, and earthquake parameter estimation using smartphone accelerometer data—are strongly connected by their shared reliance on advanced Bayesian techniques. Together, they highlight the unifying power of these methods to extract meaningful insights from complex and noisy data, providing solutions in high-stakes contexts.

Each application addresses phenomena that are inherently spatial, temporal, or both, requiring methods capable of accounting for dependencies and uncertainties in the data. In Bayesian clustering for environmental data, these methods are used to identify patterns and groupings in air quality metrics, helping to understand the underlying drivers of pollution and their spatial distribution. Similarly, spatial disease mapping employs Bayesian frameworks to quantify and visualize health disparities, identify areas of heightened risk, and guide interventions, often relying on shared principles of spatial correlation and smoothing.

The connection extends to earthquake parameter estimation, where Bayesian methods enable the integration of sparse, noisy, and irregular accelerometer data to detect and characterize seismic events. This task, like the others, depends on the ability of Bayesian models to incorporate prior knowledge, handle uncertainty, and balance complexity with interpretability.

By employing Bayesian techniques across these diverse fields, this thesis highlights their adaptability and power to model complex systems, account for uncertainty, and produce actionable insights. Furthermore, it underscores a unifying theme: the capacity of Bayesian methods to address challenges in understanding, predicting, and mitigating risks, whether they arise from environmental pollution, public health threats, or natural disasters. Together, these projects push the boundaries of Bayesian applications, demonstrating their relevance to a wide range of societal challenges.

Bayesian models stand out because of their ability to coherently model uncertainty, a key limitation in frequentist approaches. They allow for the integration of prior knowledge into the analysis (Bayarri and Berger, 2004), making them highly adaptable to situations with sparse, uncertain, or incomplete data. This is particularly relevant in public health and environmental contexts, where datasets often suffer from these limitations (Chiolero et al., 2023; Ahkola et al., 2024). Bayesian techniques enable the combination of available data with expert knowledge, historical trends, or external information, leading to more reliable estimates and predictions.

For example, Bayesian models are particularly effective for analyzing environmental data, which is often high-dimensional, characterized by non-linear relationships, and involves unknown clustering patterns. In public health, spatial models (Banerjee et al., 2014; Cressie and Wikle, 2015) benefit from Bayesian nonparametric methods (Hjort et al., 2010; Ghosal and van der Vaart, 2017), which allow for modeling complex disease spread without imposing rigid assumptions. Furthermore, Bayesian survival models (Ibrahim et al., 2005; Klein and Moeschberger, 2006) are continuously updated with new data, which is crucial in dynamic, real-time scenarios such as earthquake monitoring.

The three projects in this thesis highlight different strengths of Bayesian methods: clustering (Wade, 2023), spatial modeling (Banerjee et al., 2014), and survival analysis (Ibrahim et al., 2005). The first project (Chapter 1) provides a comprehensive review of Bayesian nonparametric clustering methods, focusing on their application to spatio-temporal environmental data. It begins by presenting the most common models for point-referenced spatio-temporal data, followed by a synthesis of recent advancements in Bayesian clustering, with particular emphasis on nonparametric models for spatio-temporal data. These methods are demonstrated through their application to an air quality dataset, and the chapter concludes by discussing potential directions for future research. This chapter extends the work presented in Aiello, Legramanti and Paci (2024).

The second project (Chapter 2) illustrates the model developed in Aiello and Banerjee (2023). The work introduces a Bayesian nonparametric model for spatial disease mapping, a crucial tool for public health authorities in tracking disease outbreaks and planning interventions (Rao, 2023). Traditional models often assume static spatial relationships, but disease spread is influenced by various factors such as environmental conditions and socioeconomic status (see Lawson, 2018). This model employs a graphical approach (Lauritzen, 1996) to capture complex, disease-varying spatial dependencies, improving the flexibility and accuracy of disease mapping. Covariates such as population age, poverty, and smoking rates are incorporated to enhance interpretability, and uncertainty in disease counts is explicitly modeled, which is crucial in areas prone to under-reporting.

The third project (Chapter 3) illustrates the approach proposed by Aiello, Argiento, Finazzi and Paci (2023). The work develops a Bayesian survival model to estimate earthquake parameters using smartphone accelerometer data. Earthquake modeling demands real-time analysis of parameters such as location, depth, and origin time. While traditional seismic networks provide accurate data, they are limited by geographic coverage and costs (Given et al., 2014). The rise of smartphones with built-in accelerometers offers a new source of real-time crowd-sourced data (Finazzi, 2016; Kong et al., 2016). By framing this as a survival modeling problem, the Bayesian approach can incorporate prior knowledge and account for the noisy and variable nature of smartphone data. This method enhances the accuracy and timeliness of earthquake parameter estimation, representing a novel application of survival modeling in seismology.

In conclusion, this thesis demonstrates the wide-ranging applicability of Bayesian methods to diverse challenges in public health and environmental risk assessment. The three projects collectively showcase the adaptability and power of Bayesian techniques, whether through clustering, spatial modeling, or survival analysis. By leveraging the flexibility of Bayesian models, this work opens up new possibilities for research and practical applications in these critical areas.

Chapter 1

Bayesian nonparametric clustering for spatio-temporal data

In the last three decades, the increasing availability of datasets indexed by both space and time has driven the development of stochastic models that account for both spatial and temporal dependencies. From a methodological perspective, analyzing such spatio-temporal data requires accounting for spatial correlation, temporal correlation, and how space and time interact. For instance, in air pollution studies, we are not only concerned with the spatial distribution of a pollutant but also with how this distribution changes over time.

Spatial data are conventionally classified into three types: point-referenced (geostatistical) data, which consist of measurements taken at specific locations (e.g., air quality levels at different monitoring stations); areal (lattice) data where measurements are aggregated over defined spatial regions (e.g., disease rates by administrative areas); and point-pattern data, which refer to the spatial occurrence of events (e.g., the locations of earthquakes or disease outbreaks). For a comprehensive overview of spatial data, refer to Banerjee et al. (2014) and Cressie and Wikle (2015). A similar distinction can be made for the temporal scale, where time can be treated either as continuous (over $\mathbb{R}^+$ or a subinterval) or discrete (e.g., hourly, daily, etc.). In the continuous case, measurements are considered to occur at every moment. In the discrete case, we must determine whether each measurement represents an average over a time interval or whether it corresponds to a single measurement, such as a count within a specific time interval, analogous to areal unit measurements in spatial data.

This work focuses on models for point-referenced data recorded at specific times. In this context, each observation corresponds to specific coordinates (e.g., latitude and longitude) and time points, and the data represents a sample observed at fixed times on a continuous spatial domain. Denote with $D \subseteq \mathbb{R}^2$ such a continuous domain, and let $\mathbf{s} \in D$ be a point within it. We denote with $\{Y_t(\mathbf{s}) : \mathbf{s} \in D;\ t = 1, \dots, T\}$ a spatio-temporal process that can be seen as a process generating a time series indexed by $t = 1, \dots, T$ at each location $\mathbf{s} \in D$. For other ways of defining a spatio-temporal process, see Banerjee et al. (2014). Measurements at each location can be continuous, binary, or counts, such as pollution levels, the presence/absence of a species, or its abundance at each location.

A customary model for continuous data is

$$y_t(\mathbf{s}) = \mathbf{x}_t(\mathbf{s})^\top \boldsymbol{\beta}_t(\mathbf{s}) + w_t(\mathbf{s}) + \epsilon_t(\mathbf{s}), \tag{1.1}$$

where $\mathbf{x}_t(\mathbf{s})^\top \boldsymbol{\beta}_t(\mathbf{s})$ represents the mean structure depending on spatio-temporal covariates, $w_t(\mathbf{s})$ is a mean-zero spatio-temporal process capturing the spatio-temporal dependence not explained by the covariates, and $\epsilon_t(\mathbf{s})$ is an error term – typically a Gaussian white noise process – accounting for the residual spatio-temporal variability in the data.

In principle, the regression coefficient $\boldsymbol{\beta}_t(\mathbf{s})$ may depend on both space and time. However, it can be simplified to $\boldsymbol{\beta}_t(\mathbf{s}) = \boldsymbol{\beta}$ (Arima et al., 2015; Hefley et al., 2017; Laurini, 2019; Wang et al., 2024), or at least to $\boldsymbol{\beta}_t(\mathbf{s}) = \boldsymbol{\beta}_t$ (Kottas et al., 2008; Berrocal, 2016; Lee et al., 2021) or $\boldsymbol{\beta}_t(\mathbf{s}) = \boldsymbol{\beta}(\mathbf{s})$ (Torabi, 2014; Li et al., 2016; Peluso et al., 2020). More details on the representation of $\boldsymbol{\beta}_t(\mathbf{s})$ can be found in Banerjee et al. (2014).

As for the spatio-temporal process $w_t(\mathbf{s})$, different modeling choices are available in the literature, including an additive form $w_t(\mathbf{s}) = \alpha_t + \psi(\mathbf{s})$ (Waller et al., 1997; Knorr-Held and Besag, 1998; Cameletti et al., 2011), a multiplicative form $w_t(\mathbf{s}) = \alpha_t \psi(\mathbf{s})$ (Paci et al., 2013) or a nested form with spatial effects nested within time, i.e., independent spatial processes over time (Arima et al., 2012). In all cases, a popular approach to model $\psi(\mathbf{s})$ is via a Gaussian process (Rasmussen and Williams, 2005) equipped with a spatial covariance structure. Another popular choice for $w_t(\mathbf{s})$, is to assume a dynamic evolution of the spatial process over time, namely

$$w_t(\mathbf{s}) = \phi w_{t-1}(\mathbf{s}) + \nu_t(\mathbf{s}), \tag{1.2}$$

where the $\nu_t(\mathbf{s})$ are independent spatial processes over time. Altogether, (1.1) and (1.2) constitute the well known dynamic linear model (Harrison and Stevens, 1976; West and Harrison, 2006), often referred to as the state-space model in the time-series literature. This approach allows for modeling temporal components such as trends, seasonal effects, and autoregressive behavior. For instance, Stroud et al. (2001) applied this model to two large-scale environmental datasets, specifically tropical rainfall levels and sea surface temperatures in the Atlantic Ocean.

This chapter is structured as follows: Section 1.1 provides an overview of Bayesian clustering methods and their application to spatio-temporal data, with a special focus on mixture models under a nonparametric approach. Section 1.2 explores the Exchangeable Product Partition Model and the spatial Product Partition Model. In Section 1.3, we illustrate the Bayesian clustering methods using air quality data. Section 1.4 offers some concluding remarks.

1.1 Bayesian cluster analysis

Cluster analysis refers to grouping a set of observations so that those belonging to the same cluster are more similar to each other than those belonging to different clusters. Clustering is widely used across various fields, including environmental sciences. Recent applications to environmental data include air quality monitoring (Cheam et al., 2017), identifying minefields or seismic faults (Dasgupta and Raftery, 1998), food-searching behaviour of sandhoppers (Ranalli and Maruotti, 2020), coastal rainfall patterns (Paton and McNicholas, 2020), and fish biodiversity (Vanhatalo et al., 2021), just

to name a few. In many applications, the focus is on the cluster allocation of the observations, as well as the patterns within each cluster.

One possible classification of clustering methods is into heuristic (or algorithmic) and model-based (or probabilistic) strategies. Heuristic methods include, e.g., the popular k-means algorithm (Macqueen, 1967), which partitions the data into k clusters by iteratively assigning each data point to the cluster with the closest mean. For a detailed discussion on this and other distance-based clustering algorithms, see Sarang (2023). Heuristic strategies have the advantage of being fast and able to tackle big data. On the other hand, they return a single partition without quantifying the uncertainty in the clustering structure, that is, the uncertainty about any observation's group membership. In contrast, model-based approaches (Fraley and Raftery, 2002) allow quantifying the uncertainty associated with the clustering structure by postulating a statistical model for the population from which the data are sampled. This also brings the advantage of answering questions such as how many clusters are present, and how to detect and treat outliers. See Bouveyron et al. (2019) for a comprehensive overview of model-based clustering, and Wade (2023) for a review of Bayesian approaches to it.

To be more specific, let us introduce some of the notation that will be used in this paper. First, note that grouping n objects into K clusters is equivalent to specifying the allocation vector $\mathbf{c} = (c_1, \dots, c_n)^\top$, where each $c_i \in \{1, \dots, K\}$ and $c_i = k$ means that the i-th object is assigned to the k-th cluster. Such a vector induces a partition of the n considered objects (or equivalently of the first n integers), denoted with $\rho_n := \{S_1, \dots, S_K\}$, where $S_k = \{i : c_i = k\}$. Note that, while each allocation vector induces a unique partition, the viceversa does not hold. In fact, the same partition can be induced by different allocation vectors. As a simple example, let n be even and consider the case where the first half of the objects are assigned to cluster 1 and the second half to cluster 2. This obviously induces the same partition as the allocation assigning the first half of the objects to cluster 2, and the second half to cluster 1. This phenomenon – known as *label switching* – must be accounted for in many clustering methods.

While the typical output of clustering algorithms is a single value of the allocation vector $\mathbf{c}$, i.e., a single partition ρ_n, the actual object of interest in cluster analysis is in evaluating the uncertainty of the partition ρ_n. Model-based clustering enables inference on ρ_n by postulating a model for the observable data $\mathbf{y} = \{y_1, \dots, y_n\}$ that depends on the partition ρ_n. We will denote such a model with $p(\mathbf{y} \mid \rho_n)$. In particular, Bayesian approaches allow to quantify uncertainty on ρ_n *a posteriori* – i.e., after observing the data – by first posing a prior distribution on ρ_n, and then computing its posterior distribution through the Bayes theorem $p(\rho_n \mid \mathbf{y}) \propto p(\mathbf{y} \mid \rho_n)\ p(\rho_n)$. The prior $p(\rho_n)$ encodes what the analyst knows or expects about the partition (e.g., about the cardinality of each cluster) before observing the data $\mathbf{y}$. The prior for ρ_n is customary assigned via an Exchangeable Partition Probability Function (EPPF). For a comprehensive review of possible models for ρ_n, we refer to Quintana (2006). However, before going into details about the prior, let us focus on the model $p(\mathbf{y} \mid \rho_n)$.

Typically, instead of directly assigning the model as $p(\mathbf{y} \mid \rho_n)$, one first assigns a mixture model $p(\mathbf{y} \mid P)$, where P is a random measure, almost surely discrete. Thus, given $\mathbf{c}$, the law of $\mathbf{y}$ conditioned on ρ_n is obtained by integrating out, i.e., marginalizing, P from the law of $\mathbf{y}$, with P conditioned on $\mathbf{c}$. Further details on this procedure will follow. A popular class of $p(\mathbf{y} \mid \rho_n)$ is represented by mixture models, where observations are assumed to belong to one of the possible groups, which could even be infinitely many in the big data limit. Each group is suitably modeled by a density,

referred to as a component of the mixture, which is weighted by the relative frequency of the group in the population. For instance, in a (multivariate) Gaussian mixture model, each component is a (multivariate) Gaussian density. For a review of mixture models see Bouveyron et al. (2019), Grün (2019), McLachlan et al. (2019) and Frühwirth-Schnatter et al. (2019).

In finite mixture modeling, each observation is assumed to come from one of the $M < \infty$ mixture components. Namely, the model is given by

$$f(y \mid P) = \int_{\Theta} f(y \mid \theta) P(d\theta) = \sum_{m=1}^{M} \pi_m f(y \mid \gamma_m), \tag{1.3}$$

where $\{f(y \mid \theta) : \theta \in \Theta\}$ is a parametric family of densities on $\mathcal{Y}$, referred to as the *mixture kernel*, while P is an a.s.-discrete measure on Θ, supported on $\{\gamma_m \in \Theta : m = 1, \dots, M\}$, referred to as the *mixing measure*. The choice of the kernel depends on the nature of the data and on the analysis goals. For continuous data, while Gaussian kernels stand out as the most popular choice, skewness or outlier robustness can be achieved using, e.g., multivariate skew-normals or Student's t-distributions (Frühwirth-Schnatter and Pyne, 2010; Lee and McLachlan, 2014), shifted asymmetric Laplace distributions (Franczak et al., 2013), or normal-inverse Gaussian distributions (O'Hagan et al., 2016). For categorical data, one may employ latent class models, which rely on mixtures of Bernoulli or multinomial distributions (Goodman, 1974; Argiento et al., 2024). For count data, mixtures of Poisson (Karlis and Xekalaki, 2005; Krnjajić et al., 2008), negative binomial (Liu et al., 2024), or zero-inflated Poisson and negative-binomial distributions are useful for handling sparse counts (Wu and Luo, 2022).

In equation (1.3), each component is weighted by π_m, the relative frequency of the m-th group in the population, for $m = 1, \dots, M$. Note that M denotes the total number of components at a population level, i.e., the maximum number of possible clusters as $n \to \infty$. In contrast, $K \leq M$ refers to the number of non-empty clusters, i.e., those clusters that contain at least one of the $n < \infty$ observations (see, e.g., Argiento and De Iorio, 2022). Obviously, K is also upper bounded by n, while M can be larger than n and even infinite, as we will see.

In Bayesian model-based clustering, two primary approaches are prevalent: (i) fixing M (either finite or infinite), (ii) treating M as a random parameter and inferring it from the data; see, e.g., Wade (2023) and Grazian (2023). In the second scenario, Nobile (2004) emphasizes the distinction between K and M, observing that the posterior distribution of M might concentrate on higher values with respect to the posterior of K. In contrast, when M is set to infinity, K is the parameter of interest in posterior inference. However, even when M is fixed and finite, we still need to estimate K (Rousseau and Mengersen, 2011).

As customary in mixture modeling, a hierarchical parametrization can be employed to facilitate the computation. To this end, we assume that the allocation variables c_i are conditionally independent given $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)$, where $\pi_m = \Pr(c_i = m \mid \boldsymbol{\pi})$, and that $c_i \mid \boldsymbol{\pi}, M \overset{iid}{\sim} \text{Multinomial}_M(\pi_1, \dots, \pi_M)$. The complete specification becomes

$$
\begin{aligned}
y_i \mid c_i, \gamma &\overset{ind}{\sim} f(y \mid \gamma_{c_i}), && i = 1, \dots, n, \\
c_i \mid \boldsymbol{\pi}, M &\sim \text{Multinomial}_M(1, \pi_1, \dots, \pi_M), && i = 1, \dots, n, \\
\boldsymbol{\pi} \mid M &\sim \text{Dirichlet}_M(\alpha/M, \dots, \alpha/M), \\
M &\sim q_M, \\
\gamma_m &\overset{iid}{\sim} P_0, && m = 1, \dots, M,
\end{aligned}
\tag{1.4}
$$

where $f(y \mid \gamma)$ is a parametric density on $\mathcal{Y}$, depending on a vector of parameters γ. For each component $m = 1, \dots, M$, the corresponding vector of parameters $\gamma_m \in \Theta \subset \mathbb{R}^d$ is assigned a non-atomic prior P_0.

For computational convenience, a common choice for P_0 is a conjugate prior to the kernel $f(y \mid \gamma)$. The hyperparameters of the base measure P_0 can be specified based on prior knowledge, set empirically, or inferred through additional hyperpriors. Alternatively, data-dependent or non-informative priors can be used for P_0 (Rousseau et al., 2019). To encourage well-separated components, the independence assumption on the atoms γ_m can be relaxed by employing repulsive priors (Petralia et al., 2012; Xie and Xu, 2020; Beraha et al., 2022), determinantal point processes (Xu et al., 2016), or non-local priors (Fúquene et al., 2019).

Conditionally on M, the vector of weights $\boldsymbol{\pi}$, which represents the probability of belonging to each mixture component, is given a Dirichlet prior. Typically, all the M parameters of the Dirichlet prior are set to the same value, say α/M, thus obtaining a symmetric Dirichlet distribution. A small value of α/M encourages sparsity in the weights. In other words, as $\alpha/M \to 0$, all the prior mass is concentrated at the vertices of the simplex, resulting in all weight being assigned to a single component.

Beyond the Dirichlet distribution, other distributions may also be considered, such as the generalized Dirichlet distribution (Connor and Mosimann, 1969), non-informative Jeffreys priors (Bernardo and Girón, 1988), the Pitman–Yor (PY) process (Perman et al., 1992; Pitman and Yor, 1997), the multinomial PY process (Lijoi et al., 2020), the Gibbs-type priors (De Blasi et al., 2015), Bertoin, Fujita, Roynette, and Yor (BFRY) priors (Lee et al., 2016), or normalized jumps of a finite point process (Argiento and De Iorio, 2022). A recent work by Ascari et al. (2024) explores additional distributions, presenting a novel multivariate regression model for constrained responses, which is grounded in the extended flexible Dirichlet distribution introduced by Ongaro et al. (2020). Finally, the number of components is given a prior q_M whose typical choices include a discrete uniform on some finite space, a Negative Binomial (Grazian et al., 2020), a Poisson (Stephens, 2000; Nobile, 2004) or a Uniform distribution over $\{1, \dots, K_{max}\}$ (Miller and Harrison, 2018).

The mixture model in (1.4) induces a clustering among the observations. Namely, when an observation is sampled from a component, it will be assigned to the corresponding cluster. Naturally, when sampling $n < \infty$ observations, not all of the M components will be necessarily sampled from. This is obvious when $M = \infty$, but is absolutely possible even when M is finite, especially if large. As observed before $p(\mathbf{y} \mid \rho_n) = p(\mathbf{y} \mid \mathbf{c})$ is derived by integrating out the parameter $P \equiv (\boldsymbol{\gamma}, M, \boldsymbol{\pi})$ from the joint distribution of model (1.4), see Argiento and De Iorio (2022) for mathematical details. We also note that the marginalization just described is achieved since model (1.4) belongs to the wider class of species sampling mixture models Pitman (1996), which are largely adopted in Bayesian

nonparametric frameworks; see, among others, Ishwaran and James (2003); Miller and Harrison (2018); Argiento and De Iorio (2022).

1.1.1 Posterior inference for Bayesian clustering

Over the years, computational methods for finite mixtures, mainly based on Markov chain Monte Carlo (MCMC) algorithms, have significantly advanced. In particular, posterior inference for finite mixture models with random M needs a transdimensional algorithm that accommodates jumps between parameter spaces of different dimensions (according to the number of mixing components). An early breakthrough was made by Green (1995); Richardson and Green (1997), who introduced the Reversible Jump MCMC algorithm for univariate Gaussian mixtures. The algorithm is quite popular, but it requires the design of good reversible jump moves, posing challenges in applications, particularly with high-dimensional parameter spaces.

On the other hand, posterior inference for infinite mixture models has evolved through key algorithmic developments, which can be classified into two groups: (i) *conditional algorithms*, providing full Bayesian inference on both mixing parameters and the clustering structure (Papaspiliopoulos and Roberts, 2008; Kalli et al., 2011); and (ii) *marginal algorithms*, which simplify computation by integrating out the mixture parameters, focusing solely on clustering (MacEachern and Müller, 1998). Early progress was marked by the introduction of the Pólya urn Gibbs sampler (Escobar, 1994; MacEachern, 1994; Escobar and West, 1995; MacEachern, 1998), followed by the application of the stick-breaking representation of the Dirichlet process by Ishwaran and James (2001), which enabled posterior inference via Gibbs sampling.

Recently, exploiting the link between finite and infinite mixture models, algorithms developed for Bayesian nonparametric models have been adapted to finite mixture models. Examples are the Chinese restaurant process sampler in Miller and Harrison (2018), the telescoping sampling developed by Frühwirth-Schnatter et al. (2021), and the two augmented Gibbs samplers proposed by Argiento and De Iorio (2022). These advances have enhanced the computational efficiency and applicability of finite mixture models in practice.

Bayesian mixture models deliver a posterior distribution over the entire space of partitions, revealing the uncertainty of the clustering structure given the data. All the MCMC sampling schemes mentioned above provide an approximation of such posterior. However, a natural issue in this setting is how to summarize the posterior distribution, e.g., how to provide an appropriate posterior point estimate of the clustering structure and of the cluster-specific parameters based on the MCMC output. The problem is twofold; first, to estimate the clustering and the group-specific parameters, the mixture model must be identified to avoid label switching (Jasra et al., 2005). Second, the high dimension of the partition space and the fact that many of these partitions are usually quite similar (e.g., differing only in a few data points) make the posterior spread out across a large number of partitions.

A possible solution to get a point estimate of the partition from its posterior samples is to assign data points to the cluster with maximum posterior probability. More generally, from a decision-theoretic perspective, the goal is to find the partition minimizing the expected loss under the posterior. In this framework, the Maximum a Posteriori (MAP) described above is optimal under a binary loss, but may not fully explore the partition space or account for partial similarities between partitions.

An alternative approach is the Binder loss (Binder, 1978), which imposes a penalty when pairs of observations that should be grouped together are placed in different clusters, and when pairs that should be separated are incorrectly clustered together. Wade and Ghahramani (2018) instead propose using the Variation of Information (VI) loss (Meilă, 2007), which compares the information shared between two partitions. Although the VI loss is more computationally demanding and is sensitive to the cluster initialization, it provides a more nuanced comparison. To obtain a posterior point estimate of the partition, the posterior distribution of the cluster assignments can be derived using a greedy, stochastic search approach Dahl et al. (2022). This method minimizes a chosen loss function, such as Binder loss or VI, as previously discussed.

1.1.2 Clustering spatio-temporal data

This work focuses on Bayesian nonparametric clustering for spatio-temporal data, with an emphasis on environmental applications. The class of finite mixture models discussed in Section 1.1 appears particularly useful for clustering complex objects like spatio-temporal series.

In the Bayesian nonparametric context, one prominent method is the spatial Dirichlet process (Gelfand et al., 2005), which poses a distribution on the atoms that takes into account the spatial dependence. This model was extended by Gelfand et al. (2007) and Duan et al. (2007), whose main contribution was incorporating spatially dependent mixture weights. Another approach is the spatial stick-breaking process (Reich and Fuentes, 2007; Rodrıguez and Dunson, 2011), which enables spatially varying allocation probabilities without requiring the presence of replications. Additionally, the Dirichlet labeling process (Nguyen and Gelfand, 2011) has been proposed as a flexible model for spatial clustering. While these methods are highly appealing, their performance can be limited when covariates are available to aid in cluster identification.

A Bayesian semiparametric mixture model within a state-space framework was proposed by Nieto-Barajas and Contreras-Cristán (2014) for clustering of time series data. Several research challenges have emerged from finite mixture models for time series, including detecting dynamic changes in spatio-temporal patterns, identifying anomalies, monitoring environmental changes, real-time spatial process control, and recognizing complex geospatial patterns over time.

Notable examples of this approach include Fernández and Green (2002), who developed a spatial mixture model for areal data with a variable number of mixing components and spatially dependent mixing weights. Frühwirth-Schnatter and Kaufmann (2008) proposed a clustering method using finite mixtures of dynamic regression models, allowing for information pooling within clusters. Similarly, Viroli (2011) applied a finite mixture model to analyze three-way data structures, including spatio-temporal features. Neelon et al. (2014) introduced a finite mixture model with spatial random effects for each component, specifically designed to analyze multivariate areal-referenced data. Additionally, Hossain et al. (2014) employed a space-time mixture of Poisson regression models to tackle relabeling algorithms and model selection challenges, while Paci and Finazzi (2018) proposed a dynamic space-time mixture model to identify level-based clusters in spatio-temporal data and track their temporal evolution.

When clustering spatio-temporal data, the focus is typically on the spatio-temporal component $w_t(\mathbf{s})$ in equation (1.1). It can be seen either as n time series $\mathbf{w}(\mathbf{s}_i) = (w_1(\mathbf{s}_i), \dots, w_T(\mathbf{s}_i))^\top$ for $i = 1, \dots, n$, one for each location, or as T spatial surfaces $\mathbf{w}_t = (w_t(\mathbf{s}_1), \dots, w_t(\mathbf{s}_n))^\top$, one for each time

point. A first option would be to cluster together locations $\mathbf{s}_i$ and $\mathbf{s}_j$ if the corresponding time series take the same values over the whole time interval, i.e., $w_t(\mathbf{s}_i) = w_t(\mathbf{s}_j)$ for each $t = 1, \dots, T$; see, e.g., Berrocal (2016) and Nieto-Barajas and Contreras-Cristán (2014). Alternatively, one may want to cluster together two time points t and t' if the corresponding surfaces coincide at all considered locations, i.e. $w_t(\mathbf{s}_i) = w_{t'}(\mathbf{s}_i)$ for each $i = 1, \dots, n$; see, e.g.,Gelfand et al. (2005) and Page et al. (2022). A third option consists of clustering both over time and space meaning that the pairs $(\mathbf{s}_i, t)$ and $(\mathbf{s}_j, t')$ are assigned to the same cluster if $w_t(\mathbf{s}_i) = w_{t'}(\mathbf{s}_j)$; see, e.g., Duan et al. (2007), Nguyen and Gelfand (2011), and Mastrantonio et al. (2022).

Finally, clustering can be based on the parameters governing the spatio-temporal random effect $w_t(\mathbf{s})$; see, e.g., Mastrantonio et al. (2019), Bucci et al. (2022) and Musau et al. (2022). For instance, if we assume that $w_t(\mathbf{s}_i) = \phi_i w_{t-1}(\mathbf{s}_i) + \nu_t(\mathbf{s}_i)$ with $\nu_t(\mathbf{s}_i) \sim N(0, \tau_i^2)$ as in equation (1.2), we can cluster the locations $\mathbf{s}_i$ based on the parameters $\gamma_i = (\phi_i, \tau_i^2)$. Indeed, this is the approach that we follow next.

1.2 Product Partition Models

1.2.1 Exchangeable Product Partition Models

Our Bayesian clustering framework aims at making inference on the partition ρ_n given the data. In this context, a key role is played by the prior on ρ_n, for which a common choice is represented by Gibbs-type priors. Such priors can be seen as generalizations of the popular Dirichlet process (De Blasi et al., 2015), and offer analytical and computational tractability (Lijoi et al., 2007a). In particular, such processes can be seen as "seating mechanisms", similar to the Chinese restaurant metaphor for the Dirichlet Process. For examples of tractable Gibbs-type priors, see De Blasi et al. (2013), Lijoi et al. (2007a,b), and Miller and Harrison (2018).

As shown in Lijoi et al. (2007b), the Gibbs-type priors are a sub-class of Product Partition Models (PPM, Hartigan, 1990). In particular, they coincide with exchangeable PPMs. The latter are characterized by the following general structure

$$p(\rho_n) \propto \prod_{k=1}^{K} C(S_k), \tag{1.5}$$

which factorizes over clusters and incorporates the cohesion function $C(\cdot)$, which gauges the degree of "closeness" among the elements in each cluster. Each Gibbs type prior is then characterized by a specific choice for the cohesion function. There are some options for defining the cohesion function. For example, the Dirichlet Process is induced by the cohesion function $C(S_k) = \alpha \times \Gamma(|S_k|) = \alpha \times (|S_k| - 1)!$ with $|S_k|$ denoting the cardinality of S_k and α the total mass parameter of the process (Quintana and Iglesias, 2003). For details on the cohesion functions of other Gibbs-type priors such as the Pitman-Yor process, see Gnedin and Pitman (2006), Lijoi et al. (2008), Gnedin (2010) and De Blasi et al. (2015).

Gibbs sampling for PPMs is a widely used approach for clustering data into distinct groups based on an underlying similarity structure. The process begins with the initialization of the model parameters, which may include variables such as variance components, autoregressive parameters

(if time-series data is involved), and parameters related to the cohesion function that defines how clusters are formed. These initial values are typically chosen based on prior distributions or some reasonable initial guesses.

The core of Gibbs sampling in the PPM context involves updating the partition structure, which defines how observations are grouped into clusters. At each iteration, the algorithm samples the group membership of each observation from the posterior distribution of possible clusterings. This probability is informed by the likelihood of the data, the cohesion between observations, and the prior imposed by the Product Partition Model. As part of this process, a co-clustering matrix is often calculated to measure how frequently observations are assigned to the same cluster.

Once the partition is updated, the next step involves updating the parameters associated with each cluster and the overall likelihood function. These parameters, which may include the mean, variance, and autoregressive coefficients for each cluster, are sampled from their full conditional distributions. The sampling is performed based on the data assigned to each cluster, reflecting the local information captured by the partition. Additionally, hyperparameters, such as those governing the overall variance or the concentration parameter of the Dirichlet Process (if used), are updated based on the current partition and data. These updates ensure that the model parameters reflect the latest partition structure and adapt to the clustering patterns discovered during the iteration.

Once the sampling is complete, the Gibbs sampler provides a set of posterior samples that reflect the distribution of clustering configurations and model parameters. As mentioned in Section 1.1.1 the final clustering configuration can be selected by minimizing a posterior loss function, such as the VI, or by choosing the configuration that best fits the data according to the posterior samples. In this way, Gibbs sampling for PPM yields a flexible clustering framework that accounts for uncertainty in both the partitions and the underlying model parameters, offering a rich framework for posterior inference.

1.2.2 Spatial Product Partition Model

PPMs can be extended to directly incorporate covariate information into the clustering process. This extension, known as the PPM with covariates (PPMx), was introduced by Müller et al. (2011). PPMx is defined by augmenting the PPM with an additional factor – referred to as *similarity function* – that induces the desired dependence on the covariates; see function $g(\cdot)$ in equation (1.6) below. When the covariates are spatial locations, this model is referred to as spatial PPM (sPPM, Page and Quintana, 2016). More specifically, let $s_k^* = \{\mathbf{s}_i : \ i \in S_k\}$ and $y_k^* = \{y_i : \ i \in S_k\}$ be the covariates (i.e., the spatial coordinates) and the data for the individuals in cluster k, and denote with $g(\cdot)$ a non-negative function taking higher values on sets of covariates judged to be more similar. The model on the partition ρ_n is then conditioned on the value taken by all the locations $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$. Specifically, model (1.5) becomes

$$p\left(\rho_n \mid \mathbf{s}_1, \dots, \mathbf{s}_n\right) \propto \prod_{k=1}^{K} g(s_k^*) C\left(S_k\right). \tag{1.6}$$

In the context of the sPPM, the specification of a similarity function plays a critical role in how data points are grouped into clusters. Several alternative formulations of the similarity function are

proposed, as outlined by Müller et al. (2011) and Page and Quintana (2016). These formulations, though distinct in their construction, all aim to reflect the spatial proximity and relatedness of data points within a cluster.

Following Page and Quintana (2016), the first similarity function, $g_1(s_k^*)$, includes an indicator function for singleton clusters ($|S_k| = 1$) and a term that depends on the sum of distances $\mathcal{D}_k$ between points within a cluster and its centroid. Namely,

$$g_1(s_k^*) = \begin{cases} 1 & \text{if} \quad |S_k| = 1 \\ \frac{1}{\Gamma(\omega \mathcal{D}_k)\mathbb{1}[\mathcal{D}_k \geq 1] + \mathcal{D}_k \mathbb{1}[\mathcal{D}_k < 1]} & \text{if} \quad |S_k| > 1 \end{cases}$$

This formulation makes intuitive sense in spatial modeling since a smaller $\mathcal{D}_k$ implies that the points in the cluster are tightly packed, leading to a higher similarity. The term ω serves as a tuning parameter, giving flexibility in how sensitive the model is to spatial spread.

The second function, $g_2(s_k^*)$, is based on pairwise distances within a cluster, with a threshold parameter a. Specifically,

$$g_2(s_k^*) = \prod_{i,j \in S_k} \mathbb{1}\left[d(\mathbf{s}_i, \mathbf{s}_j) \leq a\right].$$

This approach sets a stricter condition for similarity, only considering pairs of points that are sufficiently close. It emphasizes strong spatial cohesion within clusters by discarding any relationships where points are too distant from one another.

For $g_3(s_k^*)$ and $g_4(s_k^*)$, the similarity function becomes more probabilistic, incorporating a density function $q(\mathbf{s}_i \mid \boldsymbol{\xi}_k)$, where $\boldsymbol{\xi}_k$ represents hyperparameters specific to the cluster. To be more precise,

$$\begin{aligned} g_3(s_k^*) &= \int \prod_{i \in S_k} q(\mathbf{s}_i \mid \boldsymbol{\xi}_k)\, q(\boldsymbol{\xi}_k)\, d\boldsymbol{\xi}_k, \\ g_4(s_k^*) &= \int \prod_{i \in S_k} q(\mathbf{s}_i \mid \boldsymbol{\xi}_k)\, q(\boldsymbol{\xi}_k \mid s_k^*)\, d\boldsymbol{\xi}_k. \end{aligned} \tag{1.7}$$

The use of a probability density function enables a more flexible representation of spatial clusters. By integrating over the hyperparameters, we introduce a measure of correlation within the cluster, meaning that clusters with more spatially similar points will yield higher values for $g(s_k^*)$. This approach is especially valuable in spatial modeling, as it accounts not only for the distances between points but also for the underlying spatial distribution patterns that emerge. Although these two similarity measures appear to have the same form, $g_4(s_k^*)$ differs by employing a posterior predictive conjugate model, in contrast to the prior predictive conjugate model used by $g_3(s_k^*)$.

The probabilistic framework underlying $g_3(s_k^*)$ and $g_4(s_k^*)$ has significant implications. It introduces a natural clustering behavior where data points are grouped together based on their spatial similarity, with clusters exhibiting higher internal coherence receiving greater likelihoods. This aligns well with the goals of spatial partitioning in models like the sPPM, where spatial dependencies and proximities need to be accounted for in a mathematically rigorous way. Typically the choice for $q(\mathbf{s}_i \mid \boldsymbol{\xi}_k)$ in equation (1.7) is a multivariate normal with hyperparameters $\boldsymbol{\xi}_k = (\mathbf{m}_k, V_k)$, representing the mean and covariance of the spatial locations, and a conjugate normal-inverse-Wishart distribution for $q(\boldsymbol{\xi}_k)$. The closed form of $g_3(\cdot)$, resulting from this choice, is detailed in Appendix A.1.

The similarity function leverages these parameters to finely tune the clustering process, ensuring that the spatial correlation within clusters is adequately captured.

In summary, the choice of a similarity function in the sPPM has profound implications for how spatial clusters are formed and interpreted. By carefully specifying this function, we can ensure that clusters reflect meaningful spatial patterns, accounting for both proximity and underlying spatial dependencies. This ultimately leads to a more nuanced understanding of spatial structures, which is essential in fields like environmental science and public health, where spatial relationships play a crucial role in the phenomena being studied.

1.3 Analysis of air quality data

1.3.1 Data description

Particulate matter (PM) consists of a variety of solid and liquid particles that remain suspended in the atmosphere for extended periods, enabling them to undergo diffusion and transport processes. These particles originate from both natural sources, such as soil erosion, volcanic activity, and pollen dispersal, and human activities, including industrial production, heating, and vehicular traffic. Unlike other pollutants, PM is not a single chemical substance but a complex mixture of particles with varying properties, and has significant environmental impact, contributing to climate change, soil and water contamination, and posing serious health risks to living organisms. Continuous monitoring of PM levels is therefore essential to prevent concentrations from exceeding critical thresholds that are deemed harmful to health. This is achieved by using monitoring stations distributed across the region. Previous Bayesian approaches to PM data include Sahu et al. (2006), which introduced two random effects components for rural and urban PM levels, Cameletti et al. (2011), where a comparison method was proposed to select a model for PM10 data based on goodness of fit, interpretability, and parsimony, and Hamm et al. (2015), which presented a spatially varying coefficients model, allowing the regression coefficients to vary spatially according to an estimated covariance function.

Our dataset consists of daily PM10 measurements (concentration of PM with a diameter smaller than $10\mu m$) collected from 162 monitoring stations located in Northern Italy during the period from January 1 to April 10, 2019. The dataset comes from the European Environmental Agency (EEA) and is freely available at `https://eeadmz1-downloads-webapp.azurewebsites.net/`. Northern Italy is characterized by poor air quality. For instance, Figure 1.1 shows that, during 2019, 77 out of the 162 stations in Northern Italy exceeded the daily limit established by the EEA (European Commission, 2008), i.e., 50 $\mu g/m^3$ on more than the 35 days per year. In some particularly affected areas, stations registered up to 80 days above the threshold. This underscores the unevenness of air quality across the study region.

Figure 1.2 presents the time series of the PM10 concentrations median and interquantile ranges for the study period, illustrating both the overall levels of pollution and the variability across the stations. The plot reveals notable temporal variation in PM10 concentrations, with peaks during colder months where levels often exceed 50 $\mu g/m^3$. This pattern aligns with well-known winter factors, including increased heating emissions, reduced atmospheric dispersion, and stagnant air conditions (Pietrogrande et al., 2022). In contrast, during the warmer period, PM10 concentrations are generally lower, frequently falling below 50 $\mu g/m^3$. Additionally, the time series shows substantial variability

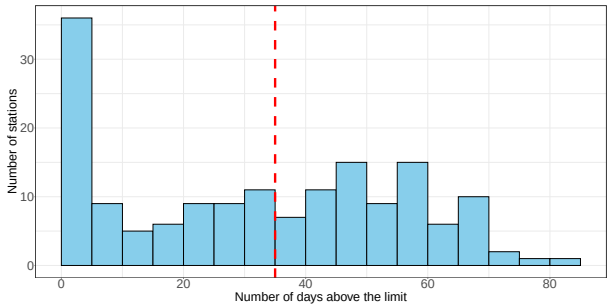


Figure 1.1: Number of stations above the daily limit in 2019.

across monitoring stations, with some stations experiencing higher peaks and deeper troughs.

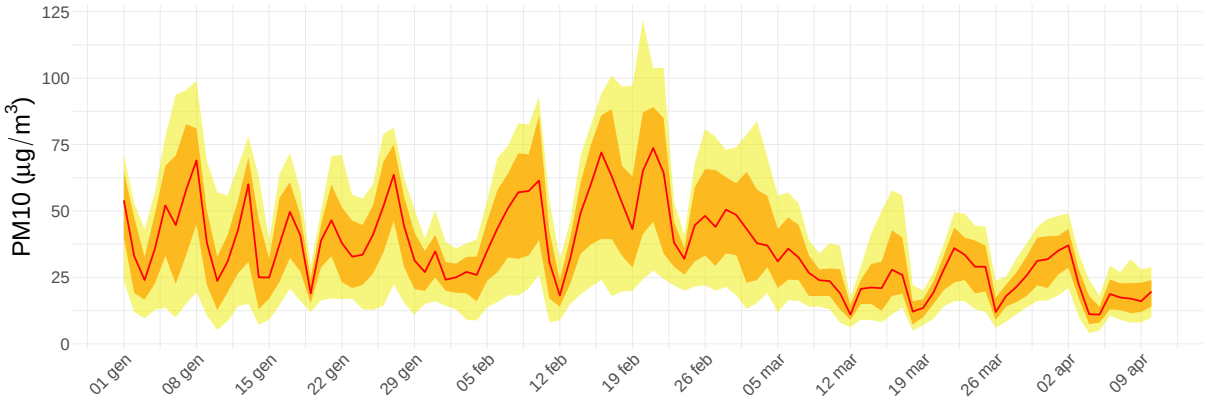


Figure 1.2: PM10 concentration station-wise median (red), 50% interquartile (orange) and 90% interquartile range (yellow) over the study period.

Figure 1.3 illustrates the spatial distribution of the mean and standard deviation of the PM10 concentration over the time period. In particular, panel (a) presents the mean PM10 concentrations recorded across different stations, showing the highest values – often exceeding 50 $\mu g/m^3$ – in urban and industrial areas like Turin, Milan, Brescia, Verona, and Venice. In contrast, lower concentrations (below 20 $\mu g/m^3$) are found in northern and coastal regions, such as Bolzano and Genoa. The southern Po Valley exhibits moderate concentrations, generally between 20 and 50 $\mu g/m^3$. Panel (b)

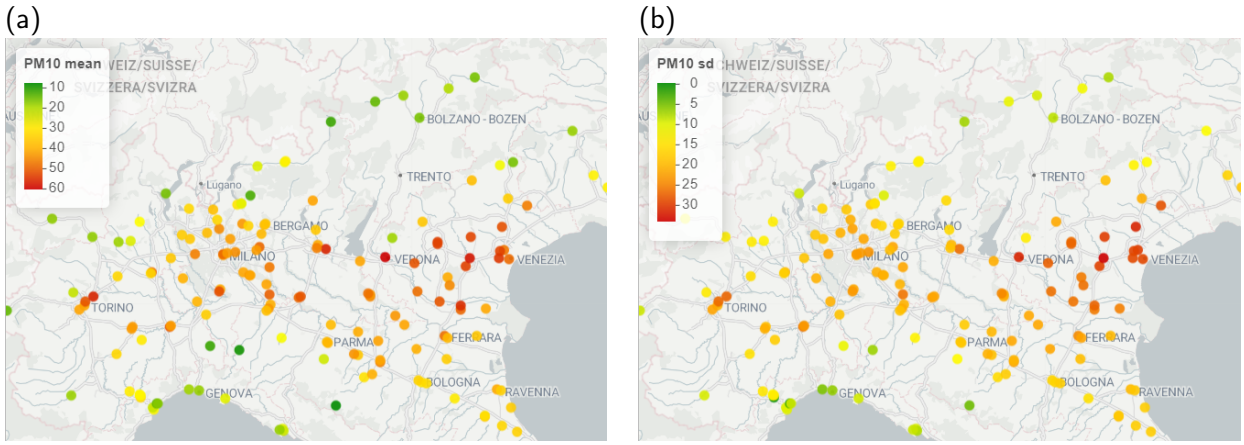


Figure 1.3: Maps with mean (a) and standard deviation (b) of PM10 concentration time series.

shows the standard deviation, highlighting the spatial pattern of the temporal variance. Regions with higher mean concentrations, particularly urban and industrial areas, also tend to experience greater temporal fluctuations, while areas with lower mean concentrations show less temporal variability.

Finally, to gain a better understanding of the data and identify which parameters may be useful for clustering, we fit an AR(1) model, independently for each time series of PM10 concentrations. In particular, the AR(1) model assumes that each day's PM10 level can be predicted based on the previous day's level, capturing short-term persistence in air quality at each monitoring station. The results are shown in Appendix A.2 and reveal a clear spatial pattern of the estimated autoregressive coefficients and residual variances, with similar spatial evidence in the mean levels. These findings support our decision to focus the clustering structure on temporal persistence and residual variability of the time series.

1.3.2 Data analysis

In order to cluster monitoring stations, we adopt a model similar to the one described in Nieto-Barajas and Contreras-Cristán (2014). Specifically, the PM10 concentration recorded at the i-th monitoring station on day t, for $i = 1, \dots, n$ and $t = 1, \dots, T$, is modeled as

$$\begin{aligned} y_t(\mathbf{s}_i) &= \mathbf{z}_t^\top \boldsymbol{\beta}_i + w_t(\mathbf{s}_i) + \epsilon_t(\mathbf{s}_i), \qquad \text{with} \quad \epsilon_t(\mathbf{s}_i) \overset{ind}{\sim} \mathcal{N}\left(0, \sigma_i^2\right), \\ w_t(\mathbf{s}_i) &= \phi_i w_{t-1}(\mathbf{s}_i) + \nu_t(\mathbf{s}_i), \qquad \text{with} \quad \nu_t(\mathbf{s}_i) \overset{ind}{\sim} \mathcal{N}\left(0, \tau_i^2\right), \end{aligned} \tag{1.8}$$

$\mathbf{z}_t$ is a vector containing dummy variables that account for the seasonality of the data, and $\boldsymbol{\beta}_i$ is the vector of location-specific regression parameters. Specifically, the vector $\mathbf{z}_t$ includes an intercept term and additional terms that indicate the season corresponding to the time point t. For example, if there are four seasons, $\mathbf{z}_t$ will consist of four components: the intercept, an indicator for spring, an indicator for summer, and an indicator for autumn. If all seasonal indicators are zero, the time point corresponds to winter. Unlike the formulation presented in equation (1.1), the model in equation (1.8) does not include station-specific covariates. From the second line of (1.8), it follows that the joint distribution of the spatio-temporal random effects $\mathbf{w}(\mathbf{s}_i) = (w_1(\mathbf{s}_i), \dots, w_T(\mathbf{s}_i))^\top$ is

$$\mathbf{w}(\mathbf{s}_i) \sim \mathcal{N}_T\left(\mathbf{0}, \mathbf{R}\left(\boldsymbol{\gamma}_i\right)\right) \quad \text{with} \quad \left[\mathbf{R}\left(\boldsymbol{\gamma}_i\right)\right]_{t,t'} = \frac{\tau_i^2}{1-\phi_i^2}\phi_i^{|t-t'|}$$

where $\boldsymbol{\gamma}_i = \left(\phi_i, \tau_i^2\right)$. The clustering of the monitoring stations is then based on the parameters driving the autoregressive process of the random effect $w_t(\mathbf{s})$. Specifically, we employ the sPPM (Page and Quintana, 2016), as described in Section 1.2.2, to group stations according to the values of the parameters $\boldsymbol{\gamma}_i$. In particular, we use the similarity function $g_3(\cdot)$ in equation (1.7) to account for the spatial dependence. Clustering based on the autoregressive coefficient ϕ_i and variance τ_i^2 involves grouping observations whose spatio-temporal random effects are driven by temporal processes with the same persistence and variability. This approach differs from clustering the entire time series, as discussed in Section 1.1.2. Here, we allow different time series to belong to the same cluster if they share the same temporal process parameters.

Before proceeding with posterior inference, we need to set the remaining prior distributions:

$$\begin{aligned}
&\boldsymbol{\beta}_i \sim \mathcal{N}_p\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right) \quad \text{for} \quad i = 1, \dots, n \quad \text{with} \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \text{diag}\left(\zeta_1^2, \dots, \zeta_p^2\right) \\
&\zeta_k^2 \overset{iid}{\sim} \text{IG}\left(a_\beta, b_\beta\right) \quad \text{for} \quad l = 1, \dots, p \\
&\sigma_i^2 \overset{iid}{\sim} \text{IG}\left(a_\epsilon, b_\epsilon\right) \quad \text{for} \quad i = 1, \dots, n \\
&\boldsymbol{\gamma}_k^* \overset{iid}{\sim} G_0 \quad \text{for} \quad k = 1, \dots, K \quad \text{with} \quad G_0 = \text{IG}\left(a_\tau, b_\tau\right) \times \text{Beta}\left(a_\phi, b_\phi\right)
\end{aligned} \tag{1.9}$$

where $\text{IG}(a, b)$ denotes an inverse gamma distribution with mean $b/(a-1)$, $\text{Beta}(a, b)$ denotes a beta distribution with mean $a/(a+b)$, and $\boldsymbol{\gamma}_1^*, \dots, \boldsymbol{\gamma}_K^*$ denote the cluster specific parameters so that $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_{c_i}^*$.

For our model specification, we selected hyperparameters to balance incorporating prior information and allowing the data to dominate the posterior inference. Specifically, for the inverse gamma prior on the variance of the regression coefficients, ζ_k^2, we set $a_\beta = 2$ and $b_\beta = 1$, i.e., a rather vague prior distribution with mean 1 and infinite variance. Similarly, for the residual variance σ_i^2, we chose $a_\epsilon = 2$ and $b_\epsilon = 1$, using a prior structure comparable to that of ζ_k^2. This is a reasonable assumption in many practical applications where the variance of the residuals is expected to be of the same order of magnitude as that of the regression coefficients. The hyperparameters of the base measure G_0 are set as follows. We selected $a_\tau = 0.5$ and $b_\tau = 0.5$; this choice corresponds to an undefined prior mean and a non-informative prior that allows for potentially large variances in the autoregressive process. Given that $w_t(\mathbf{s}_i)$ captures temporal dependence in the model, we allow substantial flexibility to account for strong dynamics where necessary. In fact, by not constraining the prior mean, the model is free to accommodate varying levels of temporal variability across spatial locations, which is especially useful in situations where temporal correlations may differ significantly. Finally, we set $a_\phi = 1$ and $b_\phi = 1$, corresponding to a Uniform distribution between 0 and 1. This reflects the assumption of stationarity in the univariate temporal process and excludes negative autocorrelation, which is consistent with our prior belief.

The full posterior inference is provided through an MCMC sampling scheme. In particular, we adapted the Gibbs sampler outlined in Section 1.2.1 to approximate the joint posterior distribution under model (1.8)-(1.9), see Algorithm 1. We then employ the greedy stochastic search approach of Dahl et al. (2022) to obtain a posterior point estimate of the partition from the posterior distribution of the cluster assignment of each station.

Results

Algorithm 1 was run for 5,000 iterations, with the first 2,000 discarded as burn-in, achieving a computational speed of 8.78 iterations per second. The code to reproduce the results in this section is available at https://github.com/lucaaiello/time_series_clustering/tree/main.

Panel (a) of Figure 1.4 presents the clustering results for the 162 PM10 monitoring stations across Northern Italy, based on their station-specific time series parameters. In particular, the estimated partition in four clusters is obtained using the VI loss function. Each cluster corresponds to a color, and the median values of the autoregressive coefficient and the variance of each cluster are indicated in the legend. In particular, we used the final MCMC iteration to estimate the cluster-specific parameters, calculating the mean parameters for each cluster based on the assigned observations.

Algorithm 1 Gibbs Sampling step for the sPPM model on time-series parameters.

1: update, for $i = 1, \dots, n$, $\boldsymbol{\beta}_i \mid \sigma_i^2, \gamma_i, \boldsymbol{\Sigma}_\beta, \mathbf{Z}, \mathbf{y}_i$ from

$$\mathcal{N}_p\left(\mathbf{V}_{\beta_i}\mathbf{Z}^\top\mathbf{Q}_i^{-1}\mathbf{y}_i, \mathbf{V}_{\beta_i}\right)$$

where $\mathbf{Q}_i = \sigma_i^2\mathbf{I} + \mathbf{R}(\gamma_i)$ and $\mathbf{V}_{\beta_i} = \left(\mathbf{Z}^\top\mathbf{Q}_i^{-1}\mathbf{Z} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right)^{-1}$

2: update, for $i = 1, \dots, n$, $\mathbf{w}_i \mid \sigma_i^2, \gamma_i, \boldsymbol{\beta}_i, \mathbf{Z}, \mathbf{y}_i$ from

$$\mathcal{N}_T\left(\sigma_i^{-2}\mathbf{V}_{\mathbf{w}_i}(\mathbf{y}_i - \mathbf{Z}\boldsymbol{\beta}_i), \mathbf{V}_{\mathbf{w}_i}\right)$$

where $\mathbf{V}_{\mathbf{w}_i} = \left(\sigma_i^{-2}\mathbf{I} + \mathbf{R}(\gamma_i)^{-1}\right)^{-1}$

3: update, for $i = 1, \dots, n$, $\sigma_i^2 \mid a_\epsilon, b_\epsilon, \boldsymbol{\beta}_i, \mathbf{w}_i, \mathbf{Z}, \mathbf{y}_i$

$$\mathrm{IG}\left(a_\epsilon + \frac{T}{2}, b_\epsilon + \frac{1}{2}(\mathbf{y}_i - \mathbf{Z}\boldsymbol{\beta}_i - \mathbf{w}_i)^\top(\mathbf{y}_i - \mathbf{Z}\boldsymbol{\beta}_i - \mathbf{w}_i)\right)$$

4: update, for $l = 1, \dots, p$, $\zeta_l^2 \mid a_\beta, b_\beta, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n$

$$\mathrm{IG}\left(a_\beta + \frac{n}{2}, b_\beta + \frac{1}{2}\sum_{i=1}^{n}\alpha_{il}^2\right)$$

5: **for** $i = 1, \dots, n$ **do**

6: update $(\gamma_1^*, \dots, \gamma_{K^{(-i)}}^*)$ with the $K^{(-i)}$ unique values in $\gamma^{(-i)}$

7: update $(\gamma_{K^{(-i)}+1}^*, \cdots, \gamma_{K^{(-i)}+K_{aux}}^*)$ from

$$G_0 = \mathrm{IG}\left(a_\tau, b_\tau\right) \times \mathrm{Beta}\left(a_\phi, b_\phi\right)$$

8: update c_i with the following probability

$$\Pr\left(c_i = k \mid -\right) \propto \begin{cases} f(\mathbf{w}_i \mid \gamma_k^*, -)\frac{C\left(S_k^{(-i)} \cup \{i\}\right) g\left(s_k^{*(-i)} \cup \mathbf{s}_i\right)}{C\left(S_k^{(-i)}\right) g\left(s_k^{*(-i)}\right)} & k = 1, \dots, K^{(-i)} \\ f(\mathbf{w}_i \mid \gamma_k^*, -) C\left(\{i\}\right) g\left(\mathbf{s}_i\right) & k = K^{(-i)} + 1, \dots, K^{(-i)} + K_{aux} \end{cases}$$

where $f(\mathbf{w}_i \mid \gamma_k^*, -)$ is the probability density function of $\mathbf{w}_i$ evaluated at the k-th cluster specific parameters, and K_{aux} is the additional number of clusters, see Neal (2000) for details.

9: set $\gamma_i = \gamma_{c_i}^*$

10: **end for**

The blue cluster has the lowest autocorrelation (0.10) and a variance of 1.05, while the green cluster

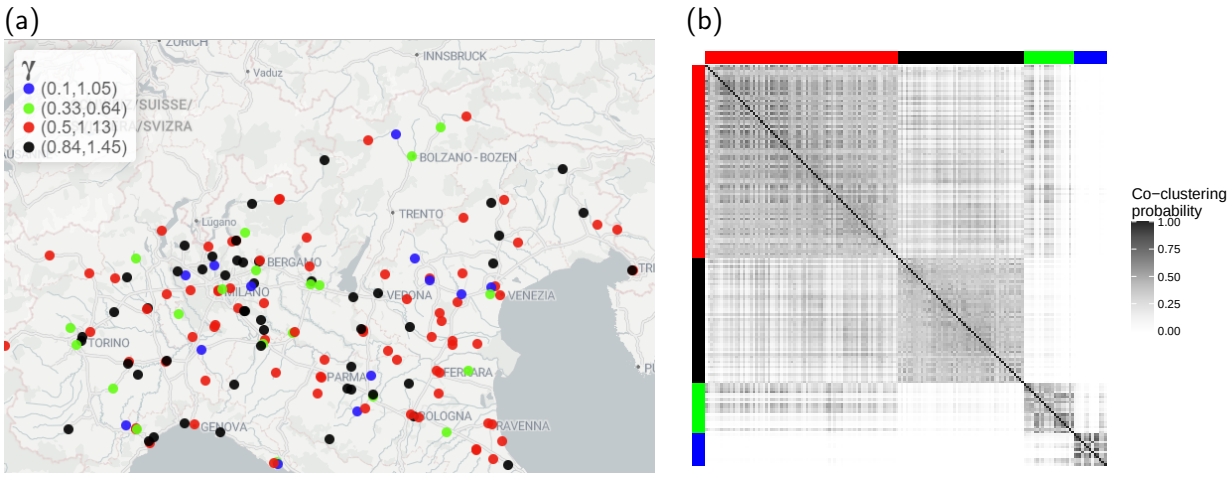


Figure 1.4: Estimated partition of the monitoring stations (a); posterior co-clustering matrix (b).

shows slightly higher autocorrelation (0.33) with a variance of 0.64. The red cluster presents an autocorrelation of 0.5 and a higher variance of 1.13, indicating stations with more fluctuating PM10 levels. The black cluster, instead, shows the highest autocorrelation (0.84) and variance (1.45). The numerical values in each cluster provide insights into PM10 temporal behavior across the region. Stations with higher autocorrelation, often in urban and industrial areas, likely experience much more variability in the pollution levels, while some of those with lower variances also have lower levels of persistence.

Beyond providing a point estimate of the partition, panel (b) of Figure 1.4 shows the heatmap of the co-clustering matrix, where each entry represents the proportion of MCMC iterations in which two individuals (monitoring stations) were clustered together based on posterior samples. The stations are ordered according to the estimated partition. The heatmap clearly shows a block structure for the green and blue clusters, indicating a strong alignment between such clusters and their co-clustering probabilities. In contrast, the red and black clusters show more uncertainty, with relatively high probabilities also appearing off the block diagonal. Darker cells correspond to pairs of stations frequently clustered together, often located in geographically contiguous areas, highlighting the influence of spatial proximity on air quality trends.

1.4 Summary

In this work, we provided an overview of Bayesian nonparametric clustering for spatio-temporal environmental data. We first review models for spatio-temporal data, followed by a detailed discussion on cluster analysis, focusing on Bayesian nonparametric methods. We then summarized posterior inference for Bayesian clustering and explored the literature on Bayesian nonparametric clustering for spatio-temporal data, covering relevant models and applications. Finally, we discussed the PPM and its spatial extensions, which are later used in our case study.

Clustering is a fundamental problem in statistics, particularly in the context of environmental and spatio-temporal data, where the clustering of both space and time must be accounted for. Model-based clustering offers the advantage of providing probabilistic allocations of observations to clusters

and probabilistic definitions of the number of clusters. In a Bayesian framework, a popular approach for model-based clustering is to use a mixture model with an unknown number of clusters, often incorporating either a finite or infinite number of components. This results in a random partitioning of observations, making it suitable for modeling environmental data that exhibits both spatial and temporal variation.

However, classic models have been shown to be inconsistent for estimating the number of clusters, especially in spatio-temporal settings where environmental factors can influence cluster structure. Moreover, in finite mixture models, the prior on the number of components strongly affects cluster estimation. This can be problematic when modeling environmental data with complex spatial and temporal dependencies, such as pollution levels that change across regions and seasons.

Recent developments in Bayesian clustering have also extended to clustering populations or spatial regions, addressing the need for population-based or region-based clustering in environmental risk assessment. For this, Gibbs-type priors have gained attention as they extend traditional mixture models to better handle spatial clusters across populations of observations, such as air quality measurements across different regions.

An alternative is to directly model the PPM, which is particularly suited to environmental data. Although PPMs offer flexibility in modeling partitions, they do not take into account spatial dependence, which is particularly pivotal when dealing with spatio-temporal datasets.

Finally, summarizing the posterior distribution of the partition is a crucial aspect of Bayesian clustering, particularly for spatio-temporal data, where environmental and geographical factors play a key role. Various decision-theoretic approaches have been proposed to achieve optimal summarization, comparing different methods for obtaining a representative partition from the posterior distribution.

The application presented in this work demonstrates a Bayesian model-based clustering algorithm applied to environmental data. We employed a multi-level hierarchical linear mixture model, incorporating a first-order autoregressive process to account for temporal effects, while clustering was driven by the discreteness of the nonparametric prior. Given the relatively heterogeneous pollution levels across Northern Italy, we developed a model emphasizing persistence and variability. A sPPM was used as a prior for the autoregressive process parameters, enabling spatial clustering of monitoring stations. The results indicated that persistence was higher in urban areas and lower in vegetated or coastal regions with frequent breezes.

Future research in Bayesian nonparametric clustering for spatio-temporal environmental data could focus on exploring alternative cohesion and similarity functions within the sPPM framework. Additionally, extending the sPPM approach to handle areal data would represent a novel contribution. Another important avenue is refining hyperparameters selection to enhance model performance. Lastly, there appears to be a gap in the literature concerning the predictive capabilities of these models. While they have demonstrated strong inferential properties – each with its own limitations and requirements – there is limited work addressing their predictive performance in both spatial and temporal contexts.

Chapter 2

Detecting spatial health disparities using disease maps

Health disparities, or inequities, broadly refer to sections of the population being deprived of fair and equal opportunities to seek healthcare or exhibiting different disease incidences and risks (see, e.g., the recent text by Rao, 2023). Spatial disparities in health manifest as variations in health outcomes over geographic regions and are often visually represented using disease maps (Koch, 2005). Spatial data analysis in epidemiological investigations is extensively documented (see, e.g., Waller and Gotway, 2004; Waller and Carlin, 2010; Lawson, 2013; Lawson et al., 2016, and further references therein).

Detecting "boundaries" on disease maps is a specialized exercise directly relevant to assessing health disparities. In public health, such *difference boundaries*, or *wombling boundaries* (so called after Womble, 1951), indicate significantly different disease mortality and incidence between neighbors and assists in decision-making for disease prevention and control, geographic allocation of resources, and so on (see Jacquez and Greiling, 2003b,a; Lu and Carlin, 2005; Li et al., 2011; Ma and Carlin, 2007; Fitzpatrick et al., 2010, for some algorithmic approaches with applications). Model-based approaches aiming for full probabilistic uncertainty quantification have also received attention and include, but are not limited to, developments in Lu et al. (2007); Ma et al. (2010); Li et al. (2015, 2012); Hanson et al. (2015); Corpas-Burgos and Martinez-Beneito (2020); Gianella et al. (2023); Gao et al. (2023), and Pavani and Quintana (2024).

This chapter builds on the above developments in what may be described as still a fledgling area. We devise a probabilistic learning mechanism for adjacency boundaries on a map when such information may become available. An administrative or political boundary separating two adjacent regions need not delineate them in terms of the health outcomes measured there. In fact, in spatial smoothing the customary approach is to assume that neighboring regions are similar to each other. This assumption runs counter to detecting health disparities, where we seek out differences and not similarities. Therefore, modeling spatial disparities needs to balance underlying processes generating similarities based upon geographic proximity with processes generating significant differences between neighbors. We specifically pursue an effective manner of modeling the adjacency relations on a map.

Our contribution is summarized in the context of the two broad themes. The first builds stochastic models for the adjacency matrices (Lu et al., 2007; Ma et al., 2010; Lee and Mitchell, 2012; Liang, 2019; Corpas-Burgos and Martinez-Beneito, 2020) that, in principle, can accommodate information

from explanatory variables in ascertaining the presence of edges. Based upon the complexity of the resulting models and numerical difficulties in estimating such edges, a different approach forgoes introducing variables in the adjacency, but detects edges using multiple comparisons by estimating the spatial random effects and, subsequently, testing how many such pairs are significantly different. A Bayesian approach (Li et al., 2015, 2012; Gao et al., 2023) holds appeal as the posterior distribution of spatial effects produce an exact Bayesian false discovery rate (FDR) (Müller et al., 2004) without recourse to asymptotic assumptions that are inappropriate in areal settings.

We offer a framework that melds the two themes. We build upon Gao et al. (2023) by embedding directed acyclic graphical autoregression (DAGAR) models (Datta et al., 2019) within a hierarchical Bayesian nonparametrics framework. This nonparametrics specification is critical as it introduces discrete probability masses on the spatial random effects and allows us to meaningfully calculate the posterior probabilities of two neighboring spatial effects being equal. This sets up boundary detection using Bayesian FDR. We further enrich the model by allowing the DAGAR adjacency matrix to learn from explanatory variables that may carry information regarding whether two neighboring regions are similar. Thus, each element of a binary adjacency matrix is modeled using an exponential threshold devised by Lee and Mitchell (2012). We develop joint models for multiple diseases (cancers), where the diseases are dependent and accommodate disease-specific learning of adjacency relationships. We allow difference boundaries to vary by health outcomes and explain the impact of explanatory variables on such boundaries. Here, we consider unstructured disease dependencies and those posited by conditional dependencies in graphical models.

The data analytic innovation and merits of our approach are best appreciated in comparison to approaches that adhere to one, but not both, of the boundary detection themes discussed above. Methods that allow uncertainty through purely parametric adjacency relations (including regression models for adjacency matrices) within widely used conditional autoregression models often encounter numerical difficulties in convergence of iterative estimation algorithms, such as Markov chain Monte Carlo (MCMC). On the other hand, methods offering FDR based boundary detection without allowing adjacency learning from explanatory variables can lead to inflated detection rates of boundaries by failing to account for additional uncertainties in adjacency relations. We demonstrate this phenomenon in the cancer boundary detection rates reported recently by Gao et al. (2023).

The remainder of this chapter is organized as follows. Section 2.1 provides a detailed description of the dataset, which is sourced from the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results (SEER) database. This dataset includes records of four cancer types across California counties, along with explanatory variables for modeling outcomes and geographic boundaries. Section 2.2 introduces the likelihood model used to describe the response variable, incorporating an areal stick-breaking process. Section 2.3 details the DAGAR model used to account for spatial dependencies, explaining how the adjacency structure is modeled. In Section 2.4, we extend the framework by developing a multivariate Bayesian model using the MDAGAR approach to capture disease dependencies. Section 2.5 outlines the full model specification, including the prior distributions for all model parameters. Section 2.6 explains the FDR procedure for detecting boundaries. Section 2.7 discusses the computational aspects of our method, highlighting its computational advantages. In Section 2.8, we present simulation experiments that assess the performance of the proposed model using various metrics. Section 2.9 applies the spatial boundary detection method to the SEER

dataset. Finally, Section 2.10 provides concluding remarks. Additional details on the MCMC scheme and supplementary results from the simulation experiments and real data analysis are presented in Appendix B.

2.1 Data

We explore an areal data set extracted from the SEER*Stat database using the SEER*Stat statistical software (National Cancer Institute, 2019). The data records the occurrence of four potentially interrelated types of cancers: lung, esophageal, larynx, and colorectal. The occurrences of these cancers are aggregated from January 2012 through December 2016 and presented as counts in each of the 58 counties of California. Previous research has shown that lung and esophageal cancers share common risk factors (Agrawal et al., 2018) and metabolic mechanisms (Shi and Chen, 2004). Furthermore, it has been found that lung cancer is one of the most frequent second primary cancers in patients with colon cancer (Kurishima et al., 2018). Patients with laryngeal cancer are also at a high risk of developing second primary lung cancer (Akhtar et al., 2010). Hence, we seek to account for the dependence between the cancers due to non-spatial factors.

Let (y_{id}) be the observed counts for cancer type d $(d = 1, 2, 3, 4)$ in county i $(i = 1, 2, \ldots, 58)$. Our model, developed in Section 2.2, accounts for the expected number of cases (E_{id}) by adjusting for age-sex demographics in each county. Specifically, we calculate the expected age-sex adjusted number of cases in county i for cancer d as $E_{id} = \sum_{k=1}^{m} c_d^k N_i^k$ (Jin et al., 2005), where $c_d^k = (\sum_{i=1}^{58} y_{id}^k)/(\sum_{i=1}^{58} N_i^k)$ is the age-sex specific incidence rate in age-sex group k for cancer d across all California counties; y_{id}^k represents the incidence count in age-sex group k of county i for cancer d, and N_i^k represents the population in age-sex group k of county i. The age groups have been determined based on 5-year intervals up to 85 years or older. These age intervals are as follows: less than 1 year, 1–4 years, 5–9 years, 10–14 years, and so on, up to 80–84 years and 85+ years. This results in a total of $m = 19 \times 2 = 38$ age-sex groups.

Figure 2.1 displays a map of California's counties, illustrating the age-sex adjusted standardized incidence ratios $(SIR_{id} = y_{id}/E_{id})$ for different types of cancer. The map reveals a notable pattern, where regions exhibiting comparable SIRs tend to cluster together geographically, forming distinct groups. This clustering is evident in the group of regions with the highest incidence rates for each cancer. The high SIR values for all of the cancer types are concentrated in northern counties, with the only exception being that of a high SIR for colorectal cancer in a southern county (San Bernardino). Additionally, in northern counties, we notice that several counties with high SIRs are situated next to areas with low SIRs. Our proposed model, which detects boundaries, provides a robust, adaptable, interpretable and clear analytical tool for identifying these geographical clusters.

We used Pearson's correlation coefficient as an exploratory tool to evaluate the relationships between various types of cancers. In a preliminary analysis, we treated the SIRs from different counties as independent samples. Our findings indicate a significant association in the incidence of lung cancer with each of esophageal (correlation coefficient 0.58), larynx (0.40), and colorectal (0.50) cancers. We also measured a correlation coefficient of 0.42 between esophageal and larynx cancer. Moreover, to assess the spatial relationships of each cancer type, we employed the Moran's I statistic, which was computed based on the r-th order neighbors for each cancer. These calculations

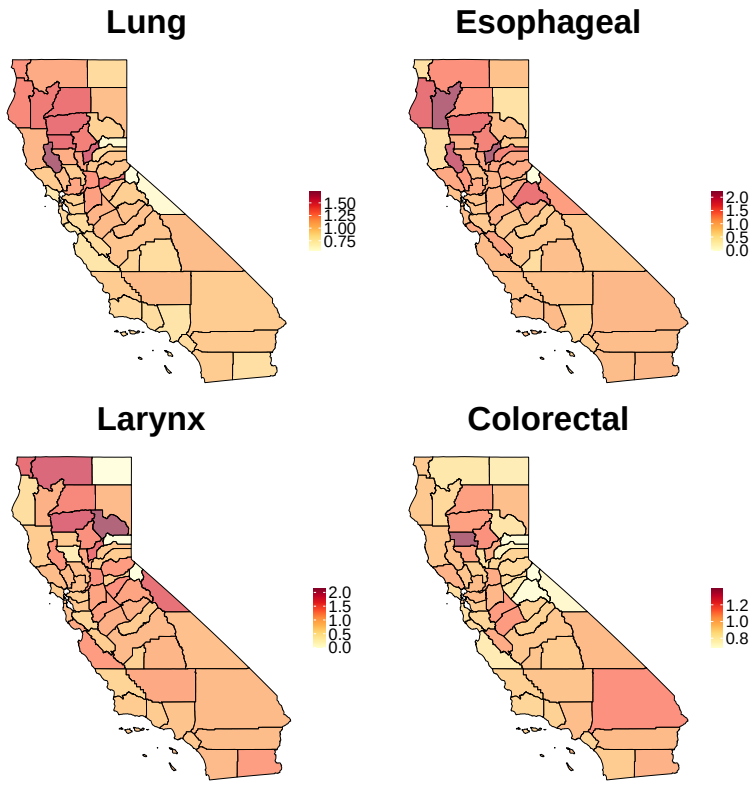


Figure 2.1: Maps of age-sex adjusted standardized incidence ratios (SIR) for lung, esophageal, larynx, and colorectal cancer in counties of California, 2012–2016.

were then used to generate an areal correlogram, as described by Banerjee et al. (2014). Here, we construct 11 distance intervals or bins $(0, d_1], (d_1, d_2], (d_2, d_3], \ldots$ using by setting each $d_r = 0.01 \times r$ for $r = 1, \ldots, 11$. We compute Euclidean distances between the county centroids from an Albers map projection of California and define r-th order neighbors as regions that fall within the bin $(d_{r-1}, d_r]$ of each other, i.e., they are separated by a distance greater than d_{r-1} but within a distance of d_r. These neighbors are then used to calculate Moran's I statistic. Figure 2.2 shows the spatial associations for lung, esophageal, colorectal, and larynx cancers. As the order of neighbors (r) increases, the spatial associations diminish notably for lung, esophageal, and colorectal cancers, but the observed pattern is less pronounced and clear for larynx cancer.

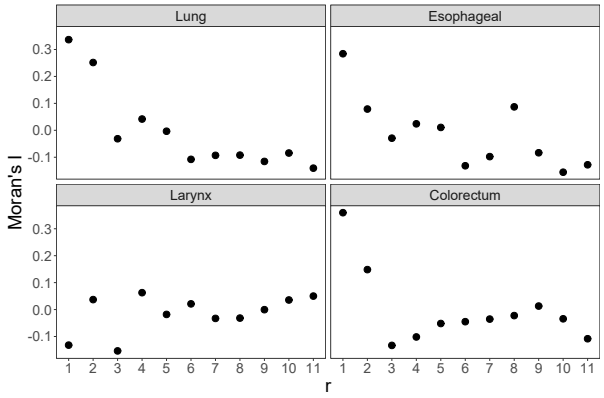


Figure 2.2: Moran's I statistic using r-th order neighbors for four cancers.

We introduce explanatory variables into our model to capture the mean structure, account for risk factors and identify neighboring areas with significant differences in rates. A detailed analysis is presented in Section 2.9. Figure 2.3 presents rates (in percentage scale) for smoking; population over 65; and below poverty threshold rates for each county, arranged from left to right. These plots reveal notable patterns: the northern counties exhibit a higher concentration of smoking rates, with a few exceptions in the central counties; both northern and eastern counties have higher rates of individuals over 65; and the northern and central counties have higher rates below the poverty threshold. Conversely, the coastal regions demonstrate relatively lower rates across all the covariates, indicating a distinct contrast between different geographical areas (see Doll et al., 2005, for similar findings). Interestingly, central and southern counties exhibit lower rates specifically in the over 65 variable. The significance and implications of these findings will be discussed in greater detail in Section 2.9.

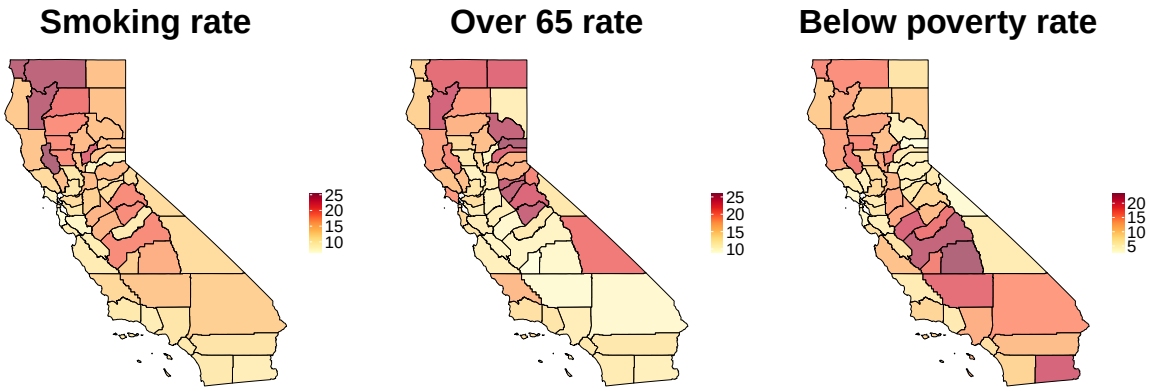


Figure 2.3: California's county-level rates (as percentages of population) for smokers, elderly population (over 65) and individuals below poverty line in terms of annualized income.

2.2 Likelihood model

We construct a generalized linear mixed model for the disease outcome, y_{id}, using a distribution from the exponential family with a canonical link,

$$g(E(y_{id})) = \mathbf{x}_{id}^{\top}\boldsymbol{\beta}_d + \phi_{id}, \tag{2.1}$$

for each areal unit $i = 1, \ldots, n$ and disease $d = 1, \ldots, q$. The large scale trend is modeled using a regression, where $\mathbf{x}_{id}$ is a vector of explanatory variables and $\boldsymbol{\beta}_d$ is the corresponding vector of slopes, while ϕ_{id} is the spatial random effect for disease d in region i. For detecting difference boundaries, we seek $P(\phi_{id} \neq \phi_{jd} \,|\, \mathbf{y}, i \sim j)$, which offers a fully model-based stochastic mechanism to calibrate difference boundaries, as in Li et al. (2015), where $\mathbf{y}$ is the collection of observed y_{id}'s and $i \sim j$ means that regions i and j are neighbors. While risks tend to fluctuate continuously across space, neighboring regions may exhibit discontinuities attributable to disparities arising from vastly different underlying factors (e.g., socioeconomic characteristics). This raises an issue about the prior distribution for the ϕ_{id}'s. A continuous probability distribution is inappropriate because the resulting posterior probability is $P(\phi_{id} \neq \phi_{jd} \,|\, \mathbf{y}, i \sim j) = 1$, which classifies all possible geographic boundaries

as difference boundaries. This limitation highlights the need for a model that accounts for spatial dependence while assigning discrete probability masses on spatial random effects. Such a model enables including non-zero probabilities for the random effects being equal. Indeed, our approach will ensure non-null probabilities of equality between the random effects, resulting in a smaller number of boundaries detected compared to n. The Dirichlet Process (DP) offers a natural way to cluster regions by endowing possibly positive mass on $P(\phi_{id} = \phi_{jd} \,|\, \mathbf{y}, i \sim j)$. The DP is especially appealing here as it still captures spatial associations in ϕ_{id} through the baseline covariance (or precision) matrix. The model we develop here belongs to a subclass of stick-breaking process priors and includes the DP as a special instance.

Let $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^\top, \dots, \boldsymbol{\phi}_q^\top)^\top$ be the $N \times 1$ vector, where $N = nq$ and $\boldsymbol{\phi}_d = (\phi_{1d}, \dots, \phi_{nd})^\top$ for each $d = 1, \dots, q$. Following Gao et al. (2023) we let $\{1, \dots, n, \dots, (q-1)n+1, \dots, N\}$ be an enumeration of the pairwise (i, d) indices corresponding to the ordering of $\boldsymbol{\phi}$, thereby dealing with a unique $N \times 1$ vector $\boldsymbol{\phi}$. We let $\boldsymbol{\phi} \sim G_N$, where G_N is an unknown distribution further specified by $G_N = \sum_{u_1,\dots,u_N} \pi_{u_1,\dots,u_N} \delta_{(\theta_{u_1},\dots,\theta_{u_N})}$, where $u_1, \dots, u_N \in \{1, \dots, K\}$, $\theta_k \,|\, \tau \overset{iid}{\sim} \mathcal{N}(0, 1/\tau)$ for $k = 1, \dots, K$, $\delta_{(\theta_{u_1},\dots,\theta_{u_N})}$ is the Dirac measure located at $(\theta_{u_1}, \dots, \theta_{u_N})$ and

$$\pi_{u_1,\dots,u_N} = P\left(\sum_{k=1}^{u_1-1} p_k < F^{(1)}(\gamma_1) < \sum_{k=1}^{u_1} p_k, \dots, \sum_{k=1}^{u_N-1} p_k < F^{(N)}(\gamma_N) < \sum_{k=1}^{u_N} p_k\right) \tag{2.2}$$

with $\boldsymbol{\gamma} \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{\Sigma}_\gamma)$. To clarify, the weights in equation (2.2) are constructed using the marginal cumulative distribution function of the elements of $\boldsymbol{\gamma}$. This construction integrates two approaches: the hybrid Dirichlet process (Petrone et al., 2009), which leverages point-referenced continuous spatial copulas to model dependencies among the weights, and the latent stick-breaking process (Rodríguez et al., 2010), which employs ordered atoms. Within the probability operator, there are N elements, each corresponding to a specific component of $\boldsymbol{\gamma}$. Changes in these components influence the value of $\pi_{u_1,\dots,u_N}$, which subsequently determines the realization of the process G_N. Marginally, each $F^{(\cdot)}(\gamma_\cdot)$ follows a $\mathcal{U}(0, 1)$ distribution, but the components are dependent through the joint distribution of $\boldsymbol{\gamma}$. Moreover, the distribution of each spatial random effect $\phi_\cdot$ is governed by a univariate Dirichlet Process (DP), with spatial dependence between these DPs introduced via a copula representation for the weights. Priors for model parameters are specified as $\boldsymbol{\beta}_d \overset{ind}{\sim} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\beta_d})$, while the precision (inverse variance) $1/\tau \sim \text{IG}(a, b)$, where a is shape and b is scale. We specify the stick-breaking weights by $p_1 = V_1$ and $p_j = V_j \prod_{k<j}(1 - V_k)$ for $j = 2, \dots, K$, where $V_k \overset{iid}{\sim} \text{Beta}(1, \alpha)$ and $F^{(1)}(\cdot), \dots, F^{(N)}(\cdot)$ are cumulative distribution functions of the marginal distribution of the corresponding $\boldsymbol{\gamma}$ elements indexed in accordance with the elements of $\boldsymbol{\phi}$.

The marginal covariance between spatial random effects ϕ_g and ϕ_h can then be expressed as

$$\text{Cov}(\phi_g, \phi_h) = \frac{b}{a-1} \sum_{k=1}^{K} \pi_{kk}^{(g,h)}, \tag{2.3}$$

where $g, h = 1, \dots, N$ and

$$\pi_{kk}^{(g,h)} = P\left(\sum_{t=1}^{k-1} p_t < F^{(g)}(\gamma_i) < \sum_{t=1}^{k} p_t, \sum_{t=1}^{k-1} p_t < F^{(h)}(\gamma_j) < \sum_{t=1}^{k} p_t\right) \tag{2.4}$$

Equation (2.3) and (2.4) evinces how the spatially dependent elements of γ induce associations among the elements of ϕ. The covariance of ϕ also depends on the hyperparameters of the precision parameter prior of the base distribution. Higher values of the probabilities, i.e., $\pi_{kk}^{(g,h)}$, produce higher correlation between the random effects. Unsurprisingly, a higher likelihood of ϕ_g and ϕ_h exhibiting similar values implies a higher covariance in (2.3). We next attend to modeling the spatial covariance matrix $\mathbf{\Sigma}_\gamma$.

The choice of the independent and identically distributions for the atoms base distribution stems from the necessity of establishing ties across both regions and cancers. For example, if we were to assign different means to different cancers, we could still accommodate ties within each cancer, wherein different regions' random effects assume the same value. However, we would be unable to establish ties across cancers as the underlying distribution from which the atoms are sampled would differ.

In theory, $K = \infty$ yields a fully non-parametric model, but we follow the standard practice of replacing the infinite sum with the sum of the initial K ($K \leq n$) terms, as the probability masses diminish rapidly. Should concerns about truncation bias arise, an alternative method called slice-sampling (described in Kalli et al., 2011) is available for exact sampling. In particular, for areal data the value of K is naturally constrained by the number of areal units rendering the infinite representation redundant. See Müller et al. (2015) for further details on nonparametrics models and their scope of inference.

2.3 Spatial dependence

This section outlines how spatial dependence is incorporated into our model, and is structured as follows. Section 2.3.1 provides an overview of the DAGAR model, detailing its theoretical basis and emphasizing its advantages over traditional spatial models. In Section 2.3.2, we discuss how adjacency relationships between spatial units are integrated into the DAGAR framework, highlighting its flexibility in capturing spatial structures. Lastly, Section 2.3.3 offers a computational comparison between the DAGAR and CAR models, focusing on their respective efficiencies and computational benefits.

2.3.1 DAGAR review

We briefly review the univariate DAGAR which uses a fixed set of arbitrarily ordered regions yielding a *topologically ordered* set of vertices, $\mathcal{V} = \{1, \ldots, n\}$, and a set $\mathcal{E}$ of directed edges that encode "neighbors" of every region. Rather than defining all geographically adjacent regions as neighbors, DAGAR defines neighbors of a given region i as regions that (i) are geographically adjacent to i and (ii) precede i in the topological order. The directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed from this definition. The precision matrix of a DAGAR random variable is $\mathbf{Q}(\rho) = (\mathbf{I} - \mathbf{B})^\top \mathbf{\Lambda}(\mathbf{I} - \mathbf{B})$, where $\mathbf{B} = (b_{ij})$ is $n \times n$ with elements $b_{ij} = \frac{\rho}{1+(n_{<i}-1)\rho^2}$ if there is a directed edge $(i, j) \in \mathcal{E}$ for $i \geq 2$, where $n_{<i}$ is the number of neighboring regions preceding i, and $b_{ij} = 0$ if $(i, j) \notin \mathcal{E}$. In DAGAR, $(i, j) \in \mathcal{E}$ if $j < i$ in the topological order and regions i and j are geographically adjacent. This implies that $b_{ii} = 0$ and $b_{ij} = 0$ for all $i \leq j$ so $\mathbf{B}$ is strictly lower triangular. The $n \times n$ diagonal matrix $\mathbf{\Lambda} = \text{diag}(\lambda_i)$ has elements $\lambda_i = \frac{1+(n_{<i}-1)\rho^2}{1-\rho^2}$ along its diagonal for $i = 1, \ldots, n$, where $n_{<1} = 0$.

Datta et al. (2019) showed that ρ can be interpreted as a spatial autocorrelation parameter and that the ordering is irrelevant or not impacting the analysis in most applications.

The construction of the DAGAR precision matrix is particularly appealing due to its exclusive dependence on the ρ parameter and its ability to account for all edges without any omissions, while still maintaining computational efficiency. To illustrate the advantages of the DAGAR structure, we explicitly present its precision matrix below:

$$\begin{aligned}
\mathbf{Q}(\rho) &= (\mathbf{I}-\mathbf{B})^\top \mathbf{\Lambda}(\mathbf{I}-\mathbf{B}) \\
&= \begin{bmatrix} 1 & -b_{12} & \cdots & -b_{1n} \\ 0 & 1 & \cdots & -b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -b_{12} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -b_{1n} & -b_{2n} & \cdots & 1 \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 + \sum_{i=2}^n \lambda_i b_{1i}^2 & -\lambda_2 b_{12} + \sum_{i=3}^n \lambda_i b_{1i} b_{2i} & \cdots & -\lambda_n b_{1n} \\ -\lambda_2 b_{12} + \sum_{i=3}^n \lambda_i b_{1i} b_{2i} & \lambda_2 + \sum_{i=3}^n \lambda_i b_{2i}^2 & \cdots & -\lambda_n b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\lambda_n b_{1n} & -\lambda_n b_{2n} & \cdots & \lambda_n \end{bmatrix}
\end{aligned}$$

One of the main advantages of the DAGAR formulation is the computation of the determinant and the quadratic form in the multivariate normal probability density function. Specifically:

$$|\mathbf{Q}(\rho)| = |\mathbf{I}-\mathbf{B}||\mathbf{\Lambda}||\mathbf{I}-\mathbf{B}| = \prod_{i=1}^n \lambda_i$$

$$\boldsymbol{\gamma}^\top \mathbf{Q}(\rho)\boldsymbol{\gamma} = \boldsymbol{\gamma}^\top (\mathbf{I}-\mathbf{B})^\top \mathbf{\Lambda} (\mathbf{I}-\mathbf{B})\boldsymbol{\gamma} = \lambda_1\gamma_1^2 + \sum_{i=1}^n \lambda_i(\gamma_i - \sum_{j<i}\gamma_j b_{ij})^2$$

which are way less heavy to compute from a computational perspective. In particular the determinant computation reduces to computing a product of n terms while the quadratic form is lighter since it takes into account the sparsity of the structure and the strictly lower triangularity of the $\mathbf{B}$ matrix.

Two extensions to the DAGAR are relevant to our current data analytic aims. First, we intend to model the adjacency matrix rather than specify it purely from the map. Geographic adjacency, by itself, is not enough to infer similarities or disparities and we accommodate learning about the adjacency relations using risk factors or other explanatory variables. We adopt one possible fully model-based approach for inferring about spatial disparities by statistically estimating difference boundaries. We avail of spatially oriented risk factors and explanatory variables to drive inference on difference boundaries. Second, while extending to joint models for dependent cancer occurrences, we allow the impact of the factors informing about the edges to vary by the diseases under considerations. This synthesizes the approaches in Lee and Mitchell (2012) and Gao et al. (2023).

2.3.2 DAGAR adjacency modeling

An appealing feature of the DAGAR is that the parameter ρ represents spatial autocorrelation, unlike in traditional spatial autoregression models such as simultaneous and conditional autoregression (SAR and CAR, respectively). However, it is unclear how information from explanatory variables, predictors

or risk factors can be introduced in the adjacency. We achieve this by constructing a strictly lower triangular $n \times n$ adjacency matrix, $\widetilde{\mathbf{W}} = (w_{ij})$ based upon the fixed topological order of regions to assimilate information from preceding neighboring regions. This results in $w_{ij} = 0$ for $i \leq j$ or $i \not\sim j$ (i and j not neighbors). The neighbor set for region i in DAGAR are defined as geographically adjacent regions preceding i in $\mathcal{V}$. Introducing a lower triangular adjacency matrix $\widetilde{\mathbf{W}}$ facilitates the subsequent exposition. For each region i, its neighbors are $\{j : j < i, j \sim i\} \equiv \{j : w_{ij} = 1\}$. We now introduce $\widetilde{\mathbf{W}}$ into the DAGAR precision matrix as

$$\mathbf{Q}(\rho, \widetilde{\mathbf{W}}) = (\mathbf{I} - \mathbf{B})^\top \mathbf{\Lambda} (\mathbf{I} - \mathbf{B})$$

where $\mathbf{B} = \widetilde{\mathbf{B}} \circ \widetilde{\mathbf{W}}$, $\widetilde{\mathbf{B}} = (\widetilde{b}_{ij})$ with elements $\widetilde{b}_{ij} = \frac{\rho}{1+(n_{<i}-1)\rho^2}$, $\lambda_i = \frac{1+(n_{<i}-1)\rho^2}{1-\rho^2}$, $n_{<i} = \sum_{j=1}^{n} w_{ij}$ and $\circ$ denotes the Hadamard (or elementwise) product.

The matrix $\widetilde{\mathbf{B}}$ is dense with all its elements being possibly non-zero. However, $\widetilde{\mathbf{B}} \circ \widetilde{\mathbf{W}}$ retains the sparse lower-triangular structure characteristic of DAGAR. Instead of directly setting the non-null elements of $\widetilde{\mathbf{W}}$ to fixed values, one possibility is to model each element of $\widetilde{\mathbf{W}}$ as a separate random quantity. For instance, Lu et al. (2007) employed a logistic model for the elements of a symmetric adjacency matrix in a CAR model while considering boundary-specific risk factors. In our DAGAR setting, a parallel approach would be to model the lower triangular elements $\widetilde{w}_{ij} \mid p_{ij} \overset{ind}{\sim} \text{Bernoulli}(p_{ij})$ and introducing a regression model in, for example, a logistic link $\log(p_{ij}/(1-p_{ij}))$.

Other possible approaches include adapting random adjacency models, such as in Ma et al. (2010), to DAGAR by leveraging an Ising model for elements of the adjacency. However, the information needed to estimate such models is rarely available and require informative prior distributions resulting in excessively parametrized covariance models for γ that inaccurately estimate uncertainties and produce biased inference for spatial boundaries (also see Li et al., 2011, who argue that effectively modeling geographical adjacency would require a separate parameter for every pair of neighbors resulting in overparametrized models).

In what follows, we adopt a simpler and more effective approach based on Lee and Mitchell (2012), who propose modeling the adjacency matrix as

$$\widetilde{w}_{ij}(\boldsymbol{\eta}) = \begin{cases} 1 & \text{if} \quad \exp\left(-\mathbf{z}_{ij}^\top \boldsymbol{\eta}\right) \geq 0.5 \quad \text{and} \quad i \sim j \\ 0 & \text{otherwise}, \end{cases} \tag{2.5}$$

where $\mathbf{z}_{ij}$ consist of explanatory variables specific to pairs of regions i and j and are assumed to be non-negative in each element. In practice, they could represent an absolute measure of discrepancy between the two regions with $\boldsymbol{\eta}$ being the vector of corresponding coefficients. Equation (2.5) encodes geographic neighbors so as to not represent disparities whenever $\mathbf{z}_{ij}^\top \boldsymbol{\eta} \leq \log 2$. Furthermore, $\mathbf{z}_{ij}$ excludes an intercept so regions with homogeneous populations, i.e., $\mathbf{z}_{ij} = \mathbf{0}$, are deemed adjacent. The coefficients in $\boldsymbol{\eta}$ act as disparity parameters and restricted to be nonnegative so the greater the dissimilarity between two regions, the more likely there is a difference boundary between them. Each element in $\boldsymbol{\eta}$ is assigned a Uniform prior distribution between 0 and a fixed positive constant such that at most 50% of geographic neighbors in the study region are classified as non-adjacent. This maintains some degree of smoothing; see Lee and Mitchell (2012) for further details. Higher values in $\boldsymbol{\eta}$ assist in staying below the threshold, which results in a loss of a geographical boundary in the

adjacencies to be smoothed over and, hence, indicates a spatial disparity.

A salient feature of the preceding development is assimilating information from explanatory variables $\mathbf{x}_{id}$ in (2.1) and $\mathbf{z}_{ij}$ in (2.5). The former informs about the observed values of the health outcomes and captures large scale trends through the mean structure, while the latter provides information on disparities among geographic neighbors. The vectors $\mathbf{x}_{id}$ and $\mathbf{z}_{ij}$ can share variables, but need to be indexed differently. For example, one could include smoking rate for a region i in $\mathbf{x}_{id}$ and the difference in smoking rates between regions i and j in $\mathbf{z}_{ij}$. The coefficients in $\boldsymbol{\eta}$ indicate if the corresponding variable in $\mathbf{z}_{ij}$ signifies disparity. They can be map-based such as the distance between the centroids of regions i and j or the discrepancies in the percentage of common boundaries shared by the two regions compared to their total geographical boundaries. Topographical data, such as the presence of challenging natural features like mountain ranges or rivers that hinder travel between the regions, can be included. Other examples include socio-demographic information comparing the percentages of urban area between the two regions or the standardized absolute difference in a specific regional characteristic, such as the proportion of residents who smoke or even the expected age-adjusted disease count for the region. By including these covariates in $\mathbf{z}_{ij}$, the model accounts for factors that influence the neighboring relationship between regions i and j.

2.3.3 DAGAR vs CAR

When dealing with spatial random effects, the computations involving covariance matrices can be cumbersome and impracticable for large datasets. Specifically, calculating the determinant and the quadratic form in the multivariate Gaussian density are computationally expensive. In the CAR model, where the precision matrix is $\mathbf{Q} = (\mathbf{D} - \rho\mathbf{W})$, with $\mathbf{D}$ diagonal comprising the number of neighbors for each region as its elements and $\mathbf{W}$ is the full (geographic) adjacency matrix, a common trick is to compute the eigenvalues ω_i of $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. The determinant $|\mathbf{Q}|$ is expressed as $|\mathbf{Q}| \propto \prod_{i=1}^{n}(1 - \rho\omega_i)$, which significantly simplifies computing the quadratic form. If the adjacency matrix is fixed, this is efficient because the eigenvalues are computed only once and only ρ needs to be updated at each iteration of the MCMC. However, if the adjacency structure changes at each iteration, the spectral decomposition must be recomputed in every iteration of the estimation algorithm. The computational complexity for the determinant of $\mathbf{Q}$ would then be $\mathcal{O}(n^3)$ since it would involve the $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ spectral decomposition.

On the other hand, in the DAGAR model, the precision matrix is defined as $\mathbf{Q} = (\mathbf{I} - \mathbf{B})^\top\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})$ and $|\mathbf{Q}|$ is the product of the diagonal elements of $\boldsymbol{\Lambda}$. The computational complexity of the determinant reduces to $\mathcal{O}(n)$. Although $\mathbf{B}$ and $\boldsymbol{\Lambda}$ need to be updated at each iteration using the current value of $\widetilde{\mathbf{W}}$, the spectral decomposition is not required, which yields very efficient density evaluation. These advantages become even more substantial when addressing multiple diseases (see Section 2.7).

2.4 Disease dependence

Recalling that our current data analytic goals involve accounting for dependence among multiple diseases and spatial dependence for each cancer, we build upon a rich literature on multivariate areal models (Mardia, 1988; Gelfand and Vounatsou, 2003; Carlin and Banerjee, 2003; Jin et al., 2005,

2007; Zhang et al., 2009; Banerjee, 2016; MacNab, 2018; Gao et al., 2022). Given the computational benefits accrued by DAGAR detailed in Section 2.7, we build multivariate DAGAR (MDAGAR) models for γ using three different inter-disease association graphs for disease dependence.

2.4.1 Unstructured graph

We introduce dependence among outcomes using linearly transformed latent variables resulting in an unstructured graph (left panel of Figure 2.4). Therefore, $\gamma_1 = a_{11}\mathbf{f}_1$, $\gamma_d = a_{d1}\mathbf{f}_1 + \cdots + a_{dd}\mathbf{f}_d$ for each $d = 2, \ldots, q$, where a_{dh}, $h = 1, \ldots, d$, are unknown coefficients that associate spatial components for different diseases and each $\mathbf{f}_d \overset{ind}{\sim} \mathcal{N}_n(\mathbf{0}, \mathbf{Q}(\rho_d, \widetilde{\mathbf{W}}_d)^{-1})$, where ρ_d and $\widetilde{\mathbf{W}}_d$ are the spatial autocorrelation parameter and the lower triangular spatial adjacency matrix, respectively, for disease d. Each γ_d is a linear combination of independent DAGAR random variables. As demonstrated by Jin et al. (2007) for the MCAR model and extended by Gao et al. (2022) to the MDAGAR framework, this construction ensures independence from the ordering of diseases. The association among diseases is captured by $\mathbf{A}\mathbf{A}^\top$, where $\mathbf{A}$ is a lower-triangular matrix with elements a_{dh}. The prior distribution for $\mathbf{A}\mathbf{A}^\top$ is given below in equation (??). The precision matrix of γ is

$$\boldsymbol{\Sigma}_\gamma^{-1} = (\mathbf{A}^{-\top} \otimes \mathbf{I}_n) \left[\bigoplus_{d=1}^{q} \mathbf{Q}(\rho_d, \widetilde{\mathbf{W}}_d) \right] (\mathbf{A}^{-1} \otimes \mathbf{I}_n) \tag{2.6}$$

where $\otimes$ is the Kronecker product and $\widetilde{\mathbf{W}}_d$ is specified through (2.5) with a different $\boldsymbol{\eta}_d$ for each disease d and $\bigoplus_{d=1}^{q}$ defines a block diagonal matrix. Therefore, we are incorporating a random adjacency matrix into the model for each type of cancer. We evaluate the model's performance by accounting for the particular cancer under consideration, as well as the explanatory variables utilized in model (2.5) that informs the adjacency structure, potentially varying for each cancer type. Finally, we model $\mathbf{A}\mathbf{A}^\top$ using an inverse-Wishart distribution with $\mathbf{A}$ identifying as the unique lower-triangular Cholesky factor.

Alternatively, we introduce a graphical model for multivariate dependencies across diseases (see, e.g., Cox and Wermuth, 1993, 1996). Such models distinguish between "directed" and "undirected" graphs (central and right panels in Figure 2.4, respectively). The lack of edges in the undirected graph immediately reveals conditional independence $\gamma_2 \perp \gamma_4 \,|\, \{\gamma_1, \gamma_3\}$ and $\gamma_1 \perp \gamma_3 \,|\, \{\gamma_2, \gamma_4\}$. The directed acyclic graph, however, only imposes the first relation and not the second because γ_2 and γ_4 share a common parent (γ_1).

2.4.2 Directed graph

Let $\mathcal{G}_{dis} = \{\mathcal{V}_{dis}, \mathcal{E}_{dis}\}$ be a directed acyclic disease graph (the suffix *dis* distinguishes this from the graph of the spatial map), where $\mathcal{V}_{dis}$ consists of q nodes and $\mathcal{E}_{dis}$ comprises directed edges from parents to children (center panel of Figure 2.4 with $q = 4$). Let

$$\gamma_1 = \mathbf{f}_1, \quad \gamma_d = \mathbf{A}_{d1}\gamma_1 + \cdots + \mathbf{A}_{dd-1}\gamma_{d-1} + \mathbf{f}_d, \quad \text{for } d = 2, \ldots, q\,, \tag{2.7}$$

where each $\mathbf{f}_d \overset{ind}{\sim} \mathcal{N}_n(\mathbf{0}, \mathbf{Q}(\rho_d, \widetilde{\mathbf{W}}_d)^{-1})$ and $\mathbf{A}_{dd'}$ is $n \times n$ with $\mathbf{A}_{dd'} = \mathbf{O}$, the matrix with all null entries, whenever $d' \geq d$. The equations in (2.7) encode conditional distributions $p(\gamma_d \,|\, \gamma_1, \ldots, \gamma_{d-1})$

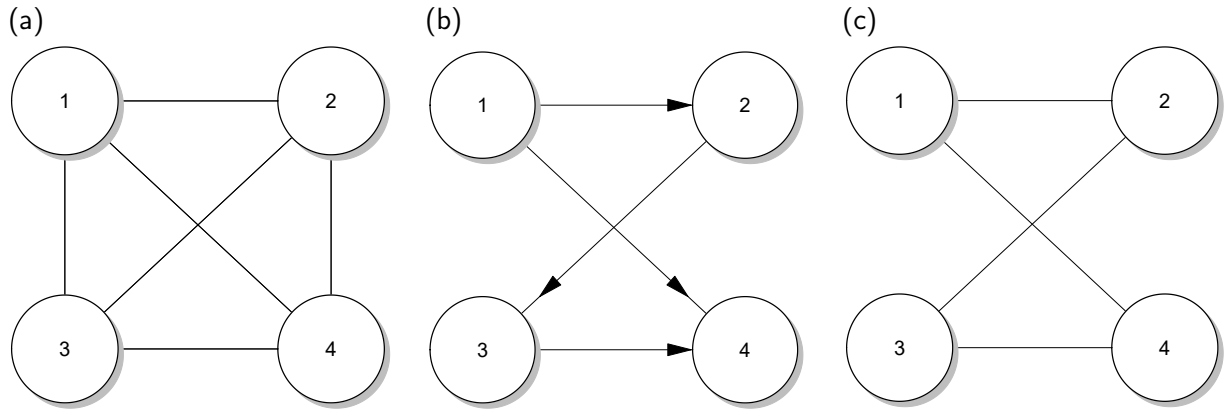


Figure 2.4: Unstructured graph (a), directed acyclic graph (b), undirected graph (c).

for $d = 2, \ldots, q$. Setting $\mathbf{A}_{dd'} = \mathbf{O}$ whenever node d' is not a parent of d, i.e., there is no directed edge from d' to d, ensures that the model conforms to the conditional independence among diseases posited in $\mathcal{G}_{dis}$. If $\mathbf{A} = (\mathbf{A}_{dd'})$ is the $N \times N$ block matrix with the $n \times n$ matrix $\mathbf{A}_{dd'}$ occupying block (d, d'), then $\mathbf{A}$ is strictly lower triangular, hence $\mathbf{I}_N - \mathbf{A}$ is nonsingular and

$$\boldsymbol{\Sigma}_{\gamma}^{-1} = (\mathbf{I}_N - \mathbf{A}^{\top}) \left[\bigoplus_{d=1}^{q} \mathbf{Q}(\rho_d, \widetilde{\mathbf{W}}_d) \right] (\mathbf{I}_N - \mathbf{A}) \tag{2.8}$$

is positive definite no matter how we specify $\mathbf{A}_{dd'}$ for all $(d, d') \in \mathcal{E}_{dis}$ (d' is a parent of d).

Setting $\mathbf{A}_{dd'} = \alpha_{0dd'}\mathbf{I}_n + \alpha_{1dd'}\mathbf{W}$, where $\mathbf{W}$ is the fixed geographic map, introduces parameters $\alpha_{0dd'}$ and $\alpha_{1dd'}$ that allow dependence between diseases d and d' to vary across space. This flexibility is desirable because inter-disease dependencies, or lack thereof, are often manifestations of lurking shared risk factors that remain unaccounted for. Finally, inference from inter-disease DAGAR without depending upon disease ordering is possible by using Bayesian model averaging over all topological orders (Gao et al., 2022), but the method is computationally demanding and not our current focus.

2.4.3 Undirected graph

In contrast to directed acyclic graphs, "undirected" graphs model relationships among nodes using conditional dependencies. An edge signifies conditional dependence given remaining nodes, while the absence of an edge indicates conditional independence given all other nodes (right panel in Figure 2.4 with $q = 4$). We specify a Markov random field using full conditional distributions,

$$\boldsymbol{\gamma}_d \,|\, \boldsymbol{\gamma}_{(-d)} \sim \mathcal{N}_n \left(\sum_{h=1}^{q} \mathbf{A}_{dh}\boldsymbol{\gamma}_h, \mathbf{Q}(\rho_d, \widetilde{\mathbf{W}}_d)^{-1} \right) , \tag{2.9}$$

where $\boldsymbol{\gamma}_{(-d)}$ is the collection of all random effects except for the d-th disease type. Each $\mathbf{A}_{dh}$ is $n \times n$ and encodes the relationships between various diseases, while $\mathbf{Q}(\rho_d, \widetilde{\mathbf{W}}_d)$ refers to the univariate precision matrix of the DAGAR model applied to the d-th cancer, capturing the spatial correlation for each cancer. By exploiting a multivariate extension of Brook's Lemma (see, e.g., Mardia, 1988; Banerjee et al., 2014), the collection of distributions in (2.9) yield a possible joint Gaussian distribution

with zero mean and precision matrix

$$\boldsymbol{\Sigma}_{\gamma}^{-1} = \left[\bigoplus_{d=1}^{q} \mathbf{Q}(\rho_d, \widetilde{\mathbf{W}}_d)\right](\mathbf{I}_N - \mathbf{A})$$

provided this matrix is symmetric and positive definite. This is achieved by carefully specifying the $N \times N$ block matrix $\mathbf{A} = (\mathbf{A}_{dh})$.

Let $\boldsymbol{\Lambda}_{dis} = (\mathbf{D}_{dis} - \rho_{dis}\mathbf{W}_{dis})$ be $q \times q$, where $\mathbf{W}_{dis}$ is the $q \times q$ binary adjacency matrix of the inter-disease graph, $\mathbf{D}_{dis}$ is diagonal with the number of edges incident on each node as its diagonal element, and $\rho_{dis} \in (1/\zeta_{min}, \zeta_{max})$ with $\zeta_{min} < 0$ and $\zeta_{max} = 1$ denoting the smallest and largest eigenvalues of $\mathbf{D}_{dis}^{-1/2}\mathbf{W}_{dis}\mathbf{D}_{dis}^{-1/2}$, respectively; all eigenvalues are real, $\zeta_{max} = 1$ since $\mathbf{D}_{dis}^{-1}\mathbf{W}_{dis}$ is row-stochastic and $\zeta_{min} < 0$ since the trace of $\mathbf{D}_{dis}^{-1/2}\mathbf{W}_{dis}\mathbf{D}_{dis}^{-1/2}$ is zero. The interval for ρ_{dis} ensures that $\boldsymbol{\Lambda}_{dis}$ is positive definite. Setting $\mathbf{U}_d = (\mathbf{I} - \mathbf{B}_d)^{\top}\boldsymbol{\Lambda}_d^{1/2}$ and $\mathbf{A}_{dh} = -\lambda_{dis,dh}\mathbf{U}_d^{-\top}\mathbf{U}_h^{\top}$ yields

$$\boldsymbol{\Sigma}_{\gamma}^{-1} = \left[\bigoplus_{d=1}^{q} \mathbf{U}_d\right](\boldsymbol{\Lambda}_{dis} \otimes \mathbf{I}_n)\left[\bigoplus_{d=1}^{q} \mathbf{U}_d^{\top}\right], \tag{2.10}$$

which is symmetric and positive definite. Since $\mathbf{U}_d\mathbf{U}_d^{\top} = \mathbf{Q}(\rho_d, \widetilde{\mathbf{W}}_d)$, we are using the square-root of the DAGAR precision $\mathbf{Q}(\rho_d, \widetilde{\mathbf{W}}_d)$ for each disease. This avoids the Cholesky decomposition. Constructing precision matrices avoids matrix inversion. The joint distribution $\boldsymbol{\gamma} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}_{\gamma})$ serves as a proper prior for the spatial effects conforming to the conditional independence posited by the undirected disease-graph.

2.4.4 Remarks on multivariate models

The constructions of $\boldsymbol{\Sigma}_{\gamma}^{-1}$ in (2.6), (2.8) and (2.10) offer some appealing features for analyzing multivariate spatial data. All three introduce spatial dependence and the latter two conform to dependencies posited by graphs. They retain DAGAR's computationally efficient structure and accommodate further modeling of the spatial adjacency matrix as in (2.5). The structure affords parsimony in storage and computation. To the practicing spatial analyst, these structures offer an inferential tool to aid them in understanding localized spatial disease patterns. This tool becomes particularly valuable when adjacencies are identified despite visible disparities. In such cases, investigators can delve deeper into these regions to explore additional contributing factors to the observed disparity.

Notably, our implementation minimizes computational burden compared to the MCAR model, by avoiding the need to invert matrices, except in the case of the unstructured graph where only a $q \times q$ lower triangular matrix inversion is required, and by simplifying a lot the computation of determinants. In most applications, typically, the value of q is not an issue for computational efficiency. Section 2.7 offer further details on computational benefits of these approaches.

2.5 Model implementation

Letting $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^{\top}, \ldots, \boldsymbol{\eta}_q^{\top})^{\top}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\top}, \ldots, \boldsymbol{\beta}_q^{\top})^{\top}$ for the unstructured disease graph we evaluate the following joint posterior distribution,

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}, \tau, \mathbf{V}, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\eta}, \mathbf{A} \,|\, \mathbf{y}) \propto p(\boldsymbol{\theta}, \boldsymbol{\beta}, \tau, \mathbf{V}, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\eta}, \mathbf{A}) \times \prod_{i=1}^{n} \prod_{d=1}^{q} p(y_{id} \,|\, \boldsymbol{\beta}_d, \phi_{id}) \,, \tag{2.11}$$

where $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_q^\top)^\top$ is $N \times 1$ with $N = nq$, n is the number of geographic regions, and $\mathbf{y}_d = (y_{1d}, \ldots, y_{nd})^\top$ and each y_{id} being modeled independently conditional on the spatial random effects using a member from the exponential family. For example, in our simulation experiments in Section 2.8 we assume $y_{id} \,|\, \boldsymbol{\beta}_d, \phi_{id}, \tau_d \overset{ind}{\sim} \mathcal{N}(\mathbf{x}_{id}^\top \boldsymbol{\beta}_d + \phi_{id}, 1/\tau_d)$ and specify the joint prior distribution, $p(\boldsymbol{\beta}, \tau_d, \boldsymbol{\theta}, \mathbf{V}, \tau, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\eta}, \mathbf{A})$ proportional to

$$\begin{aligned} &\mathcal{N}_{qp}(\boldsymbol{\beta} \,|\, \mathbf{0}, 1/\tau_\beta \mathbf{I}_{qp}) \times \prod_{d=1}^{q} \mathsf{IG}(1/\tau_d \,|\, a_\epsilon, b_\epsilon) \times \prod_{k=1}^{K} \{\mathcal{N}(\theta_k \,|\, 0, 1/\tau) \times \mathsf{Beta}(V_k \,|\, 1, \alpha)\} \\ &\quad \times \mathsf{IG}(1/\tau \,|\, a, b) \times \mathcal{N}_N(\boldsymbol{\gamma} \,|\, \mathbf{0}, \boldsymbol{\Sigma}_\gamma) \times \mathcal{W}^{-1}(\mathbf{A}\mathbf{A}^\top \,|\, \nu, \boldsymbol{\Psi}) \times \left| \frac{\partial \mathbf{A}\mathbf{A}^\top}{\partial a_{dh}} \right| \\ &\quad \times \prod_{d=1}^{q} \left\{ \prod_{r=1}^{R_d} \mathcal{U}(\eta_{dr} \,|\, 0, M_r) \times \mathcal{U}(\rho_d \,|\, 0, 1) \right\} \,, \end{aligned} \tag{2.12}$$

where $\left| \frac{\partial \mathbf{A}\mathbf{A}^\top}{\partial a_{dh}} \right| = 2^q \prod_{d=1}^{q} a_{dd}^{q-d+1}$ is the Jacobian transformation for the prior on $\mathbf{A}\mathbf{A}^\top$ in terms of the Cholesky factor $\mathbf{A}$ (Olkin and Sampson, 1972), and R_d is the dimension of $\boldsymbol{\eta}_d$. We employ Markov Chain Monte Carlo (MCMC) with Gibbs sampling and random walk Metropolis implemented in `RcppArmadillo` (Eddelbuettel and Sanderson, 2014) statistical computing environment, to draw samples from the posterior distribution, as defined in (2.11).

The different graphical models discussed above provide parametric structures for $\mathbf{A}$. Therefore, the $\mathcal{W}(\mathbf{A}\mathbf{A}^\top \,|\, \cdot, \cdot)$ in (2.12) is replaced with priors on the parameters in $\mathbf{A}$. In the directed graph, we further model $\mathbf{A}_{dh}$ in terms of parameters $\{\alpha_{0dd'}\}$ and $\alpha_{1dd'}$. These parameters are assigned independent Gaussian prior distributions $\alpha_{0dd'} \overset{ind}{\sim} \mathcal{N}(\cdot, \cdot)$ and $\alpha_{1dd'} \overset{ind}{\sim} \mathcal{N}(\cdot, \cdot)$, respectively, resulting in a joint prior distribution for $\boldsymbol{\alpha} = (\alpha_{0dd'}, \alpha_{1dd'})_{dd'}$, i.e., $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, 100\mathbf{I})$. In the undirected model, the unknown parameter $\rho_{dis} \overset{ind}{\sim} \mathcal{U}(\cdot, \cdot)$ completely specifies $\mathbf{A}$. Further insights and computational details are supplied in Section B.1 in the Appendices.

2.6 Difference boundaries through FDR

We treat spatial boundary analysis as an exercise in multiple hypothesis testing, although a formal null hypothesis is not required in Bayesian inference. Instead, we pursue full stochastic quantification to derive a threshold for detecting cancer-specific spatial disparities. For each pair of adjacent regions, i.e., $i \sim j$, and each cancer d, we seek the posterior probability of $\phi_{id} = \phi_{jd'}$ and its complement $\phi_{id'} \neq \phi_{jd'}$ for region-disease pairs (i, d) and (j, d'). We designate edge (i, j) as a difference boundary if the posterior probability of $\phi_{id} \neq \phi_{jd'}$ given $\mathbf{y}$ surpasses a predefined threshold t.

The most general difference boundary we report is the *cross-difference* boundary based upon $v_{(i,d)(j,d')} = P(\phi_{id} \neq \phi_{jd'}, \phi_{id'} \neq \phi_{jd} \,|\, \mathbf{y})$. We derive a threshold t that controls the false discovery

rate (FDR) below a level $\zeta = 0.05$. To this end, we define

$$\text{FDR}_{d,d'}(t) = \frac{\sum_{i\sim j} I(\phi_{id} = \phi_{jd'}) I(v_{(i,d)(j,d')} > t)}{\sum_{i\sim j} I(v_{(i,d)(j,d')} > t)} \tag{2.13}$$

every disease pair d and d'. The quantity in (2.13) is evaluated as

$$\widehat{\text{FDR}}_{d,d'}(t) = \mathbb{E}[\text{FDR}_{d,d'} \mid \mathbf{y}] = \frac{\sum_{i\sim j}(1 - v_{(i,d)(j,d')}) I(v_{(i,d)(j,d')} > t)}{\sum_{i\sim j} I(v_{(i,d)(j,d')} > t)} \,. \tag{2.14}$$

Following Müller et al. (2004), we define

$$t^* = \sup\left\{ t : \widehat{\text{FDR}}_{d,d'}(t) \leq \zeta \right\} \,, \tag{2.15}$$

which is based upon the optimal decision that minimizes the estimated false negative rate (FNR), $\widehat{\text{FNR}}_{d,d'}(t) = \frac{\sum_{i\sim j} v_{(i,d)(j,d')}(1 - I(v_{(i,d)(j,d')} > t))}{m - \sum_{i\sim j} I(v_{(i,d)(j,d')} > t)}$, where m is the total number of geographic boundaries, subject to $\widehat{\text{FDR}} \leq \zeta$ (also see Sun et al., 2015, who proffer a similar approach). This estimation is based on a bivariate loss function $L_{2R} = (\widehat{\text{FDR}}, \widehat{\text{FNR}})$. The single disease setting is obtained by substituting $v_{i,j}^{(d)} = P(\phi_{id} \neq \phi_{jd'}, \phi_{id'} \neq \phi_{jd} \mid \mathbf{y})$ in the above expressions. This is analogous to Li et al. (2015), but with the added flexibility of modeling the adjacency matrices and inter-disease dependence models.

2.7 Computational details

In this section, we provide the computational details of the MDAGAR model, specifically within the context of the multiple-disease framework. Our model allows for the following evaluation of the multivariate Gaussian log-density:

$$\log\left(p(\boldsymbol{\gamma} \mid \boldsymbol{\Sigma}_\gamma)\right) \propto \log\left(|\boldsymbol{\Sigma}_\gamma|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{\gamma} \right) \right) \propto \log|\boldsymbol{\Sigma}_\gamma^{-1}| - \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{\gamma}$$

To simplify this expression, we leverage the structure of the MDAGAR model in three different cases: the unstructured disease graph (Section 2.7.1), the directed disease graph (Section 2.7.2), and the undirected disease graph (Section 2.7.3). For each case, we derive the simplified forms of the log-density and highlight the computational advantages of the proposed approaches (Section 2.7.4).

2.7.1 Unstructured graph

We first need to compute $|\boldsymbol{\Sigma}_\gamma^{-1}|$:

$$|\boldsymbol{\Sigma}_\gamma^{-1}| = |\mathbf{A}^\top \otimes \mathbf{I}|^{-1} \prod_{d=1}^{q} |\mathbf{Q}_d| |\mathbf{A} \otimes \mathbf{I}|^{-1} = |\mathbf{A}|^{-2n} \prod_{d=1}^{q} |\mathbf{Q}_d| = \left(\prod_{d=1}^{q} \mathbf{A}_{dd} \right)^{-2n} \prod_{d=1}^{q} \prod_{i=1}^{n} (\boldsymbol{\Lambda}_d)_{ii}$$

where $|\mathbf{Q}_d| = |\mathbf{I} - \mathbf{B}_d|^2 |\boldsymbol{\Lambda}_d| = |\boldsymbol{\Lambda}_d|$ is motivated by the strictly lower-triangularity of $\mathbf{B}_d$ and $|\mathbf{A}| = \prod_{d=1}^{q} \mathbf{A}_{dd}$ by the lower-triangularity of $\mathbf{A}$. Then we need to compute $\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{\gamma}$:

$$\gamma^\top \boldsymbol{\Sigma}_\gamma^{-1}\gamma = \gamma^\top \left(\mathbf{A}^{-1}\otimes \mathbf{I}\right)^\top \left[\bigoplus_d^q \mathbf{Q}_d\right]\left(\mathbf{A}^{-1}\otimes \mathbf{I}\right)\gamma = \left[\left(\mathbf{A}^{-1}\otimes \mathbf{I}\right)\gamma\right]^\top \left[\bigoplus_d^q \mathbf{Q}_d\right]\left[\left(\mathbf{A}^{-1}\otimes \mathbf{I}\right)\gamma\right]$$

Here we need some algebra to simplify the expression. Let's start by defining $\boldsymbol{\Gamma} = [\gamma_1 \mid \cdots \mid \gamma_q]$ and by exploiting some properties of the Kronecker product:

$$\left(\mathbf{A}^{-1}\otimes \mathbf{I}\right)\gamma = \left(\mathbf{A}^{-1}\otimes \mathbf{I}\right)\text{vec}\left(\boldsymbol{\Gamma}\right) = \text{vec}\left(\mathbf{I}\boldsymbol{\Gamma}\mathbf{A}^{-\top}\right) = \text{vec}\left(\boldsymbol{\Gamma}\mathbf{A}^{-\top}\right) = \text{vec}\left(\mathbf{C}\right)$$

Here, matrix $\mathbf{C}$ is computed as:

$$\begin{aligned}\mathbf{C} &= [\gamma_1 \mid \gamma_2 \mid \cdots \mid \gamma_q]\begin{bmatrix}\widetilde{a}_{11} & \widetilde{a}_{21} & \cdots & \widetilde{a}_{q1}\\ 0 & \widetilde{a}_{22} & \cdots & \widetilde{a}_{q2}\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \widetilde{a}_{qq}\end{bmatrix} = \left[\widetilde{a}_{11}\gamma_1 \mid \widetilde{a}_{21}\gamma_1 + \widetilde{a}_{22}\gamma_2 \mid \cdots \mid \sum_{d=1}^q \widetilde{a}_{qd}\gamma_d\right]\\ &= [\mathbf{c}_1 \mid \mathbf{c}_2 \mid \cdots \mid \mathbf{c}_q]\end{aligned}$$

where $\widetilde{a}_{dh}$ are the elements of the lower triangular matrix $\mathbf{A}^{-1}$. Hence,

$$\begin{aligned}\gamma^\top \boldsymbol{\Sigma}_\gamma^{-1}\gamma &= \left[\mathbf{c}_1^\top, \mathbf{c}_2^\top, \ldots, \mathbf{c}_q^\top\right]\begin{bmatrix}\mathbf{Q}_1 & \mathbf{O} & \cdots & \mathbf{O}\\ \mathbf{O} & \mathbf{Q}_2 & \cdots & \mathbf{O}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{Q}_q\end{bmatrix}\begin{bmatrix}\mathbf{c}_1\\ \mathbf{c}_2\\ \vdots\\ \mathbf{c}_q\end{bmatrix}\\ &= \sum_{d=1}^q \mathbf{c}_d^\top \mathbf{Q}_d \mathbf{c}_d\end{aligned}$$

The log-density of γ finally results in:

$$\log\left(p(\gamma \mid \boldsymbol{\Sigma}_\gamma)\right) \propto -n\sum_{d=1}^q \log\left(\mathbf{A}_{dd}\right) + \frac{1}{2}\sum_{d=1}^q\sum_{i=1}^n \log\left((\boldsymbol{\Lambda}_d)_{ii}\right) - \frac{1}{2}\sum_{d=1}^q \mathbf{c}_d^\top \mathbf{Q}_d \mathbf{c}_d \tag{2.16}$$

This log-density evaluation is computationally efficient because it requires only one matrix inversion, which pertains to $\mathbf{A}$. In applications similar to this work, the dimensions of $\mathbf{A}$ are typically small ($q \leq 10$).

2.7.2 Directed graph

In this case the determinant of $\boldsymbol{\Sigma}_\gamma^{-1}$ is very simple:

$$|\boldsymbol{\Sigma}_\gamma^{-1}| = |\mathbf{I}_N - \mathbf{A}|\prod_{d=1}^q |\mathbf{Q}_d||\mathbf{I}_N - \mathbf{A}| = \prod_{d=1}^q |\mathbf{Q}_d| = \prod_{d=1}^q\prod_{i=1}^n (\boldsymbol{\Lambda}_d)_{ii}$$

where we have exploited the strictly lower triangularity of the $N \times N$ block matrix $\mathbf{A}$. Now we need to compute $\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{\gamma}$:

$$\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{\gamma} = \boldsymbol{\gamma}^\top \left(\mathbf{I}_N - \mathbf{A}\right)^\top \left[\bigoplus_{d=1}^{q} \mathbf{Q}_d\right] \left(\mathbf{I}_N - \mathbf{A}\right) \boldsymbol{\gamma}$$

where

$$\left(\mathbf{I}_N - \mathbf{A}\right) \boldsymbol{\gamma} = \begin{bmatrix} \mathbf{I}_n & \mathbf{O} & \cdots & \mathbf{O} \\ -\mathbf{A}_{21} & \mathbf{I}_n & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{A}_{q1} & -\mathbf{A}_{q2} & \cdots & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_q \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 - \mathbf{A}_{21}\boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_q - \sum_{h=1}^{q-1} \mathbf{A}_{qh}\boldsymbol{\gamma}_h \end{bmatrix}$$

Hence

$$\begin{aligned} \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{\gamma} &= \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 - \mathbf{A}_{21}\boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_q - \sum_{h=1}^{q-1} \mathbf{A}_{qh}\boldsymbol{\gamma}_h \end{bmatrix}^\top \begin{bmatrix} \mathbf{Q}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{Q}_2 & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{Q}_q \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 - \mathbf{A}_{21}\boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_q - \sum_{h=1}^{q-1} \mathbf{A}_{qh}\boldsymbol{\gamma}_h \end{bmatrix} \\ &= \sum_{d=1}^{q} \left[\left(\boldsymbol{\gamma}_d - \sum_{h=1}^{d-1} \mathbf{A}_{dh}\boldsymbol{\gamma}_h \right)^\top \mathbf{Q}_d \left(\boldsymbol{\gamma}_d - \sum_{h=1}^{d-1} \mathbf{A}_{dh}\boldsymbol{\gamma}_h \right) \right] \end{aligned}$$

In this case the log-density is defined as

$$\log\left(p(\boldsymbol{\gamma} \mid \boldsymbol{\Sigma}_\gamma)\right) \propto \sum_{d=1}^{q} \sum_{i=1}^{n} \log\left((\boldsymbol{\Lambda}_d)_{ii}\right) - \sum_{d=1}^{q} \left[\left(\boldsymbol{\gamma}_d - \sum_{h=1}^{d-1} \mathbf{A}_{dh}\boldsymbol{\gamma}_h \right)^\top \mathbf{Q}_d \left(\boldsymbol{\gamma}_d - \sum_{h=1}^{d-1} \mathbf{A}_{dh}\boldsymbol{\gamma}_h \right) \right] \quad (2.17)$$

where, it is clear that there is no need to invert any matrices.

2.7.3 Undirected graph

Again, the first thing we need to compute is $|\boldsymbol{\Sigma}_\gamma^{-1}|$:

$$|\boldsymbol{\Sigma}_\gamma^{-1}| = \left| \bigoplus_{d=1}^{q} \mathbf{U}_d \right| |\boldsymbol{\Lambda}_{dis} \otimes \mathbf{I}_n| \left| \bigoplus_{d=1}^{q} \mathbf{U}_d^\top \right| = |\boldsymbol{\Lambda}_{dis}|^n \left(\prod_{d=1}^{q} |\mathbf{U}_d| \right)^2 = |\boldsymbol{\Lambda}_{dis}|^n \prod_{d=1}^{q} \prod_{i=1}^{n} (\boldsymbol{\Lambda}_d)_{ii}$$

where for the computation of $|\mathbf{U}_d|$ we have exploited $\mathbf{U}_d = \left(\mathbf{I} - \mathbf{B}_d\right)^\top \boldsymbol{\Lambda}_d^{1/2}$ and consequently $|\mathbf{U}_d| = |\mathbf{I} - \mathbf{B}_d||\boldsymbol{\Lambda}_d^{1/2}| = \prod_{i=1}^{n} \sqrt{(\boldsymbol{\Lambda}_d)_{ii}}$. Then, it is the turn of $\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{\gamma}$:

$$\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{\gamma} = \boldsymbol{\gamma}^\top \left[\bigoplus_{d=1}^{q} \mathbf{U}_d \right] (\boldsymbol{\Lambda}_{dis} \otimes \mathbf{I}_n) \left[\bigoplus_{d=1}^{q} \mathbf{U}_d^\top \right] \boldsymbol{\gamma} = \left(\left[\bigoplus_{d=1}^{q} \mathbf{U}_d^\top \right] \boldsymbol{\gamma} \right)^\top (\boldsymbol{\Lambda}_{dis} \otimes \mathbf{I}_n) \left(\left[\bigoplus_{d=1}^{q} \mathbf{U}_d^\top \right] \boldsymbol{\gamma} \right)$$

In this case, to simplify the expression we need to develop the computations regarding

$$\left[\bigoplus_{d=1}^{q}\mathbf{U}_d^\top\right]\boldsymbol{\gamma} = \begin{bmatrix}\mathbf{U}_1^\top & \mathbf{O} & \cdots & \mathbf{O}\\ \mathbf{O} & \mathbf{U}_2^\top & \cdots & \mathbf{O}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{U}_q^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\gamma}_1\\ \boldsymbol{\gamma}_2\\ \vdots\\ \boldsymbol{\gamma}_q\end{bmatrix} = \begin{bmatrix}\mathbf{U}_1^\top\boldsymbol{\gamma}_1\\ \mathbf{U}_2^\top\boldsymbol{\gamma}_2\\ \vdots\\ \mathbf{U}_q^\top\boldsymbol{\gamma}_q\end{bmatrix}$$

Hence,

$$\begin{aligned}\boldsymbol{\gamma}^\top\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma} &= \left[\boldsymbol{\gamma}_1^\top\mathbf{U}_1, \boldsymbol{\gamma}_2^\top\mathbf{U}_2, \dots, \boldsymbol{\gamma}_q^\top\mathbf{U}_q\right]\begin{bmatrix}\lambda_{dis,11}\mathbf{I} & \lambda_{dis,12}\mathbf{I} & \cdots & \lambda_{dis,1q}\mathbf{I}\\ \lambda_{dis,21}\mathbf{I} & \lambda_{dis,22}\mathbf{I} & \cdots & \lambda_{dis,2q}\mathbf{I}\\ \vdots & \vdots & \ddots & \vdots\\ \lambda_{dis,q1}\mathbf{I} & \lambda_{dis,q2}\mathbf{I} & \cdots & \lambda_{dis,qq}\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{U}_1^\top\boldsymbol{\gamma}_1\\ \mathbf{U}_2^\top\boldsymbol{\gamma}_2\\ \vdots\\ \mathbf{U}_q^\top\boldsymbol{\gamma}_q\end{bmatrix}\\ &= \sum_{d=1}^{q}\sum_{h=1}^{q}\lambda_{dis,dh}\boldsymbol{\gamma}_d^\top\mathbf{U}_d\mathbf{U}_h^\top\boldsymbol{\gamma}_h\end{aligned}$$

Finally, also in this case we exploit the DAGAR construction of each spatial precision matrix, so no matrix inversion is required since the log-density results in:

$$\log\left(p(\boldsymbol{\gamma}\,|\,\boldsymbol{\Sigma}_\gamma)\right) \propto \sum_{d=1}^{q}\sum_{i=1}^{n}\log\left((\boldsymbol{\Lambda}_d)_{ii}\right) + n\log|\boldsymbol{\Lambda}_{dis}| - \sum_{d=1}^{q}\sum_{h=1}^{q}\lambda_{dis,dh}\boldsymbol{\gamma}_d^\top\mathbf{U}_d\mathbf{U}_h^\top\boldsymbol{\gamma}_h \tag{2.18}$$

2.7.4 Advantages with respect to the MCAR specification

In all the aforementioned log-density specifications (2.16), (2.17) and (2.18), the log-determinant of the spatial precision matrix, $|\mathbf{Q}_d|$, was simplified to the summation of the diagonal elements of the diagonal matrix $|\boldsymbol{\Lambda}_d|$, resulting, for all the d, in a computational complexity of $\mathcal{O}(qn)$, due to the DAGAR construction. However, this simplification does not apply to an MCAR specification, i.e., $\mathbf{Q}_d = (\mathbf{D}_d - \rho_d\mathbf{W}_d)$, where $\mathbf{D}_d$ is a diagonal matrix with elements equal to the number of neighbors for each region. For MCAR, the precision matrix determinant cannot be simplified without computing the eigenvalues ω_{di} of $\mathbf{D}_d^{-1/2}\mathbf{W}_d\mathbf{D}_d^{-1/2}$, from which $|\mathbf{Q}_d| \propto \prod_{i=1}^{n}(1-\rho_d\omega_{di})$.

In a framework with a fixed adjacency structure, this is not problematic because the eigenvalues can be precomputed. At each iteration of the MCMC, only ρ_d would need to be updated. However, if the adjacency structure changes at every iteration, the eigendecomposition would need to be recomputed each time, negating the computational advantage of precomputation. In this case, for all the d, the computational complexity is $\mathcal{O}(qn^3)$.

Only in the case of the undirected graph, the computational complexity associated with the quadratic form differs. Specifically, for the MDAGAR, we can leverage the construction method of each $\mathbf{Q}_d$, leading to a computational complexity of $\mathcal{O}(q^2n^2)$. However, for the MCAR variant, Cholesky factorization of each $\mathbf{Q}_d$ is necessary, resulting in a computational complexity of $\mathcal{O}(qn^3 + q^2n^2)$, where the first term represents the additional complexity introduced by the Cholesky decomposition. In all the other cases also the computational complexity associated to the MCAR construction of $\mathbf{Q}_d$ is $\mathcal{O}(q^2n^2)$.

2.8 Simulation experiments

We generate Gaussian data for $q = 4$ diseases using (2.1) with $y_{id} \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}_d + \phi_{id}, 1/\tau_d^2)$ over $N = 58$ counties in California. At each replicate the data are generated after simulating spatial effects $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^\top, \boldsymbol{\phi}_2^\top, \boldsymbol{\phi}_3^\top, \boldsymbol{\phi}_4^\top)^\top \sim G_N$ with parameters fixed as $K = 15$, $\alpha = 1$, $\beta_1 = (2, 1)^\top$, $\beta_2 = (1, 2)^\top$, $\beta_3 = (2, 2)^\top$, $\beta_4 = (1, 2)^\top$, $\tau_d^2 = 10$ for every d and $\tau = 0.25$. Each $\mathbf{x}_i$ is 2×1 consisting of a 1 and a value generated from a standard normal distribution. We generate 100 different replicates of data using the above parameters and generating the spatial effects ϕ_{id} using distributions corresponding to each of the 3 graphs illustrated in Figure 2.4: (i) unstructured, (ii) directed, and (iii) undirected. These correspond to the three specifications for $\boldsymbol{\Sigma}_\gamma$ in (2.6), (2.8) and (2.10). For each of the above models, $\boldsymbol{\Sigma}_\gamma$ depends upon the DAGAR parameters and the inter-disease covariances. Spatial correlation parameters for the different cancers are fixed as $\boldsymbol{\rho} = (0.2, 0.8, 0.4, 0.6)^\top$. For the lower triangular adjacency matrices, $\widetilde{\mathbf{W}}_d$, the elements z_{ij} of the $\mathbf{Z}$ matrix are computed as the standardized absolute difference of a generated covariate $|x_i - x_j|/\sigma$ where σ represents the standard deviation of all the absolute differences between the x_i values across all pairs of contiguous areas. The unknown parameters associated with these variables were fixed at $\boldsymbol{\eta} = (0.4, 0.2, 0.3, 0.5)^\top$. In addition to specifying the DAGAR parameters for each disease, for (i) we specify $\mathbf{A}$ as lower triangular with all elements set to 1. For (ii) we fix $(\alpha_{021}, \alpha_{121})^\top = (0.3, 0.5)^\top$, $(\alpha_{041}, \alpha_{141})^\top = (0.5, 0.4)^\top$, $(\alpha_{032}, \alpha_{132})^\top = (0.4, 0.4)^\top$, and $(\alpha_{043}, \alpha_{143})^\top = (0.8, 0.1)^\top$, and for (iii) we fix $\rho_{dis} = 0.25$. For each of these specifications, we produce a single instance of $\boldsymbol{\Sigma}_\gamma$, which is used to generate a different instance of $\boldsymbol{\gamma}$ at every replication. This yields the 100 sets of random effects used to generate the 100 datasets.

For data analysis, we specify the prior using $\tau_\beta = 0.001$, $a = 2$, $b = 0.1$, $a_\epsilon = 2$, $b_\epsilon = 0.1$, $\alpha = 1$, and $M = -\log(0.5)/\mathbf{Z}_{0.5}$, where $\mathbf{Z}_{0.5}$ represents the 0.5th quantile of the $\mathbf{Z}$ matrix. For the unstructured or unconstrained inter-disease graph (left panel in Figure 2.4) we use the Wishart distribution in (2.12) with $\nu = 2$ and $\boldsymbol{\Psi} = \text{diag}(0.1, 0.1, 0.1, 0.1)$. For the directed graph (center panel in Figure 2.4), $\mathbf{A}$ is determined through $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, 100\mathbf{I})$, while for the undirected graph (right panel in Figure 2.4) we specify $\rho_{dis} \sim \mathcal{U}(-1, 1)$. We fit all three models to each of the $3 \times 100 = 300$ datasets and present inference based on 10,000 posterior samples using MCMC algorithms after discarding the initial 20,000 iterations as burn-in.

Each data is analyzed by one correctly specified model and two misspecified models. We evaluate these models under correct and incorrect specifications by computing the Widely Applicable Information Criterion (Watanabe and Opper, 2010; Vehtari et al., 2017) for each model using the `loo R` package (Vehtari et al., 2024). This score measures the extent of information loss in a model with lower values indicating better performance by effectively addressing the risks of both overfitting and underfitting. Table 2.1 presents the relative differences in WAIC scores. For each disease graph model, we compute the standardized relative difference between the mean WAIC of the misspecified models and the true model, normalized by the mean WAIC of the true model. Not surprisingly, the model corresponding to the true data generating scheme has the lowest WAIC scores (denoted as 0% in standardized relative difference). However, Table 2.1 reveals substantially worse performance (higher WAIC) by the graphical models when analyzing data from the unstructured model. The scores for the two graphical models are very comparable to one another when fitted to the unstructured data. On the other hand, for data generated from a graphical model, the unstructured model excels

Table 2.1: WAIC relative differences for the 3×3 simulation grid.

True/Fit	Unstructured	Directed	Undirected
Unstructured	0.00	+6.12	+6.63
Directed	+1.28	0.00	+21.25
Undirected	+1.65	+25.62	0.00

Table 2.2: Boundary detection results in the simulation study under the three disease graph models using the unstructured graph to fit.

		Disease 1		Disease 2		Disease 3		Disease 4	
True Disease graph	T	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
Unstructured	85	0.553	0.868	0.513	0.891	0.538	0.885	0.544	0.880
	90	0.503	0.882	0.454	0.906	0.482	0.900	0.490	0.893
	95	0.433	0.906	0.402	0.921	0.428	0.913	0.434	0.904
	100	0.383	0.921	0.342	0.932	0.359	0.930	0.379	0.918
	105	0.329	0.931	0.297	0.940	0.296	0.942	0.330	0.929
Directed	85	0.575	0.863	0.461	0.918	0.382	0.936	0.434	0.928
	90	0.511	0.885	0.414	0.926	0.334	0.942	0.391	0.937
	95	0.452	0.904	0.360	0.941	0.297	0.949	0.336	0.946
	100	0.399	0.916	0.300	0.952	0.252	0.959	0.289	0.953
	105	0.338	0.933	0.243	0.964	0.215	0.967	0.241	0.964
Undirected	85	0.574	0.848	0.590	0.849	0.603	0.854	0.620	0.848
	90	0.508	0.863	0.531	0.870	0.542	0.875	0.567	0.869
	95	0.450	0.883	0.463	0.888	0.477	0.896	0.499	0.888
	100	0.401	0.899	0.405	0.904	0.420	0.910	0.431	0.902
	105	0.339	0.913	0.345	0.918	0.364	0.925	0.365	0.923

over the misspecified graphical model.

Turning to boundary detection, we compute the conditional probability, $P(\phi_{id} \neq \phi_{jd'} \,|\, \mathbf{y})$, for diseases d and d' and geographically neighboring regions i and j. We use these posterior probabilities to obtain sensitivity and specificity for all pairs of regions and for all the disease graph models. In each simulation, sensitivity and specificity are determined by selecting a fixed number of edges with the highest T posterior probabilities, where $\mathrm{T} \in \{85, 90, 95, 100, 105\}$. This method accounts for false positives across all other T values for the diseases. Table 2.2 presents the corresponding results for the unstructured model fitted with data generated from the true model (the unstructured model) and the misspecified models (the directed and undirected models). The values in Table 2.2 represent average sensitivity and specificity over 100 simulated data sets using the unstructured disease graph for model fitting under three different choices of the true disease graph.

With T set to 105, we achieve sensitivities of approximately 90% for all diseases. Despite introducing uncertainty into the adjacency matrix and adopting a much more complex simulation setting compared to Gao et al. (2023), we attain similar levels of sensitivity. This added uncertainty in the adjacency structure propagates through the model leading to a notably broader posterior distribution for spatial random effects. A wider posterior distribution indicates a greater incidence of ties between random effects across different cancers and regions. Consequently, with the detection of more boundaries, specificity decreases.

With regard to the adjacency model in (2.5), Table 2.3 presents specificity and sensitivity values for detecting adjacencies. Here, sensitivity and specificity are evaluated by comparing the adjacencies determined by the true value of $\boldsymbol{\eta}$ with estimated adjacencies using the posterior mean of $\boldsymbol{\eta}$ in (2.5). It appears that higher sensitivity values are achieved across all diseases and disease graph models, while specificity values, though slightly lower, remain at satisfactory levels. The high sensitivity

Table 2.3: Adjacencies detection for the three disease graph models using the unstructured graph to fit.

	Disease 1		Disease 2		Disease 3		Disease 4	
True Disease graph	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
Unstructured	0.386	0.986	0.598	0.928	0.600	0.959	0.261	1.000
Directed	0.340	0.988	0.604	0.919	0.645	0.957	0.264	0.999
Undirected	0.287	0.985	0.671	0.913	0.576	0.975	0.258	1.000

coupled with lower specificity suggests that our model excels in identifying true positives, but may occasionally incorrectly classify negative cases as positive (false positives). While false positives are undesirable, they do not pose a significant issue for our application, particularly considering the fact that specificity levels are not unreasonably low. This is especially relevant in the context of spatial health disparities, where policy decisions are predominantly influenced by true positives.

We conclude this section with a brief remark. For analyzing our data over the map of California, computational costs for the MDAGAR model averaged approximately 0.006 seconds per iteration for the unstructured disease graph, 0.013 for the directed and 0.009 for the unstructured. In comparison, the MCAR took about 0.043 iterations per seconds for the unstructured disease graph, 0.049 for the directed and 0.045 for the undirected graph. In summary, MDAGAR is substantially more efficient (gains of 86%, 73% and 80% for the unstructured, directed and undirected, respectively) than MCAR while offering very similar inference for boundary detection. Section B.2 in the Appendices offers additional results from the simulations including root median squared error, sensitivity and specificity tables for models fitted under the true model and their WAIC values.

2.9 SEER cancers analysis

We analyze the data set introduced in Section 2.1 using a Poisson spatial regression model. Specifically, we model the observed counts of incidence for each county and for each cancer as $y_{id} \,|\, \beta_d, \phi_{id} \overset{ind}{\sim} \text{Pois}(E_{id} \exp(\beta_d + \phi_{id}))$ for $i = 1, \ldots, 58$ and $d = 1, \ldots, 4$. Following the joint prior distribution in (2.12) (excluding τ_d as we use a Poisson likelihood) with hyper-parameter values set as described in Section 2.8, we employ the unstructured model with $\mathbf{\Sigma}_\gamma^{-1}$ in (2.6) since a graphical model prior for cancer dependence is lacking. Inference is conducted using 300,000 total iterations of the MCMC algorithm. After discarding the initial 50,000 iterations as burn-in, we retain every 5th sample from the remaining 250,000 iterations for analysis. Throughout this Section, it is worth noting, we comment on the results by directly referring to specific counties by their names. For a visual reference of the corresponding county names, we refer the reader to Appendix B and the map shown in Figure B.7. Moreover, Section B.3 in the Appendices includes plots for the 95% credible intervals.

We employ a threshold to control for FDR as specified in (2.15) for detecting difference boundaries for each cancer. Figure 2.5 plots estimated FDR against varying numbers of selected edges as difference boundaries for the four cancers individually. We observe analogous FDR curves for esophageal and larynx cancers leading us to select a similar number of boundaries for these two cancers with a threshold of $\zeta = 0.05$. We detect a much larger number of boundaries for lung cancer and a much smaller number for colorectal cancer using the same threshold indicating considerable variation in spatial disparities among these cancers.

Setting $\zeta = 0.05$ in (2.15), Figure 2.6 illustrates the identified difference boundaries (highlighted

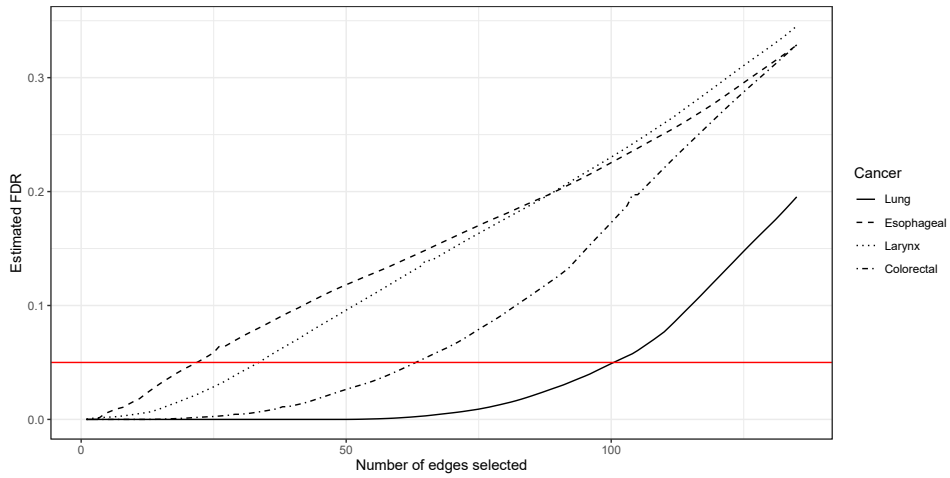


Figure 2.5: Estimated FDR curves plotted against the number of selected difference boundaries for four cancers

in blue) in the California maps for the four types of cancers, colored according to the posterior means of the corresponding ϕ_{id}. The width of the lines on the boundaries signify higher probabilities of detection. Notably, lung cancer exhibits the highest number of detected boundaries, i.e., 100, with colorectal cancer demonstrating a much higher number of boundaries, i.e., 63, compared to esophageal and larynx cancer with 21 and 33 boundaries respectively. It is worth noting that the detected boundaries for the different cancers tend to form geographical groups, which include regions with similar random effects. This was made possible by the random effect specification described in Section 2.2, which explains how the chosen discrete distribution for the random effects accounts for ties and facilitates spatial aggregation through the weights in equation (2.2). In contrast, using a continuous distribution with spatial dependence for the random effects would introduce greater smoothness, making the evaluation of boundaries more challenging and less direct.

In northern California, for example, we observe that 10 counties (Tehama, Glenn, Butte, Yuba, Lake, Colusa, Shasta, Trinity, Humboldt and Del Norte) form a "group" with high posterior medians of spatial effects for lung cancer, where "group" indicates that these counties do not exhibit disparities (or difference boundaries) among themselves but do so with other neighbors. A similar grouping appears in the central regions for Solano (Bay Area), Sacramento and San Joaquin, with elevated posterior random effect estimates for lung cancer. Moving to central California, we find Stanislaus, Merced, Tuolumne, Mariposa, Madera and Fresno counties forming a group with moderately elevated spatial effects while not exhibiting disparities among them. Colorectal cancer offers a similar story in the central counties. In contrast, maps for esophageal and larynx cancers reveal different patterns. Identifying spatial patterns for larynx and esophageal cancer is more challenging and the detected boundaries are less pronounced compared to lung cancer. However, the model identifies boundaries in Santa Clara County based on posterior means of random effects for esophageal cancer and similarly for Los Angeles County, which depicts the lowest posterior mean for the random effects and is surrounded by counties with higher values. These disparities are likely due to the inclusion of smoking-standardized differences in our adjacency model, which significantly influences lung, esophageal, and larynx cancers (Doll et al., 2005).

For distinguishing boundaries among cancers, we examine shared difference boundaries and mutual

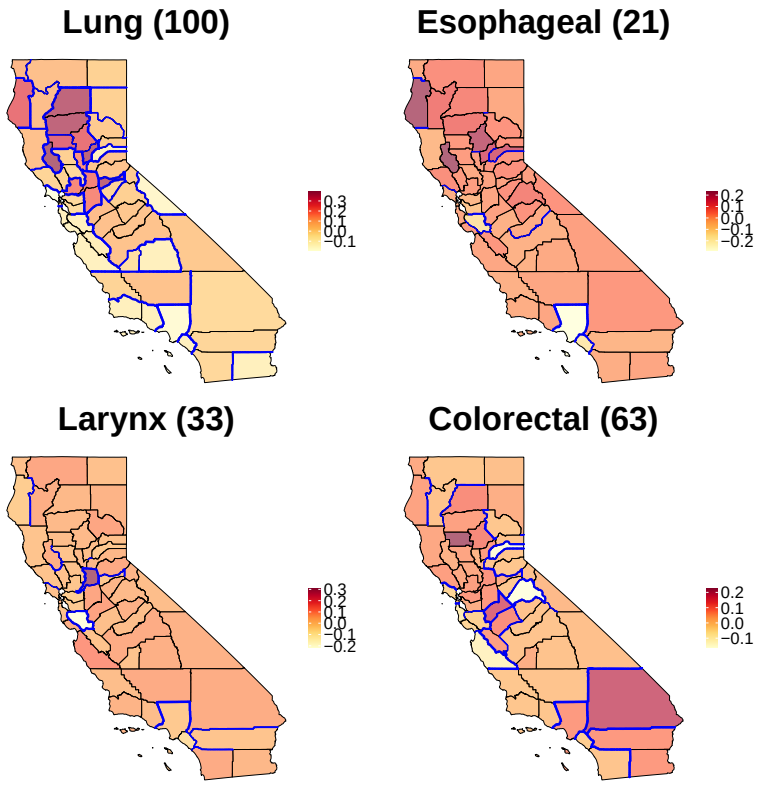


Figure 2.6: Estimated difference boundaries (numbers in parenthesis) in blue for 4 cancers among counties in California based on posterior mean of ϕ_{id} when $\zeta = 0.05$.

cross-cancer boundaries. The shared difference boundaries refer to those detected in common across different cancers. Figure 2.7 illustrates the shared boundaries for each cancer pair, denoted as $P(\phi_{id} \neq \phi_{jd}, \phi_{id'} \neq \phi_{jd'} \,|\, \mathbf{y})$ where $d \neq d'$. We have observed consistent detection of specific boundaries whenever lung is affected. For example, Los Angeles reveals disparities with Ventura, Kern and San Bernardino. Similarly, when examining all cancers, disparities are evinced between Santa Clara and Stanislaus. Notably, all the Los Angeles geographic neighbors are consistently identified.

We introduce mutual cross-cancer boundaries by evaluating $P(\phi_{id} \neq \phi_{jd'}, \phi_{id'} \neq \phi_{jd} \,|\, \mathbf{y})$, where $i \sim j$ and $i < j$, to ascertain boundaries for cross-cancer differences. This boundary effectively segregates the effects pertaining to distinct cancers across adjacent counties (see Figure 2.8). We observe analogous detection between counties for lung-esophageal and lung-larynx differences. It is clear that the number of detected borders is significantly higher for lung-esophageal, lung-larynx and lung-colorectal cancers compared to other types. Moreover, Stanislaus is clearly separated from Santa Clara and Los Angeles from Ventura, indicating notable differences when these cancer relations are taken into account.

Including explanatory variables affect difference boundary detection, leading to either larger or smaller numbers of such boundaries. This depends on whether a covariate intensifies the disparity in residual spatial effects between two adjacent counties, whence a difference boundary is more likely to be observed between them. Conversely, if a covariate explains or absorbs the disparity in residual spatial effects between two neighboring counties, then the presence of a difference boundary diminishes. The former occurs, for example, for lung, esophageal and larynx boundary detection

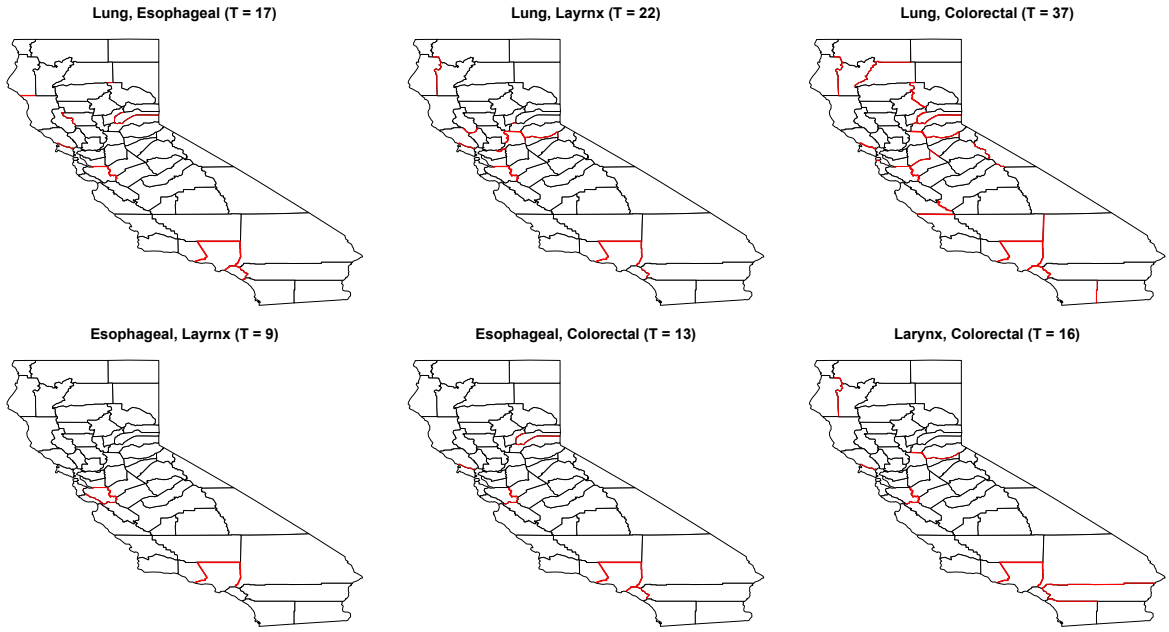


Figure 2.7: Shared difference boundaries in red detected for each pair of cancers in California map when $\zeta = 0.05$. The numbers in parenthesis indicate the difference boundaries detected.

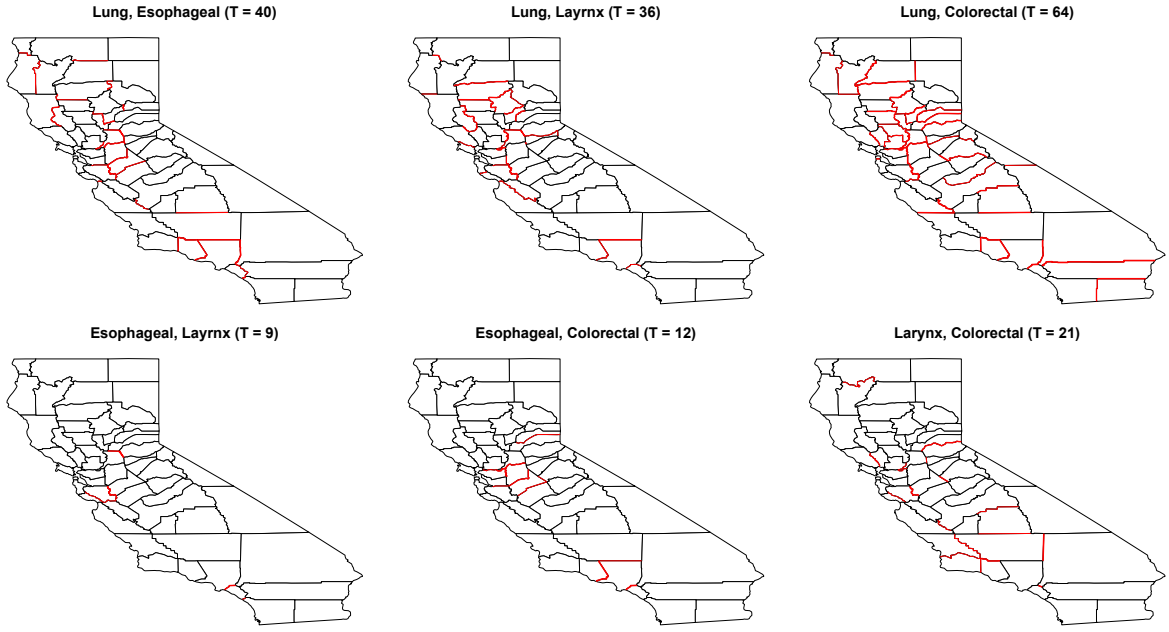


Figure 2.8: Mutual cross-difference boundaries (highlighted in red) detected by the model for each pair of cancers in California map when $\zeta = 0.05$. The values in brackets are the number of difference boundaries detected.

individually, while the latter is seen for colorectal for which we detect fewer boundaries. The impact of covariates on the number of difference boundaries depends on how they affect the variation in rates across neighboring counties. We note that while covariates always absorb some spatial effects, relationships among cancers, regions and covariates are intricate and a definitive pattern can remain elusive.

Turning to adjacency models, non-adjacency is detected based on the following:

$$P(w_{d,ij} = 0 \,|\, \mathbf{y}, i \sim j) \tag{2.19}$$

where $w_{d,ij}$ represents the (i, j)-th element of the adjacency matrix corresponding to the d-th cancer. Figure 2.9 presents a map of California where the thickness of the boundaries indicates the associated probabilities. County boundaries highlighted in blue signify that (2.19) is not negligible and are identified as "non-adjacenct". Northern and central counties exhibit higher values of (2.19). Figure 2.9 suggests that this phenomenon likely reflects the greater differences in covariates between neighboring counties in these regions than in the south. Considering all four cancers, the number of adjacencies having probability (2.19) under 0.1 are 67 for lung, 5 for esophageal, 79 for larynx and 67 for colorectal. These correspond to approximately 48%, 4%, 57% and 48% of the total number of county boundaries, respectively.

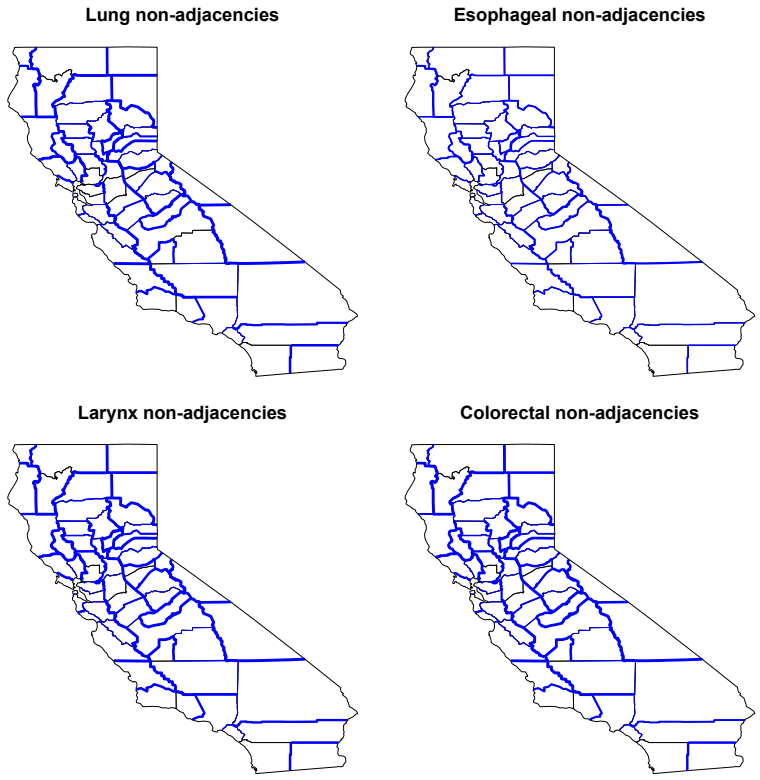


Figure 2.9: Non-adjacencies (shown in blue) over the California map. The thickness of the lines is proportional to the probability of being considered as a non-adjacency

In general, non-adjacencies with the highest probabilities are predominantly between counties with large differences in covariates, especially in smoking rates, while esophageal cancer seems to exhibit lower probabilities. Conversely, counties with lower differences in covariates exhibit zero detected

non-adjacencies. This strongly suggests that smoking, more than the other covariates, serves as a dissimilarity metric for identifying disparities as the estimated boundaries closely align with significant variations in smoking rates.

We discover strong dependence among lung, esophageal, and larynx cancers, which implicates shared risk factors at play. A compelling aspect of such associations is revealed through the visual representations in Figure 2.6. We evince similar spatial clusters for lung, esophageal, and larynx cancers, which imply similar values for the corresponding random effects. This corroborates the notion that these cancers are closely related and potentially influenced by shared factors. The spatial clusters observed in our analysis indicate that the incidence and prevalence of lung, esophageal, and larynx cancers tend to occur in close proximity to one another. This spatial pattern may be influenced by a multitude of factors including, but not limited to, geographical characteristics, socioeconomic factors, or cultural and lifestyle factors specific to certain regions. Understanding these spatial relationships provide valuable insights into the complex interplay between environmental and individual risk factors.

Sections B.4.1 and B.4.2 in the Appendices present results from two additional settings. The former introduces covariates as fixed effects and in the adjacencies, while the latter does so only in the regression while fixing the adjacencies. The difference boundary maps evince the impact of differences among covariates on *defining* the adjacency rather than fix it as a geographic neighbor. Introducing covariates as fixed effects and in the adjacency explains how specific factors contribute to the mean and the spatial patterns. Similarly, including covariates solely in the mean structure, with fixed adjacencies, enables assessing the influence of these variables on the overall mean values of the observed phenomena.

2.9.1 Monte Carlo parameter estimates and standard error

Monte Carlo estimates, whether for Markov chain Monte Carlo output or any other purpose, should be accompanied by Monte Carlo standard error as a measure of reliability of posterior inference based on the estimates from the Monte Carlo samples. If the central limit theorem applies, then the corresponding covariance matrices' estimates must be reported, either directly or indirectly. Due to the correlation between the obtained samples, these quantities necessitate more advanced tools compared to typical sample estimators. The validity of a central limit theorem is not guaranteed in all cases because it depends on the convergence rate of the underlying Markov chain. Nonetheless, this assumption is also made when employing various convergence diagnostics. We use the `mcmcse` `R` package (Flegal et al., 2021) to compute Monte Carlo standard errors and the effective sample size, based on Gong and Flegal (2016) to demonstrate that we can maintain control over the Monte Carlo error by applying a predefined threshold Vats et al. (2019).

Table 2.4 presents the estimates of the model described in Section 2.9 including only cancer-specific intercepts in the regression model and the 3 covariates described in Section 2.1 in the adjacency model. While none of the intercepts are significant, we note that lung and esophageal cancers reveal positive estimates indicating higher expected counts than the standardized rates, E_{id} defined in Section 2.1 and introduced in the model in Section 2.9, while larynx and colorectal exhibit negative estimates indicating lower expected counts than their standardized rates. The last column of Table 2.4 presents Monte Carlo standard errors with the only value higher than 0.05, the usual threshold, being 0.061. We compute the multivariate effective sample size of the chain to be 1,847

Table 2.4: Posterior estimates, posterior standard deviations and Monte Carlo standard errors for the regression coefficients, atoms and their precision in the hierarchical model described in (2.11) using the Poisson regression model described in Section 2.9.

Variable	Estimate	SD	MC SE
β_1	0.022	0.027	0.006
β_2	0.075	0.033	0.005
β_3	-0.026	0.037	0.005
β_4	-0.045	0.030	0.007
θ_1	0.072	0.030	0.007
θ_2	-0.283	0.043	0.007
θ_3	-0.098	0.029	0.006
θ_4	-0.020	0.031	0.007
θ_5	0.156	0.162	0.027
θ_6	-0.072	0.175	0.061
θ_7	0.365	0.041	0.009
θ_8	-0.115	0.085	0.020
θ_9	0.312	0.086	0.023
θ_{10}	0.178	0.062	0.019
θ_{11}	-0.149	0.170	0.029
θ_{12}	-0.158	0.028	0.006
θ_{13}	0.184	0.037	0.007
θ_{14}	0.095	0.139	0.032
θ_{15}	0.010	0.029	0.006
τ	7.246	2.368	0.024

using the formula $B\frac{|\Lambda|^{1/p}}{|\Sigma|^{1/p}}$. Here, B is the total number of iterations post burn-in, Λ represents the sample covariance matrix, and Σ is an estimate of the Monte Carlo standard error. This computation accounts for the multivariate dependence structure of the process. Comparing our effective sample size to the minimum requirement of 7,619 for achieving a precision level relative to the variability in the target distribution of 0.05, we observe that ours is significantly lower. However, it is important to note that we computed the precision level of our estimate, obtained with an estimated effective sample size of 1,847.068, which amounts approximately to 0.01 and, while higher than precision levels used in simpler models, is still deemed acceptable for the inference we draw using these complex hierarchical models.

2.10 Discussion

We address statistical methods for detecting spatial disparities based upon mapping disease rates and modeling difference boundaries. The need for formal modeling in ascertaining health disparities has been articulated in the literature (see, e.g., Rao, 2023, and references therein), but formal model-based approaches for detecting geographic disparities remain relatively scarce. The inferential framework developed here achieves full probabilistic uncertainty quantification using a flexible (nonparametrics) prior on the spatial effects that endows random effects with discrete masses to detect disparities among neighbors while also allowing spatial smoothing across neighbors as is customary in disease mapping.

As in Gao et al. (2023), we also deal with multiple diseases, possibly associated among themselves,

but now explicitly address multivariate dependencies flexibly using posited graphical models. Second, we address a possible limitation of Gao et al. (2023) where factors informing about disparities need to be evaluated using their inclusion and exclusion in models with subsequent model assessment. Instead, we introduce such factors in the adjacency model. This enriches earlier statistical frameworks for "areal wombling" that have attempted similar adjacency models with limited effectiveness since statistical significance of boundary effects is difficult to establish with such hierarchically embedded information. Our framework, on the other hand, enables flexible modeling of boundaries (as in "wombling"), but uses a Bayesian FDR approach to ascertain spatial difference boundaries.

Our application to the SEER database leads to data-based discoveries with regard to detecting spatial disparities while accounting for inter-disease dependence. Associations among lung, esophageal, and larynx cancers carry significant implications for public health and cancer prevention initiatives. These associations suggest that interventions targeting one of these malignancies could potentially have a positive impact on reducing the incidence and burden of the others. By recognizing shared patterns and risk factors, healthcare professionals and policymakers can develop more targeted and effective interventions. In response to our findings, a multifaceted approach can be employed to reduce the overall burden of these cancers. Such an approach would encompass public awareness campaigns to educate individuals about the shared risk factors and promote healthy lifestyle choices.

Finally, we note future avenues for research. We do not claim our inferential framework to be unique in detecting spatial disparities. We noted a substantial relevant literature in areal "wombling" that offers ample scope for developing alternate inferential methods based upon an explicit threshold on the difference of neighboring outcomes (referred to as "boundary likelihood values" in "wombling"). Wu and Banerjee (2024) offer a rigorous Bayesian linear regression framework that calibrates the threshold using Shannon entropy. Extending such approaches to multivariate and non-Gaussian outcomes and comparing with our current approach, while beyond the scope of the current chapter, will comprise future investigations.

Chapter 3

Survival modeling of smartphone trigger data in crowdsourced seismic monitoring

Earthquake early warning systems (EEWSs) (Gasparini et al., 2007) are deployed in seismic-prone areas to mitigate the earthquake risk. Dense networks of instruments are used to detect earthquakes in real-time, to stop critical processes, and to send alerts to people who will experience strong ground shaking (see Figure 3.1 for a symplified visualization). Construction and operating costs of classic EEWSs are in the order of the millions of euros (Given et al., 2014). This limited their widespread, especially in low-income seismic countries that could largely benefit from EEWSs.

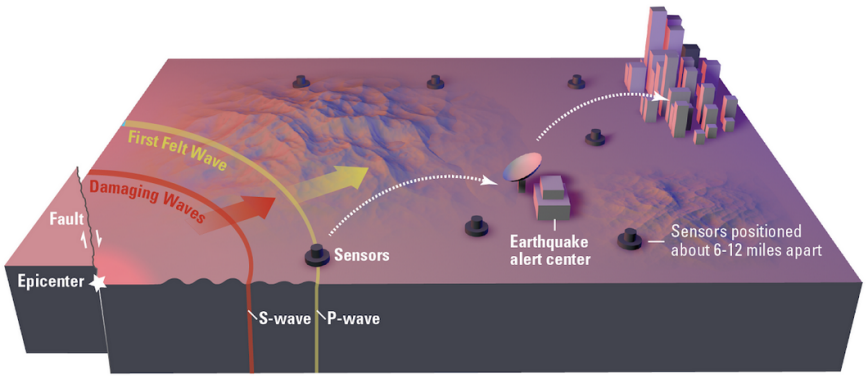


Figure 3.1: EEWS functioning scheme.

In the last two decades, researchers have proposed low-cost alternatives (Cochran et al., 2009; Clayton et al., 2011). Only with the smartphone revolution, however, new solutions (Finazzi, 2016; Kong et al., 2016) became truly accessible to the global population. The Earthquake Network (EQN) citizen science initiative (Finazzi, 2020) was the first to implement a smartphone-based EEWS able to send real-time alerts on people's smartphones. Since 2013, more than 25 million people have taken part in the initiative, and the system has contributed to sending nearly 7,600 alerts. During the 2023 Pazarcik (Turkey) earthquake, EQN provided a forewarning for up to 58 seconds to people exposed to very high levels of ground shaking (Finazzi et al., 2024), proving its efficiency in earthquake risk

mitigation.

Smartphones are less reliable than the scientific instruments used in traditional EEWS because they are subject to anthropic noise, and earthquake monitoring is based on a low-cost accelerometer. Thus, specific statistical methodologies have been developed for the real-time detection of earthquakes (Finazzi and Fassò, 2017; Massoda Tchoussi and Finazzi, 2023) and for assessing the detection performance of the smartphone-based monitoring network (Finazzi et al., 2022).

An open problem in smartphone-based earthquake monitoring is the estimation of earthquake parameters (e.g., epicenter, depth, origin time and magnitude) from noisy smartphone measurements. For each earthquake detection, EQN provides rough and often biased estimates of these quantities. In this work, we develop a statistical methodology to improve the EQN estimates of epicenter, depth and origin time. The methodology takes into account the specificities of smartphones as seismic sensors, namely the possibility of detecting non-seismic accelerations, the possibility of missing an earthquake, and the uncertainty of which seismic wave is detected by the smartphone.

We embed the estimation problem into the framework of survival data analysis by exploiting the analogy between right censored clinical data, which are naturally analyzed with survival models, and smartphone-based earthquake data. In particular, smartphones are seen as patients of a clinical study. A smartphone that detects acceleration (triggering smartphone) is considered a patient dying from a particular disease. In contrast, a smartphone that does not detect acceleration is considered a patient surviving. The earthquake plays the same role as a disease, while smartphones that fail to detect the seismic waves (*faulty smartphones*) are interpreted as cured patients. The earthquake detection by the EQN system is then the censoring event. In fact, the triggering times of (uncured/faultless) smartphones, which will eventually trigger after the earthquake detection, will not be observed.

Finally, mortality due to other causes relates to triggers unrelated to the seismic waves. In population-based clinical studies, mortality due to other causes is generally incorporated using relative survival methods (Cutler and Axtell, 1963). In this setting, *cure* occurs when the hazard rate in the diseased group of individuals returns to the same level as that expected in the general population (Peng and Taylor, 2014). In other words, a certain proportion of subjects in the population is not expected to experience the events of interest, so the relative survival curve appears to plateau after a given time. Mixture survival modeling has been widely employed to incorporate this plateau and estimate the cure fraction in clinical studies. For instance, De Angelis et al. (1999) applied a mixture survival model to analyze individual data on colon cancer and estimate both the proportion of cured patients and the failure time distribution for fatal patients. Lambert et al. (2007) compared relative survival and cure fraction estimates between mixture and non-mixture cure fraction models. Lambert et al. (2010) extended the mixture cure model by incorporating a finite mixture of parametric distributions to provide flexibility in the shape of the relative survival or excess mortality functions. A Bayesian mixture cure model with spatial random effects has been proposed by Yu and Tiwari (2012) to model county-level cancer survival data using Markov chain Monte Carlo (MCMC) methods, while Lázaro et al. (2020) implemented a Bayesian mixture cure model using the integrated nested Laplace approximation.

Summing up, the contribution of this work is to propose a Bayesian mixture model for analyzing smartphone-based early warning data in order to estimate relevant earthquake parameters, uncertainty included. The model leverages information from spatial locations and times of triggering smartphones

as well as from spatial locations of non-triggering smartphones. As a key feature, the mortality function (i.e., the smartphone triggering function) is assumed to be a mixture of two parametric densities describing the earthquake waves that may be detected by a smartphone in a given time interval. On the computational side, an efficient sampling strategy has been designed to provide a full posterior inference of the earthquake parameters. Specifically, we develop an adaptive parallel tempering MCMC algorithm to address the computational challenges associated with the multi-modality of the posterior distribution.

We illustrate the soundness and good performance through a simulation study and a set of relevant case studies. Specifically, the approach is applied to the EQN detections of the magnitude 7.8 2023 Pazarcik earthquake, of the magnitude 7.1 2019 Ridgecrest (US) earthquake and of a relatively small magnitude earthquake that occurred in Mexico with epicenter offshore, thus more challenging in terms of epicenter estimation.

The remainder of this chapter is structured as follows. Section 3.1 outlines the EQN system's operation and the dataset generated during earthquake detection, including the three datasets analyzed in this study. Section 3.2 introduces key concepts in survival analysis. Section 3.3 presents the development of the Bayesian mixture cure model and explains how earthquake physics are incorporated into the model. Section 3.4 delves deeper into how the physics of earthquakes is incorporated into the model's construction. Section 3.5 finalizes the model specification by detailing the prior distributions for all parameters. Computational aspects are discussed in Section 3.6. Section 3.7 reports the results of a simulation study. Section 3.8 demonstrates the methodology through three case studies. Finally, Section 3.9 summarizes the findings, outlines potential future work, and Appendix C includes additional results from simulation experiments and real data analyses.

3.1 EQN functioning and data generation

This section provides an overview of the functioning of the EQN system and the dataset generated every time an earthquake is detected. EQN deploys a global smartphone network for seismic monitoring. To join the EQN initiative, smartphone users must install the EQN app, which is available for Android, iOS and Huawei operating systems. Seismic monitoring starts when the smartphone is charging and not used. This implies that the network is highly dynamic, with smartphones entering and leaving the network anytime. A central server collects the monitoring data from the smartphones, and it knows the state of the network, namely which smartphones are monitoring and where they are located. On each smartphone, seismic monitoring is done through the smartphone accelerometer, which continuously measures and provides the smartphone acceleration in space. A monitoring smartphone possibly detects a seismic wave and any other acceleration induced by a local force (e.g., someone moving the smartphone or the object above which the smartphone is located). To better understand how EQN works, Figure 3.2 depicts a simplified representation of its functioning.

Earthquakes are characterized by primary (P) and secondary (S) waves. The P wave is faster than the S wave and usually produces mild ground shaking. Most of the earthquake damage is caused by the S wave. When an earthquake hits, the monitoring smartphone may detect the P or the S wave depending on the earthquake magnitude, the distance between the smartphone and the epicenter, where the smartphone is located (e.g., on which floor of the building and above which object) and

Figure 3.2: EQN functioning scheme.

on the specific accelerometer sensitivity (which may differ from a smartphone to another). The monitoring smartphone may also not detect the earthquake, either because the epicenter is too far or because the smartphone is not able to send the information to the server.

For a given region, the EQN detection algorithm Finazzi and Fassò (2017) compares the number of monitoring smartphones with the number of triggering smartphones in the last 120 seconds (with respect to the last trigger received by the server). If the number of triggers exceeds a threshold (fraction of the monitoring smartphones), an earthquake is claimed to be detected.

When the detection occurs, two lists of data are stored: the list of triggering smartphones during the last 120 seconds and the list of smartphones which were known to be monitoring at the detection time but did not trigger (active smartphones). Triggering smartphones are described by their spatial coordinates and by their triggering times, while active smartphones are only described by their spatial coordinates. Both lists cover an area of 300 km in a radius centered on the EQN detection location, provided by the detection algorithm as a rough estimate of the earthquake epicenter. This estimate, however, assumes that all smartphones are triggered because of the earthquake; it does not take into account the information of the non-triggering smartphones, and it does not consider the dynamic of the P and S waves. It follows that the EQN epicenter estimate is usually largely biased.

3.1.1 EQN datasets

To better understand their complexity and before entering into modeling details, we describe here the three EQN datasets analyzed in Section 3.8. Table 3.1 provides, for the three earthquakes, the number of triggering smartphones, the number of active smartphones that did not trigger, the hypocenter coordinates, the origin time and the earthquake magnitude retrieved by the European-Mediterranean Seismological Centre (EMSC; `https://www.emsc-csem.org`).

The first dataset is related to the recent magnitude 7.8 Pazarcik earthquake of February 6, 2023. The event was the deadliest and strongest earthquake in Turkey since 1939 and the second most powerful after the 1668 North Anatolia earthquake. Additionally, it was the deadliest earthquake in modern-day Syria. EQN detected the event with a delay of 10 seconds with respect to the earthquake origin time, allowing people living far away from the epicenter to receive a forewarning. Panel (a) of Figure 3.7 depicts the EQN dataset generated when the earthquake was detected. From a visual

Table 3.1: EQN datasets summaries and parameters of the three earthquakes analysed in this work. Earthquake parameters were retrieved from the EMSC on February 16, 2023.

	Pazarcik	Ridgecrest	Mexico
EQN triggers	16	80	150
EQN active smartphones	158	241	1115
Latitude [deg]	37.17	35.76	15.64
Longitude [deg]	37.08	-117.62	-94.97
Depth [km]	20.0	8.0	27.9
Origin time [UTC]	2023-02-06 01:17:36	2019-07-06 03:19:52	2019-07-17 06:25:48
Magnitude [M_W]	7.8	7.1	4.8

analysis of the triggers and active smartphones, it is easy to delimit the area not yet impacted by the earthquake.

The second dataset concerns the magnitude of the 7.1 Ridgecrest (US) July 6, 2019 earthquake. This was the most powerful earthquake in California in 20 years and was largely felt across the state and in Nevada. Due to the small number of smartphones in the epicentral area at the time of the event, the EQN detection occurred in Los Angeles at around 190 km from the epicenter and 41 seconds of delay (see Bossu et al. 2022 for details), limiting the EQN early warning capabilities but still providing real-time information to citizens. Panel (a) of Figure 3.8 shows the EQN dataset. Smartphones are mainly located around the biggest cities of the study region, i.e., Los Angeles and San Diego. Moreover, triggering times are noisy and do not necessarily increase with the distance from the earthquake's epicenter.

The third dataset concerns a magnitude 4.8 earthquake on July 17, 2019, offshore the Mexican coast. The event did not impact people, but assessing our methodology when the epicenter is "outside" and far from the smartphone network is useful. EQN detected the earthquake after 29 seconds at around 90 km from the epicenter, thanks to the smartphones located in Salina Cruz and Juchitán de Zaragoza (the two cities along the coast closest to the epicenter). Again, panel (a) of Figure 3.9 shows the EQN dataset. Despite the earthquake being detected on the coast, the dataset includes many triggers from areas not yet reached by the seismic waves. EQN is very popular in Mexico, and each city is a source of spurious triggers that tend to bias the epicenter estimate.

The features of the three datasets call for a statistical model that exploits all the available information (including the information of the non-triggering smartphones), and that takes into account the dynamic of seismic waves.

3.2 Survival analysis

Survival analysis, the study of time-to-event data, is widely used across various disciplines such as medicine, biology, engineering, public health, epidemiology, and economics. This type of analysis typically involves tracking a set of "individuals" and recording the time until a specific event, often referred to as failure time or lifetime. Notably, the term "individuals" can refer to people or entities like electronic components, depending on the context. Examples of survival analysis applications include estimating the time to failure of electronic devices such as refrigerators or pacemakers in industrial reliability studies, tracking the duration of unemployment periods in social sciences, measuring the

lifespan of machine parts, monitoring the time until a patient achieves remission or recovery, observing the time taken to complete a task, and evaluating the duration of economic cycles.

In all cases, the time-to-event variable is non-negative and often subject to censoring. For instance, in medical studies, a patient may be lost to follow-up for reasons unrelated to the disease being studied, making it impossible to determine whether the patient succumbed to the illness. Recent advancements in computational techniques and modeling approaches have significantly enhanced the appeal of Bayesian methods for survival analysis. As a result, Bayesian frameworks have become increasingly popular and are now widely applied in various research fields.

Let T be a continuous, nonnegative random variable representing the survival time of individuals in a population. Denote the probability density function of T by $f(t)$, with $t \in [0, \infty)$. The cumulative distribution function is given by:

$$F(t) = P(T \leq t) = \int_0^t f(u)\, du.$$

The probability that an individual survives beyond time t is described by the survivor function:

$$\begin{aligned} S(t) &= P(T > t) = 1 - F(t) \\ &= 1 - \int_0^t f(u)\, du = \int_t^\infty f(u)\, du. \end{aligned}$$

Note that $S(t)$ is a monotonically decreasing function with $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = 0$. The hazard function, $h(t)$, represents the instantaneous failure rate at time t and is defined as:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t} \\ &= \frac{1}{P(T > t)} \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} = \frac{f(t)}{S(t)}. \end{aligned}$$

Specifically, $h(t)\Delta t$ approximates the probability of failure in the interval $(t, t+\Delta t]$, given survival up to time t. The functions $f(t)$, $F(t)$, $S(t)$, and $h(t)$ provide mathematically equivalent descriptions of the distribution of T. For instance, since $f(t) = -\frac{d}{dt}S(t)$, it follows that:

$$h(t) = -\frac{d}{dt} \log(S(t)).$$

Integrating both sides of this expression and then exponentiating, we obtain:

$$S(t) = \exp\left(-\int_0^t h(u)\, du\right).$$

The exponent argument, i.e., the cumulative hazard function $H(t)$, is defined as:

$$H(t) = \int_0^t h(u)\, du,$$

and it is hence related to the survivor function by the relationship:

$$S(t) = \exp(-H(t)).$$

Given that $S(\infty) = 0$, it follows that $H(\infty) = \lim_{t\to\infty} H(t) = \infty$. Thus, the hazard function $h(t)$ has the properties:

$$h(t) \geq 0 \quad \text{and} \quad \int_0^\infty h(t)\, dt = \infty.$$

Finally, it follows that the pdf $f(t)$ can be expressed exclusively in terms of the hazard function as:

$$f(t) = h(t)S(t) = h(t)\exp(-H(t)) = h(t)\exp\left(-\int_0^t h(u)\, du\right). \tag{3.1}$$

For a comprehensive overview of survival analysis, see Klein et al. (2016), and for a detailed discussion within a Bayesian framework, refer to Ibrahim et al. (2005).

3.2.1 Right censoring

A key feature often present in time-to-event data is censoring, which occurs when the exact lifetimes of some individuals are unknown but are constrained within certain intervals, while others are known exactly. Censoring can be categorized into three types: right censoring, left censoring, and interval censoring. Right censoring is the most common form, where the exact event time for some individuals is unknown but is known to exceed a certain value. Specifically, an observation is right-censored at time t_2^* if the event time is unknown but occurs after t_2^*. Conversely, left censoring occurs when the event time is only known to be less than or equal to a certain time t_1^*. Interval censoring applies when the event time is known to fall within a specific interval (t_1^*, t_2^*). While right-censored and interval-censored data are frequent in survival studies, left-censored data are relatively rare. Here, we focus exclusively on right-censored survival data.

In Type I censoring, the event is only observed if it occurs before a specified time, which may vary for each individual. For example, in clinical trials or animal studies, a fixed number of subjects receive treatment, but due to time or budget constraints, the study ends before all subjects experience the event. In this case, if no subjects are lost or withdrawn, the censoring time equals the study's duration for all censored observations.

Type II censoring occurs when a study continues until the failure of the first k individuals, where k is a predetermined number less than the total sample size n. This method is common in equipment life testing, where all items are tested simultaneously, and the experiment concludes once k of the n items have failed. Type II censoring is often more time-efficient and cost-effective since not all items need to fail. Statistically, it simplifies analysis because it focuses on the k smallest lifetimes in the sample, enabling the use of order statistics to derive likelihoods and apply inferential techniques. In this case, k is the number of failures, $n-k$ is the number of censored observations, and the censoring time, the k-th ordered lifetime, is random. For a comprehensive review of censoring mechanisms and related topics, refer to Klein and Moeschberger (2006). As will become clear in Section 3.3.1, we leverage this framework to derive the likelihood used in our analysis.

3.2.2 Cure models

Cure rate models, which incorporate a cure fraction, are becoming increasingly important in the analysis of clinical trial data and other applications. These models are specifically designed for

censored survival data, where a subset of subjects will never experience the event of interest, regardless of the observation period. Such individuals are often referred to as "cured" or "long-term survivors". Cure rate models provide simultaneous estimates of both the proportion of individuals cured of a disease and the distribution of survival times for those who are not cured. For a comprehensive review of these models, see Peng and Taylor (2014). They are particularly valuable in contexts where a significant proportion of individuals are considered "cured", making them well-suited for modeling time-to-event data.

One challenge in analyzing survival data from subjects with the possibility of being cured is how to handle censoring. Censoring occurs either because a subject is cured or because a subject has not been followed long enough for the event to occur. In many cases, these two scenarios cannot be easily distinguished. Given that some subjects may never experience the event of interest, it is logical to consider a mixture model that accounts for two groups: the cured and the uncured. Additionally, in some population studies, only death times are available. While the length of follow-up and the nature of the disease may suggest that most deaths are due to the disease, deaths from other causes can occur and may be indistinguishable. In such cases, it is useful to define "cure" as the point at which the mortality rate for those diagnosed with the disease returns to the level expected in the general population.

In the next subsection, we will illustrate how we incorporate cure rate models and right censoring into our modeling framework.

3.3 Mixture model for relative survival

Our inferential problem is formalized within the setting of survival data analysis with the right censoring and cure. Recall that the observation period is from EQN detection back two minutes. Let T be the smartphone triggering time over the two-minute frame, and let T^* be the censoring time, equal to the time EQN detects the earthquake. Hence, the censoring corresponds to the end of the study period for all smartphones, also known as administrative censoring (Cox and Oakes, 1984). We denote $Y = \min(T, T^*)$ and $\Delta = I(T \leq T^*)$ the triggering (censoring) indicator; in other words, $\Delta = 1$ when the smartphone triggered.

As usual in survival analysis, the survival function gives the probability that a subject (a smartphone) will survive (will not trigger) past time t, i.e., $S(t) = P(Y > t) = 1 - F(t)$. In particular, in relative survival models (De Angelis et al., 1999; Lambert et al., 2010), the overall (all-cause) survival function $S(t)$ is given by the product of the expected survival function $S_0(t)$ and the relative survival function $R(t)$, that is

$$S(t) = S_0(t)R(t). \tag{3.2}$$

The relative survival function $R(t)$ can be interpreted as the net survival probability, i.e., the survival probability, when the disease (earthquake) is the only cause of death (triggering), and all other causes could be eliminated. In our setting, $S_0(t)$ represents the expected probability that a smartphone will not trigger until time t for any other causes than the earthquake.

Given Equation (3.2), the overall hazard function $h(t)$ is obtained as the sum of two components:

the expected hazard function $h_0(t)$ and the excess hazard function $r(t)$, that is

$$h(t) = h_0(t) + r(t). \tag{3.3}$$

The assumption behind Equations (3.2) and (3.3) is that the triggering associated with the earthquake is independent of triggering related to other causes, i.e., there are independent competing risks (Gamel and Vogel, 2001). Both $S_0(t)$ and $h_0(t)$ in Equation (3.2) and (3.3) are assumed to be known and equal, respectively, to the survival and hazard function of an exponential distribution with parameter λ seconds^{-1}.

As previously mentioned, as time goes to infinity, the survival curve goes to 0, i.e. $S(\infty) = 0$. In our context, however, this assumption is unrealistic since, in practice, a smartphone could never be triggered by an earthquake. This is where cure models come in. Cure models refer to a class of models for censored survival data from subjects where some will not develop the event of interest, no matter how long they are followed. The subjects who will not develop the event of interest are often referred to as cured subjects or long-term survivors. See Peng and Taylor (2014) for a review of these models.

In our setting, a smartphone that does not trigger because of the earthquake is interpreted as a *faulty smartphone* because it is not functioning properly within the EEWS, i.e. it is not detecting seismic waves. Therefore, since some smartphones will never trigger, it is natural to consider a mixture model where the population is a mixture of two groups: (i) a proportion π of cured cases with respect to the specific cause (faulty smartphones that will never trigger because of the earthquake) and thus have a mortality rate similar to that expected in the general population (i.e., smartphones that will eventually trigger due to other causes), and (ii) a fraction $(1 - \pi)$ of uncured cases, i.e., faultless smartphones that will eventually trigger due to the earthquake and whose survival function will tend to zero.

Let W be the faulty indicator with $W = 1$ if the smartphone is faulty, where $P(W = 1) = \pi$. Define $S_Q(t) = P(T > t \mid W = 0)$ the survival function of the faultless population, i.e., the survival function of smartphones that will eventually trigger because of the earthquake. The following survival function then defines the mixture model

$$S(t) = S_0(t)\{\pi + (1 - \pi)S_Q(t)\},$$

where the survival function regarding the π component is equal to one, since a faulty smartphone has probability one to not triggering up to t, i.e., it will never trigger due to the earthquake. Note that the faulty smartphone definition encoded in the π parameter is from a population perspective and does not provide any information about individual smartphones. The corresponding hazard rate can be expressed as the sum of the background triggering rate and the triggering rate associated with the earthquake, that is

$$h(t) = h_0(t) + \frac{(1 - \pi)f_Q(t)}{\pi + (1 - \pi)S_Q(t)}, \tag{3.4}$$

where $f_Q(t)$ is the probability density function associated with $S_Q(t)$. Hence, the relative survival function has an asymptote at the faulty fraction π or, equivalently, the excess hazard function $r(t)$ has an asymptote at zero.

The density $f_Q(t)$ in Equation (3.4) is modeled to capture earthquakes dynamic. In particular, we express the density $f_Q(t)$ as a mixture of two densities corresponding to the P and S waves, that is $f_Q(t) = \alpha f_P(t) + (1-\alpha) f_S(t)$, where the choice of $f_P(t)$ and $f_S(t)$ is detailed in the next Section 3.4. Hence, the survival function becomes

$$S(t) = S_0(t)\{\pi + (1-\pi)(\alpha S_P(t) + (1-\alpha) S_S(t))\}, \tag{3.5}$$

with hazard function

$$h(t) = h_0(t) + \frac{(1-\pi)(\alpha f_P(t) + (1-\alpha) f_S(t))}{\pi + (1-\pi)(\alpha S_P(t) + (1-\alpha) S_S(t))}. \tag{3.6}$$

Similar to Lambert et al. (2010), this model can be thought of as a three-component mixture: (i) a proportion π of faulty smartphones, i.e., smartphones will never trigger because of the earthquake, (ii) a proportion $(1-\pi)\alpha$ of faultless smartphones with survival distribution $S_P(t)$, i.e., the proportion of smartphones detecting the P wave and (iii) the remaining proportion $(1-\pi)(1-\alpha)$ of faultless smartphones with survival distribution $S_S(t)$, i.e., the proportion of smartphones detecting the S wave.

3.3.1 Likelihood function

In this section, we derive the likelihood function, taking into account that the censoring time is dependent on the observed data. Furthermore, since the censoring time coincides with one of the triggering events, the remaining triggering times are no longer mutually independent. This dependency arises because EQN uses only a subset of the triggering times to detect earthquakes, making the censoring time inherently dependent on the triggering times. Neglecting this dependence can lead to biased and inconsistent inferences, as noted by Davison (2003). Various strategies have been proposed to address issues of dependent censoring and cure modeling (Li et al., 2007; Megan Othus and Tiwari, 2009). In response, we derive a likelihood function that explicitly incorporates these dependencies.

Let n be the number of active smartphones in the study area over the observational window. Let $T_1, \dots, T_n$ be the survival times and T^* a common censoring time, i.e., all smartphones share the same censoring time. This setting is also referred to as administrative censoring, (see, for instance, Ibrahim et al., 2005). Let $Y_i = \min\{T_i, T^*\}$ and $\Delta_i = I(T_i \leq T^*)$ for $i = 1, \dots, n$. The observed data consists of the times (observed or censored) and of the triggering/censoring indicators, and is denoted by (y_i, δ_i) for $i = 1, \dots, n$. The likelihood function is then $P(A)$, where the A is the set

$$A = \{Y_1 \in dy_1, \dots, Y_n \in dy_n, \Delta_1 = \delta_1, \dots, \Delta_n = \delta_n\}.$$

In the setting of EQN detection, $T^* = T_{(k)}$ is the k-th order statistics of $T_1, \dots, T_n$ over a 120-second time window. Hence it coincides with one of the T_i's, and event A is equivalent to

$$A' = \left\{ \bigcap_{i \in B} (T_i \in dy_i, \Delta_i = 1), \bigcap_{i \in C} (T_i > y_i, \Delta_i = 0), T_{(k)} \in dy_{(k)} \right\}.$$

where $B = \{i : \Delta_i = 1\} \setminus \{i : T_i = T_{(k)}\}$ and $C = \{i : \Delta_i = 0\}$. Then, the likelihood function

becomes

$$\begin{aligned} P(A') &= P\left(\bigcap_{i\in B}\left(T_i \in dy_i, \Delta_i = 1\right), \bigcap_{i\in C}\left(T_i > y_i, \Delta_i = 0\right), T_{(k)} \in dy_{(k)}\right) \\ &= P\left(\bigcap_{i\in B}\left(T_i \in dy_i, \Delta_i = 1\right), \bigcap_{i\in C}\left(T_i > y_i, \Delta_i = 0\right) \mid T_{(k)} \in dy_{(k)}\right) P\left(T_{(k)} \in dy_{(k)}\right) \\ &= P\left(\bigcap_{i\in B}\left(T_i \in dy_i, \Delta_i = 1\right), \bigcap_{i\in C}\left(T_i > y_i, \Delta_i = 0\right) \mid T_{(k)} \in dy_{(k)}\right) P\left(T_{(k)} \in dy_{(k)}\right) \end{aligned} \tag{3.7}$$

Using Lemma 5 in Ahmadi and Nagaraja (2020) we obtain:

$$\begin{aligned} &P\left(\bigcap_{i\in B}\left(T_i \in dy_i, \Delta_i = 1\right), \bigcap_{i\in C}\left(T_i \in dy_i, \Delta_i = 0\right) \mid T_{(k)} \in dy_{(k)}\right) \\ &= \frac{1}{n}\prod_{i\in B}\frac{f(y_i)}{1-S(y_{(k)})}\prod_{i\in C}\frac{f(y_i)}{S(y_{(k)})} \end{aligned} \tag{3.8}$$

where $\max_{i\in B}(y_i) < y_{(k)} < \min_{i\in C}(y_i)$, $f(y_i)$ is the density function in y_i and $S(y_i)$ is the survival function in y_i. Through a straightforward integration of (3.8), we get

$$\begin{aligned} &P\left(\bigcap_{i\in B}\left(T_i \in dy_i, \Delta_i = 1\right), \bigcap_{i\in C}\left(T_i > y_i, \Delta_i = 0\right) \mid T_{(k)} \in dy_{(k)}\right) \\ &= \int_{\underset{i\in C}{\times}\{(y_i,\infty)\}} \frac{1}{n}\prod_{i\in B}\frac{f(y_i)}{1-S(y_{(k)})}\prod_{i\in C}\frac{f(t_i)}{S(y_{(k)})}\prod_{i\in C}dt_i \\ &= \frac{1}{n}\prod_{i\in B}\frac{f(y_i)}{1-S(y_{(k)})}\prod_{i\in C}\int_{(y_i,\infty)}\frac{f(t_i)}{S(y_{(k)})}dt_i \\ &= \frac{1}{n}\prod_{i\in B}\frac{f(y_i)}{1-S(y_{(k)})}\prod_{i\in C}\frac{S(y_i)}{S(y_{(k)})} \end{aligned} \tag{3.9}$$

where $\underset{i\in C}{\times}\{(y_i,\infty)\}$ is the Cartesian product of the (y_i,∞) over the indices $i \in C$ and $\max_{i\in B}(y_i) < y_{(k)} < \min_{i\in C}(y_i)$.

The other term in equation (3.7) is the density of $T_{(k)}$, which is well known to be given by

$$P\left(T_{(k)} \in dy_{(k)}\right) = \frac{n!}{(k-1)!(n-k)!}f(y_{(k)})\left[1-S(y_{(k)})\right]^{k-1}S(y_{(k)})^{n-k}. \tag{3.10}$$

Noting that $|B| = k-1$, $|C| = n-k$ and by combining (3.9) and (3.10) we get

$$\begin{aligned} P(A) &= \frac{1}{n}\prod_{i\in B}\frac{f(y_i)}{1-S(y_{(k)})}\prod_{i\in C}\frac{S(y_i)}{S(y_{(k)})}\frac{n!}{(k-1)!(n-k)!}f(y_{(k)})\left[1-S(y_{(k)})\right]^{k-1}S(y_{(k)})^{n-k} \\ &= \frac{(n-1)!}{(k-1)!(n-k)!}\frac{\prod_{i\in B}f(y_i)}{\left[1-S(y_{(k)})\right]^{k-1}}\frac{\prod_{i\in C}S(y_i)}{S(y_{(k)})^{n-k}}f(y_{(k)})\left[1-S(y_{(k)})\right]^{k-1}S(y_{(k)})^{n-k} \\ &= \frac{(n-1)!}{(k-1)!(n-k)!}\prod_{i\in B}f(y_i)\prod_{i\in C}S(y_i)f(y_{(k)}) \end{aligned}$$

$$\propto \prod_{i \in B} f(y_i) \prod_{i \in C} S(y_i) f(y_{(k)}),$$
$$= \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}$$

where the last line equality is motivated by the fact that $y_{(k)}$ is one of the y_i's, i,e., the k-th one.

Let $\boldsymbol{\theta} = (\mathbf{z}_0, d_0, t_0)$ be the vector of earthquake parameters to be estimated, i.e. the coordinates of the epicenter, $\mathbf{z}_0 = (lon_0, lat_0)$, the earthquake depth, d_0, and the event origin time, t_0. Let $\mathbf{z}_i = (lon_i, lat_i)$ represent the spatial coordinates of the i-th device. If the observation is uncensored, it contributes $f(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ to the likelihood. Conversely, if it is censored, the contribution to the likelihood is $S(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$. Specifically

$$\mathcal{L}(\boldsymbol{\theta}, \pi, \alpha; \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)^{\delta_i} S(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)^{1-\delta_i} = \prod_{i=1}^{n} h(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)^{\delta_i} S(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i), \tag{3.11}$$

where Equation (3.1) has been exploited, $\mathbf{y} = (y_1, \dots, y_n)$, $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$. Moreover, $h(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ and $S(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ are the hazard and survival functions defined in Equations (3.5) and (3.6) where we explicitly denote the dependence on $\boldsymbol{\theta}$ and $\mathbf{z}_i$.

3.4 P- and S-waves density functions

As mentioned above, earthquakes are characterized by primary and secondary waves. A higher velocity characterizes the P-wave and typically causes minimal damage, whereas the S-wave, which travels at a slower velocity, is powerful and damaging. In our analysis, we introduce the unknown velocities of the P- and S-waves, denoted v_P and v_S respectively, whose ratio in crustal rocks is assumed to be $v_P/v_S = \sqrt{3}$ (Suslick, 2001). Smartphones may detect the P- or the S-wave depending on the ground shaking intensity caused by each wave at the smartphone location.

Given the earthquake origin time t_0, the hypocenter $(\mathbf{z}_0, d_0)$ and the coordinates of the smartphone, the expected times at which the P- and S-waves reach the device can be easily calculated. Due to the operation of the detection algorithm implemented on the smartphone and the latency of the internet, the triggering can occur within 3.5 seconds of the wave arrival. Hence, a possible choice for $f_P(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ and $f_S(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ are uniform distributions over the 3.5 seconds of time support where the trigger can occur. The left bounds of the supports for the P- and S-waves are defined as follows

$$a_P = \frac{dist_H(\mathbf{z}_i, \mathbf{z_0}, d_0)}{v_P} + t_0 \qquad a_S = \frac{dist_H(\mathbf{z}_i, \mathbf{z_0}, d_0)}{v_S} + t_0,$$

where v_P and v_S are the P- and S-wave velocities, respectively, and $dist_H(\mathbf{z}_i, \mathbf{z_0}, d_0)$ is the Euclidean distance between smartphone i and the hypocenter $(\mathbf{z}_0, d_0)$, which is an approximation of the distance traveled by the seismic waves.

Unfortunately, uniform densities pose numerical challenges since the likelihood function in Equation (3.11) becomes discontinuous. Thus, any algorithm to explore the posterior parameter space will struggle to find sharp and reliable solutions. For this reason, we assume $f_P(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ and $f_S(y_i \mid \boldsymbol{\theta}, \mathbf{z}_i)$ as the densities of a normal distribution with mean $\mu_P(\boldsymbol{\theta}, \mathbf{z}_i)$ and $\mu_S(\boldsymbol{\theta}, \mathbf{z}_i)$, respectively, and standard

deviation τ, where

$$\mu_P(\boldsymbol{\theta}, \mathbf{z}_i) = \frac{dist_H(\mathbf{z}_i, \mathbf{z_0}, d_0)}{v_P} + t_0 + \mu \qquad \mu_S(\boldsymbol{\theta}, \mathbf{z}_i) = \frac{dist_H(\mathbf{z}_i, \mathbf{z_0}, d_0)}{v_S} + t_0 + \mu.$$

The parameter μ is selected to locate the mean of each component at the center of the 3.5-second interval defined by the uniform mixture. On the other hand, the parameter τ is assumed unknown and estimated from the data. To clarify, Figure 3.3 depicts both the uniform mixture and the Gaussian mixture, with weights 0.6 and 0.4 for the P- and S-waves, respectively. In the figure, different values of τ are considered, i.e., $\tau = \{0.67, 1.75, 3.5\}$, corresponding to situations where each component has approximately 99%, 68%, and 38% of the probability to fall within the 3.5 second interval. The density functions are displayed in panel (a), while panel (b) presents the corresponding survival functions.

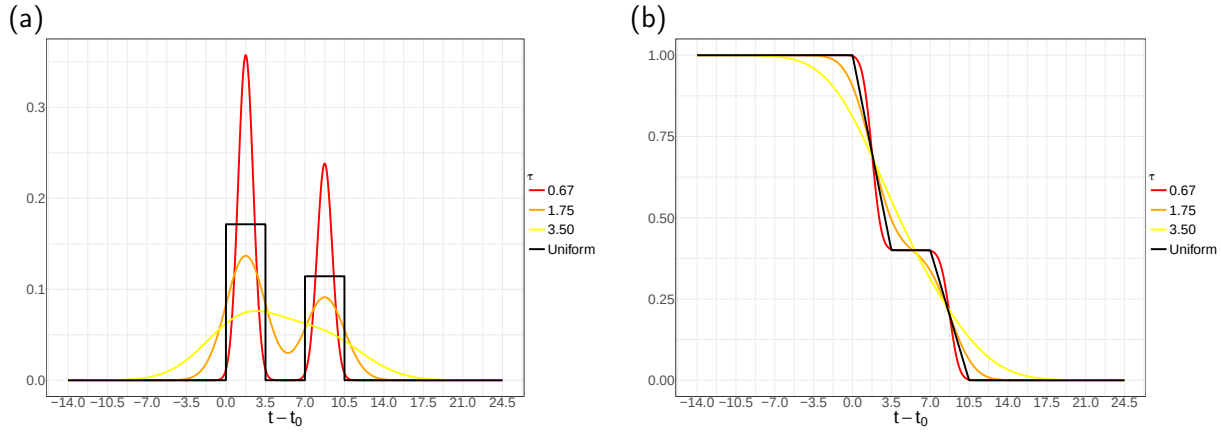


Figure 3.3: Uniform (black) vs Gaussian (red, orange and yellow) mixture with 0.6 and 0.4 weights of the P and S waves, respectively, for a smartphone located at the epicenter $\mathbf{z}_0$.

We refer to the survival model (3.5)-(3.6) together with the P- and S-waves mixture density as the Survival EarthQuake Model (SEQM).

3.5 Prior distributions

The Bayesian hierarchy of the SEQM is completed by specifying the prior distributions for all of the model parameters. Let $\mathbf{z}^* = (lon^*, lat^*)$ be the EQN detection location, which is a rough estimate of the earthquake epicenter computed as the center of mass of the triggers that contributed to the earthquake detection. For the epicenter coordinates $\mathbf{z}_0$, we place a bivariate normal distribution centered on $\mathbf{z}^*$ with diagonal covariance matrix, i.e., $\mathbf{z}_0 \sim \mathcal{N}_2\left(\mathbf{z}^*, \sigma^2 I_2\right)$, where $\sigma = 1$. We model a priori the depth parameter d_0 using a Gamma distribution whose shape and rate parameters are chosen to yield 95% of the prior probability laying between 5 and 40 kilometers, i.e., $d_0 \sim \text{Gamma}(3.94, 0.24)$. For the origin time t_0, a uniform distribution with support between 0 and 120 is assumed, i.e., the earthquake can occur at any moment between two minutes before the EQN detection and the detection time of the seismic event. As for the wave mixing weight α, we set a $\text{Beta}(1/2, 1/2)$ distribution to assume a priori that only one wave is detected by the smartphone network. We model the P-wave velocity v_P using a Gamma prior distribution. We select shape and rate parameters

to yield a prior mean velocity of 6.5 km/s and ensure that approximately 95% of the distribution falls within 5 to 8 km/s. This translates to a prior $v_P \sim \text{Gamma}(69.95, 10.94)$. For the standard deviation of the mixture components, we assume a uniform distribution, as in Gelman (2006), with values ranging from 0 to 3.5, i.e., $\tau \sim \mathcal{U}(0, 3.5)$, see Figure 3.3. Additionally, we place a Uniform distribution on the faulty fraction π with values between zero and one, i.e., $\pi \sim \mathcal{U}(0, 1)$. Finally, for the average triggering time unrelated to earthquakes λ^{-1}, we place a Uniform distribution over the range from 800 to 80,000 seconds, i.e., $\lambda^{-1} \sim \mathcal{U}(800, 80000)$. This choice allows for smartphone networks with different background noise levels (i.e., different rates of non-seismic triggering).

3.6 Computational details

To approximate the joint posterior distribution under the SEQM with the prior described in Section 3.5, we develop an MCMC sampling strategy utilizing an adaptive parallel tempering Metropolis-within-Gibbs sampler. The general concept of this approach is outlined in Section 3.6.1, while the algorithm's specific implementation for our case is detailed in Section B.1.

We denote $\phi = (\phi_1, \phi_2, \phi_3, \phi_5, \phi_5, \phi_6) = (\boldsymbol{\theta}, \alpha, v_P, \tau, \pi, \lambda)$ the vector of model parameters, where, we recall that $\boldsymbol{\theta} = (\mathbf{z}_0, d_0, t_0)$. The six blocks of parameters are a priori independent so that the posterior results in

$$p(\phi \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}) \propto \mathcal{L}(\phi; \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}) p(\phi) = \mathcal{L}(\phi; \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}) \prod_{k=1}^{6} p_k(\phi_k), \tag{3.12}$$

where $p_k(\phi_k)$ is the prior of the k-th parameter block. Moreover, to simplify the Metropolis step implementation, we map each parameter from its support to the real line and propose a new state from a Gaussian distribution in the mapped space. More details about the single transformations and how to modify the acceptance-rejection ratio are given in Section 3.6.2 of the Appendices. Finally, the adaptive strategy by Miasojedow et al. (2013) is adopted to tune the variance of the proposal distribution. For details about the adaptive MCMC schemes, we refer to Andrieu and Thoms (2008); Roberts and Rosenthal (2001); Atchadé and Rosenthal (2005).

3.6.1 Parallel tempering MCMC

A preliminary study showed how the posterior distribution, under our model, has numerous well-separated modes. Indeed, starting from different initial states, the classical random-walk Metropolis-within-Gibbs became trapped, each time, in a different mode, failing to recover the whole posterior distribution. To deal with this issue, we adopted a parallel tempering MCMC (Woodard et al., 2009), also known as Metropolis Coupled Markov Chain Monte Carlo (Altekar et al., 2004).

The algorithm tackles the multi-modality issue by performing – in parallel – several MCMC steps at various *inverse temperatures*. The latter constitute a set of L parameters $\beta_l \in (0, 1]$, $l = 1, \dots, L$ such that $1 = \beta_0 > \beta_1 > \cdots > \beta_L > 0$. In particular, each of the MCMC algorithms running in parallel is designed to draw values from a modified version of the posterior, given by

$$p(\phi \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta})^{\beta_l} \propto [\mathcal{L}(\phi; \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}) p(\phi)]^{\beta_l}.$$

In this way, as $\beta_l \to 0$, i.e., the temperature gets hotter, the multi-modal posterior flattens, and the corresponding chain better explores the posterior's support. Conversely, the coldest chain, i.e., the one with $\beta_0 = 1$, coincides with the posterior of interest. Notably, the states at various temperatures are continuously proposed to be switched during the run. A Metropolis step is employed to accept or reject the switch so that information coming from the flattest chain (β_L) is allowed to reach the original one ($\beta_0 = 1$). The parameter L, representing the number of distinct temperatures, is selected to attain an asymptotic rate of approximately 44% swaps between adjacent chains (Miasojedow et al., 2013). The coldest chain is facilitated in exploring all the modes of the posterior distribution. More precisely, at any iteration g, after the Metropolis-within-Gibbs step, for each block of the parameters, given the L current values of the chains $\left[\boldsymbol{\phi}_1^{(g)}, \dots, \boldsymbol{\phi}_L^{(g)}\right]$, the swap is proposed by uniformly sampling $\overline{l}$ from $\{1, \dots, L\}$, and swapping the current states at temperature levels $\overline{l}$ and $\overline{l}+1$ with probability given by

$$\omega\left(\boldsymbol{\phi}_{\overline{l}}^{(g)}, \boldsymbol{\phi}_{\overline{l}+1}^{(g)} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}\right) = \min\left\{1, \left(\frac{p\left(\boldsymbol{\phi}_{\overline{l}+1}^{(g)} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}\right)}{p\left(\boldsymbol{\phi}_{\overline{l}}^{(g)} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}\right)}\right)^{\beta_{\overline{l}} - \beta_{\overline{l}+1}}\right\}.$$

When implementing a tempering MCMC, choosing the inverse temperature parameters β_l, $l = 1, \dots, L$ is crucial. Indeed, selecting temperatures that are too similar can lead to inadequate exploration of the parameter space. On the contrary, selecting temperatures that are too far apart from one another may lead to significantly low acceptance rates for swaps. To avoid this issues, following Miasojedow et al. (2013), we opt for an adaptive strategy to tune a function of the temperatures spacing, i.e., $\rho_l = \log(1/\beta_{l+1} - 1/\beta_l)$. Specifically, we fix the number of temperature levels L and target a specific mean acceptance rate of the swaps between the chains in nearby temperatures. Following the guidelines by (Atchadé and Rosenthal, 2005), we set this value to 0.41. More in detail, at each iteration g, we update $\rho_{\overline{l}}$ by adding a function of the difference between the current accept/reject probability ω and the targeted one. In this way, if the swap acceptance probability is too high, the log distance between $\overline{l}$ and $\overline{l}+1$ temperatures increases, and if it is too low, it decreases.

We point out that Miasojedow et al. (2013) provide theoretical guarantees that the adaptive tempered algorithm eventually expresses the fixed acceptance rate for both the Metropolis acceptance rate within each block and the adjacent states swap. The same paper shows that the L chains drawn in parallel by the algorithm are geometrically ergodic, so both the strong law of large numbers and the central limit theorem hold for these chains. Section 3.6.2 of the Appendices provides a detailed description of the algorithm. The `R` code implementing the algorithm is available at `https://github.com/lucaaiello/SEQM`. All analyses were performed on an Intel(R) Core(TM) i7-10750H CPU processor with a base frequency of 2.60 GHz. The computational cost for running the MCMC algorithm, in seconds for iteration, is reported in Tables C.1 - C.2 and Table 3.4 for the simulation study and the real data analysis, respectively.

As a final note, it is important to consider that traditional methods to summarize multi-modal posterior densities, such as the posterior mean, median, or quantile-based credible intervals, may not be suitable. In this study, we use the highest posterior mode as a point estimate of the parameter of interest, including the epicenter. At the same time, uncertainty is evaluated using the highest posterior density regions (HPDRs). HPDRs are particularly useful in the case of multi-modal distributions, as they can include multiple disjoint subsets, one for each local mode, providing a more accurate

representation of the underlying posterior framework.

3.6.2 MCMC scheme

Recall that we have defined $\phi = (\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6) = (\boldsymbol{\theta}, \alpha, \pi, v_P, \tau, \lambda^{-1})$. From Equation (3.12), we define $p_k(\phi_k \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}) \propto \mathcal{L}(\phi; \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}) p_k(\phi_k)$ the full conditional distribution of each block of ϕ. Let $J_k\left(\widetilde{\phi}_k\right)$ be the Jacobian of the transformation function $m_k : \mathbb{R}^{d_k} \to S_k$, where $S_k = [lb_{1_k}, ub_{1_k}] \times \cdots \times [lb_{d_k}, ub_{d_k}]$, lb_{i_k} and ub_{i_k} are the bounds of the support of the i_k-th parameter within the k-th block, d_k is the dimension of the k-th block and $\widetilde{\phi}_k = m_k^{-1}(\phi_k)$. We generalize here the notation by referring to $\phi_{k,l}^{(g)}$ as the value at iteration g of the k-th subvector of $\phi_l^{(g)} = \left(\phi_{1,l}^{(g)}, \ldots, \phi_{K,l}^{(g)}\right)$ regarding the chain targeting the l-th version of the posterior, i.e., $p(\phi_l \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta})^{\beta_l}$, with $l = 1, \ldots, L$.

Algorithm 2 provides the details of the MCMC scheme we designed to approximate the joint posterior distribution under the SEQM. In our application, we assumed $K = 6$ blocks with dimension $\mathbf{d} = (d_1, d_2, d_3, d_4, d_5, d_6) = (4, 1, 1, 1, 1, 1)$ and $L = 10$. Regarding the transformations, the parameter supports are $[lb_{1_1}, ub_{1_1}] = [-90, 90]$, $[lb_{2_1}, ub_{2_1}] = [-180, 180]$, $[lb_{3_1}, ub_{3_1}] = [0, \infty)$, $[lb_{4_1}, ub_{4_1}] = [0, 120]$, $[lb_{1_2}, ub_{1_2}] = [0, 1]$, $[lb_{1_3}, ub_{1_3}] = [0, 1]$, $[lb_{1_4}, ub_{1_4}] = [0, \infty)$, $[lb_{1_5}, ub_{1_5}] = [0, 3.5]$ and $[lb_{1_6}, ub_{1_6}] = [800, 80000]$. Hence, defining $\widetilde{\phi}_1 = \left(\widetilde{lat}_0, \widetilde{lon}_0, \widetilde{d}_0, \widetilde{t}_0\right)$, $\widetilde{\phi}_2 = \widetilde{\alpha}$, $\widetilde{\phi}_3 = \widetilde{\pi}$, $\widetilde{\phi}_4 = \widetilde{v}_P$, $\widetilde{\phi}_5 = \widetilde{\tau}$ and $\widetilde{\phi}_6 = \widetilde{\lambda}^{-1}$ transformation functions are defined as:

$$
\begin{aligned}
m_1\left(\widetilde{\phi}_1\right) &= \left(g\left(\widetilde{lat}_0; -90, 90\right), g\left(\widetilde{lon}_0; -180, 180\right), e^{\widetilde{d}_0}, g\left(\widetilde{t}_0; 0, 120\right)\right)^\top \\
m_2\left(\widetilde{\phi}_2\right) &= g\left(\widetilde{\alpha}; 0, 1\right) \\
m_3\left(\widetilde{\phi}_3\right) &= g\left(\widetilde{\pi}; 0, 1\right) \\
m_4\left(\widetilde{\phi}_4\right) &= e^{\widetilde{v}_P} \\
m_5\left(\widetilde{\phi}_5\right) &= g\left(\widetilde{\tau}; 0, 3.5\right) \\
m_6\left(\widetilde{\phi}_6\right) &= g\left(\widetilde{\lambda}^{-1}; 800, 80000\right)
\end{aligned}
$$

where

$$
g(\widetilde{x}; lb, ub) = \frac{lb + ub\, e^{\widetilde{x}}}{1 + e^{\widetilde{x}}}.
$$

For each block and for $l = 1, \ldots, 10$ we set the starting values of the model parameters as $\widetilde{lat}_0^{(0)} = g^{-1}\left(lat_0^{(0)}; -90, 90\right)$, $\widetilde{lon}_0^{(0)} = g^{-1}\left(lon_0^{(0)}; -180, 180\right)$, $\widetilde{d}_0^{(0)} = \log\left(d_0^{(0)}\right)$, $\widetilde{t}_0^{(0)} = g^{-1}\left(t_0^{(0)}; 0, 120\right)$ and set $\widetilde{\phi}_{1,l}^{(0)} = \left(\widetilde{lat}_0^{(0)}, \widetilde{lon}_0^{(0)}, \widetilde{d}_0^{(0)}, \widetilde{t}_0^{(0)}\right)$. In particular $\mathbf{z}_0^{(0)} = \left(lat_0^{(0)}, lon_0^{(0)}\right)$ is sampled from a uniform centered on $\mathbf{z}^*$ with a two degree wide support, $d_0^{(0)}$ is sampled from a uniform between 5 and 100 km and $t_0^{(0)}$ is sampled from a uniform between 0 and 120s. For $\widetilde{\phi}_{2,l}^{(0)} = g^{-1}\left(\alpha^{(0)}; 0, 1\right)$, $\widetilde{\phi}_{3,l}^{(0)} = g^{-1}\left(\pi^{(0)}; 0, 1\right)$ we sample $\alpha^{(0)}$ and $\pi^{(0)}$ from a uniform distribution between 0 and 1, while for $\widetilde{\phi}_{4,l}^{(0)} = \log\left(v_P^{(0)}\right)$, $\widetilde{\phi}_{5,l}^{(0)} = g^{-1}\left(\tau^{(0)}; 0, 3.5\right)$ and $\widetilde{\phi}_{6,l}^{(0)} = g^{-1}\left(\lambda^{-1(0)}; 800, 80000\right)$ we sample $v_P^{(0)}$, $\tau^{(0)}$ and $\lambda^{-1(0)}$ from a uniform distributions with support between 3 and 8, 0 and 3.5, 800 and 80000, respectively. Moreover, we fix the target acceptance rates $\bar{\boldsymbol{\xi}} = \left(\bar{\xi}_1, \bar{\xi}_2, \bar{\xi}_3, \bar{\xi}_4, \bar{\xi}_5, \bar{\xi}_6\right) =$

$(0.23, 0.41, 0.41, 0.41, 0.41, 0.41)$, the adaptation step size sequences $\nu = 0.6$, and the target adjacent state swapping rate $\bar{\omega} = 0.41$. Finally, regarding the initial values of the adaptation parameters, we set chain means $\boldsymbol{\mu}_{k,l}^{(0)} = \widetilde{\boldsymbol{\phi}}_{k,l}^{(0)}$, scaling parameters $s_{k,l}^{(0)} = 0.1$, covariance estimates $\mathbf{R}_{1,l}^{(0)} = \text{diag}(0.1, 0.1, 10, 1)$, $\mathbf{R}_{2,l}^{(0)} = 0.1$, $\mathbf{R}_{3,l}^{(0)} = 0.1$, $\mathbf{R}_{4,l}^{(0)} = 0.1$, $\mathbf{R}_{5,l}^{(0)} = 0.1$, $\mathbf{R}_{6,l}^{(0)} = 0.1$ proposal covariances $\boldsymbol{\zeta}_{k,l}^{(0)} = \exp\left(s_{k,l}^{(0)}\right)\mathbf{R}_{k,l}^{(0)}$, and the temperatures spacing logarithm $\rho_l^{(0)} = 1$.

3.7 Simulation study

The simulation study aims to evaluate the performance of our modeling approach in estimating the earthquake parameters under different scenarios that mimic real cases. The key parameters of the simulation study are: the number of active smartphones, the location of the epicenter with respect to the smartphone network, and the detection threshold of the EQN detection algorithm (i.e., the ratio of triggering smartphones to active smartphones required to trigger an earthquake detection). To simulate realistic smartphone networks, we adopted a dataset of nearly 181,000 smartphone locations observed by the EQN in the Marmara region of Turkey (see Figure 3.6). From the dataset, random networks of $n = \{250, 500, 1000, 2000\}$ smartphone are generated. For the earthquake epicenter, two locations were considered: one inside the smartphone network and one offshore outside the smartphone network. In addition, three detection thresholds for the EQN algorithm (Finazzi and Fassò, 2017) were adopted, namely $r = \{0.10, 0.15, 0.20\}$, which determine the order statistics on the triggering times and thus the EQN detection time. This brings the total number of scenarios to 24.

For each scenario, 100 simulation runs are performed with fixed model parameters ($\alpha = 0.8$, $\pi = 0.3$, $\lambda^{-1} = 4800$ and $v_P = 7.8$). In each run, triggers are first simulated for the n smartphones. Two triggering times are simulated independently: a triggering time due to the arrival of the seismic wave and a triggering time due to causes unrelated to the earthquake. The first triggering time is generated from the mixture of uniform distributions shown in Figure 3.3, assuming a maximum triggering delay of 3.5 seconds. The second triggering time is instead simulated from an exponential distribution with parameter λ and support starting 120 seconds before the origin time. Only the minimum between the two triggering times is retained. Also, a fraction π of smartphones that triggered because of the earthquake are randomly selected as faulty, so their triggering times are discarded. All simulated triggers are ordered by their triggering time, and the EQN detection algorithm is then applied to the list of ordered triggers. This simulates earthquake detection by the EQN server as if the triggers were received one at a time. If the earthquake is detected, future triggering times are censored (with respect to EQN detection). Finally, the observational time window is shifted to start from 120 seconds before the EQN detection time, and all the triggering times and the origin time are shifted accordingly.

To illustrate how this trigger list is defined, Figure 3.4 presents an example of the observational time window within which the triggering time (t) is measured. This window extends from 120 seconds before the EQN detection to EQN time, which serves as the censoring time t^*. For reference, the figure also includes the earthquake origin time t_0.

In each simulation run, randomness affects the locations of smartphones, their trigger times (since the epicenter locations and model parameters remain fixed), and the EQN detection. For example,

Algorithm 2 Adaptive Tempered Metropolis within Gibbs algorithm

1: Initialization: the number of blocks K, their dimensions $\mathbf{d}$, the number of temperatures L, the required MCMC sample size G, burn-in period g_0, the starting values $\widetilde{\boldsymbol{\phi}}_{k,l}^{(0)}$, $\boldsymbol{\mu}_{k,l}^{(0)}$, $\boldsymbol{\zeta}_{k,l}^{(0)}$, $s_{k,l}^{(0)}$, $\mathbf{R}_{k,l}^{(0)}$ and $\rho_l^{(0)}$ for $k = 1, \ldots, K$, $l = 1, \ldots, L$, the acceptance probability rates $\bar{\boldsymbol{\xi}}$, the adaptation step size sequences ν, and the adjacent state swapping rate $\bar{\omega}$.

2: **for** $g = 1, 2, \ldots, G$ **do**

3: **for** $k = 1, \ldots, K$ **do**

4: **for** $l = 1, \ldots, L$ **do**

5: Sample $\widetilde{\boldsymbol{\phi}}'_{k,l} = \widetilde{\boldsymbol{\phi}}_{k,l}^{(g-1)} + \boldsymbol{\epsilon}_{g,k,l}$ where $\boldsymbol{\epsilon}_{g,k,l} \sim \mathcal{N}_{d_k}\left(\mathbf{0}, \boldsymbol{\zeta}_{k,l}^{(g-1)}\right)$

6: Let the next iterate be

$$\widetilde{\boldsymbol{\phi}}_{k,l}^{(g)} = \begin{cases} \widetilde{\boldsymbol{\phi}}'_{k,l} & \text{with probability } \xi_{k,l}^{(g)}\left(\boldsymbol{\phi}_{k,l}^{(g-1)}, \widetilde{\boldsymbol{\phi}}_{k,l}^{(g-1)}, \boldsymbol{\phi}'_{k,l}, \widetilde{\boldsymbol{\phi}}'_{k,l} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}, \beta_l^{(g-1)}\right) \\ \widetilde{\boldsymbol{\phi}}_{k,l}^{(g-1)} & \text{with probability } 1 - \xi_{k,l}^{(g)}\left(\boldsymbol{\phi}_{k,l}^{(g-1)}, \widetilde{\boldsymbol{\phi}}_{k,l}^{(g-1)}, \boldsymbol{\phi}'_{k,l}, \widetilde{\boldsymbol{\phi}}'_{k,l} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}, \beta_l^{(g-1)}\right) \end{cases}$$

where

$$\boldsymbol{\phi}'_{k,l} = m_k\left(\widetilde{\boldsymbol{\phi}}'_{k,l}\right)$$

$$\xi_{k,l}^{(g)}\left(\boldsymbol{\phi}_{k,l}^{(g-1)}, \widetilde{\boldsymbol{\phi}}_{k,l}^{(g-1)}, \boldsymbol{\phi}'_{k,l}, \widetilde{\boldsymbol{\phi}}'_{k,l} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}, \beta_l^{(g-1)}\right) = \min\left\{1, \left(\frac{p_k\left(\boldsymbol{\phi}'_{k,l} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}\right)\left|J_k\left(\widetilde{\boldsymbol{\phi}}'_{k,l}\right)\right|}{p_k\left(\boldsymbol{\phi}_{k,l}^{(g-1)} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}\right)\left|J_k\left(\widetilde{\boldsymbol{\phi}}_{k,l}^{(g-1)}\right)\right|}\right)^{\beta_l^{(g-1)}}\right\}$$

7: Compute

$$s_{k,l}^{(g)} = s_{k,l}^{(g-1)} + (g+1)^{-\nu}\left(\xi_{k,l}^{(g)}\left(\boldsymbol{\phi}_{k,l}^{(g-1)}, \widetilde{\boldsymbol{\phi}}_{k,l}^{(g-1)}, \boldsymbol{\phi}'_{k,l}, \widetilde{\boldsymbol{\phi}}'_{k,l} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}, \beta_l^{(g-1)}\right) - \bar{\xi}_k\right)$$

$$\mathbf{R}_{k,l}^{(g)} = \left(1 - (g+1)^{-\nu}\right)\mathbf{R}_{k,l}^{(g-1)} + (g+1)^{-\nu}\left(\widetilde{\boldsymbol{\phi}}_{k,l}^{(g)} - \boldsymbol{\mu}_{k,l}^{(g-1)}\right)\left(\widetilde{\boldsymbol{\phi}}_{k,l}^{(g)} - \boldsymbol{\mu}_{k,l}^{(g-1)}\right)^{\top}$$

$$\boldsymbol{\mu}_{k,l}^{(g)} = \left(1 - (g+1)^{-\nu}\right)\boldsymbol{\mu}_{k,l}^{(g-1)} + (g+1)^{-\nu}\widetilde{\boldsymbol{\phi}}_{k,l}^{(g)}$$

8: and

$$\boldsymbol{\zeta}_{k,l}^{(g)} = \exp\left(s_{k,l}^{(g)}\right)\mathbf{R}_{k,l}^{(g)}$$

9: **end for**

10: **end for**

11: Sample $\overline{l} \in \{1, \ldots, L-1\}$ uniformly and swap states $\overline{l}$ and $\overline{l}+1$ with probability

$$\omega_{\overline{l}}^{(g)}\left(\boldsymbol{\phi}_{\overline{l}}^{(g)}, \widetilde{\boldsymbol{\phi}}_{\overline{l}}^{(g)}, \boldsymbol{\phi}_{\overline{l}+1}^{(g)}, \widetilde{\boldsymbol{\phi}}_{\overline{l}+1}^{(g)} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}, \beta_{\overline{l}}^{(g-1)}, \beta_{\overline{l}+1}^{(g-1)}\right) = \min\left\{1, \left(\frac{p\left(\boldsymbol{\phi}_{\overline{l}+1}^{(g)} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}\right)\left|J\left(\widetilde{\boldsymbol{\phi}}_{\overline{l}+1}^{(g)}\right)\right|}{p\left(\boldsymbol{\phi}_{\overline{l}}^{(g)} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}\right)\left|J\left(\widetilde{\boldsymbol{\phi}}_{\overline{l}}^{(g)}\right)\right|}\right)^{\beta_{\overline{l}}^{(g-1)} - \beta_{\overline{l}+1}^{(g-1)}}\right\}$$

12: Compute

$$\rho_{\overline{l}}^{(g)} = \rho_{\overline{l}}^{(g-1)} + (g+1)^{-\nu}\left(\omega_{\overline{l}}^{(g)}\left(\boldsymbol{\phi}_{\overline{l}}^{(g)}, \widetilde{\boldsymbol{\phi}}_{\overline{l}}^{(g)}, \boldsymbol{\phi}_{\overline{l}+1}^{(g)}, \widetilde{\boldsymbol{\phi}}_{\overline{l}+1}^{(g)} \mid \mathbf{y}, \mathbf{Z}, \boldsymbol{\delta}, \beta_{\overline{l}}^{(g-1)}, \beta_{\overline{l}+1}^{(g-1)}\right) - \bar{\omega}\right)$$

13: Set $\beta_0^{(g)} = 1$ and

14: **for** $l = 1, \ldots, L-1$ **do**

$$\frac{1}{\beta_{l+1}^{(g)}} = \frac{1}{\beta_l^{(g)}} + \exp\left(\rho_l^{(g)}\right)$$

15: **end for**

16: **end for**

17: Return for $l = 1, \ldots, L$ draws $\boldsymbol{\phi}_l^{(g_0+1)}, \ldots, \boldsymbol{\phi}_l^{(G)}$

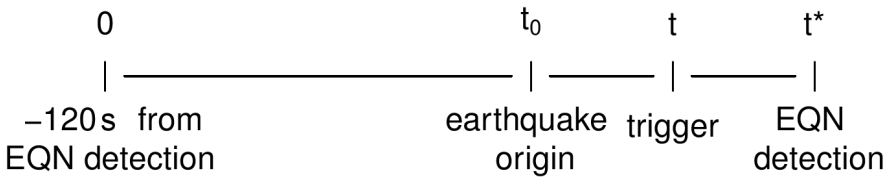


Figure 3.4: Example of the observational time window.

Table 3.2: Modes of the true origin time over the 100 simulated datasets for all scenarios with the epicenter inside and outside the network.

		$n = 250$	$n = 500$	$n = 1000$	$n = 2000$
$r = 0.10$	Inside	115	115	115	116
	Outside	105	106	106	106
$r = 0.15$	Inside	114	115	115	115
	Outside	105	105	105	105
$r = 0.20$	Inside	113	113	113	114
	Outside	104	104	104	104

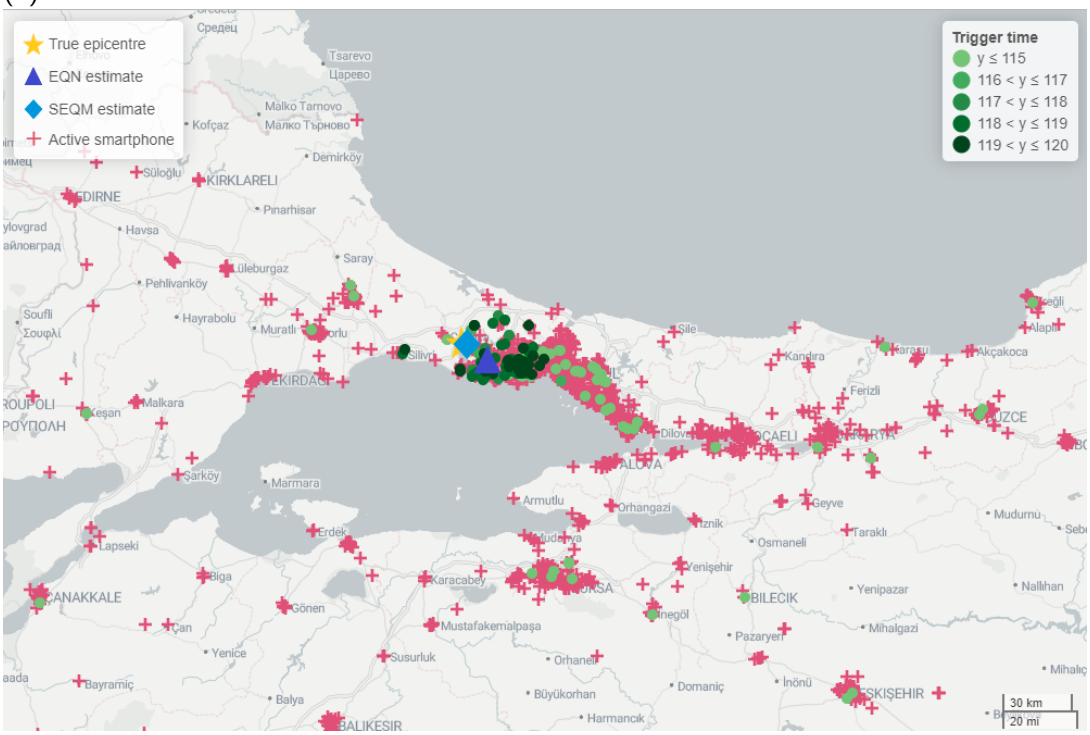
Figure 3.5 presents two simulated datasets: one with the true epicenter located inside the network and the other with the epicenter outside the network. Both figures show the EQN detection and the epicenter estimated by SEQM. As discussed in Section 3.1, the distribution of trigger times demonstrates significant variability and noise, which is also evident in Figure 3.5. Additionally, Table 3.2 highlights the modes of the simulated origin times across different scenarios. When the epicenter is inside the network, EQN detection typically occurs 4 to 5 seconds after the true origin time. In contrast, for external epicenters, detection occurs between 15 and 16 seconds after the origin time.

Model estimation is finally performed considering the list of triggers and the list of active smartphones. Specifically, the prior distributions described in section 3.5 were used. The burn-in period was set to 5,000 iterations, followed by 25,000 post-burn-in iterations. We retained one every five iterations, resulting in 5,000 effective iterations for the subsequent analysis.

The boxplots in Figure 3.6 show the distribution of the geodetic distance (error) between the estimated and simulated epicenter for EQN and SEQM for different values of n under the scenarios with EQN detection threshold set to $r = 0.20$ and when the real epicenter is inside and outside the network, respectively. As expected, both the median and the error variability tend to decrease when n increases, enlarging the benefit of SEQM over EQN. Moreover, the results show that the improvement of SEQM over EQN is much higher in the scenario where the epicenter is located outside the network, which is indeed the situation where the EQN struggles to recover the true epicenter using only the triggered smartphones. Rather, SEQM also exploits spatial locations of the non-triggering smartphones to improve the earthquake parameter estimation. Similar conclusions hold for the other scenarios with varying EQN detection thresholds, whose results are postponed to the Supporting material, see Section C.1.

Table 3.3 illustrates, for the scenario where $r = 0.20$, the median and 95% interval of the origin time error, i.e., the difference between the true and the estimated origin time. When the true epicenter is located within the network, the median origin time error typically ranges between 1.23 and 1.64 seconds, with the 95% interval encompassing zero in cases where $n = 250$ or 500. In contrast, when

(a)



(b)

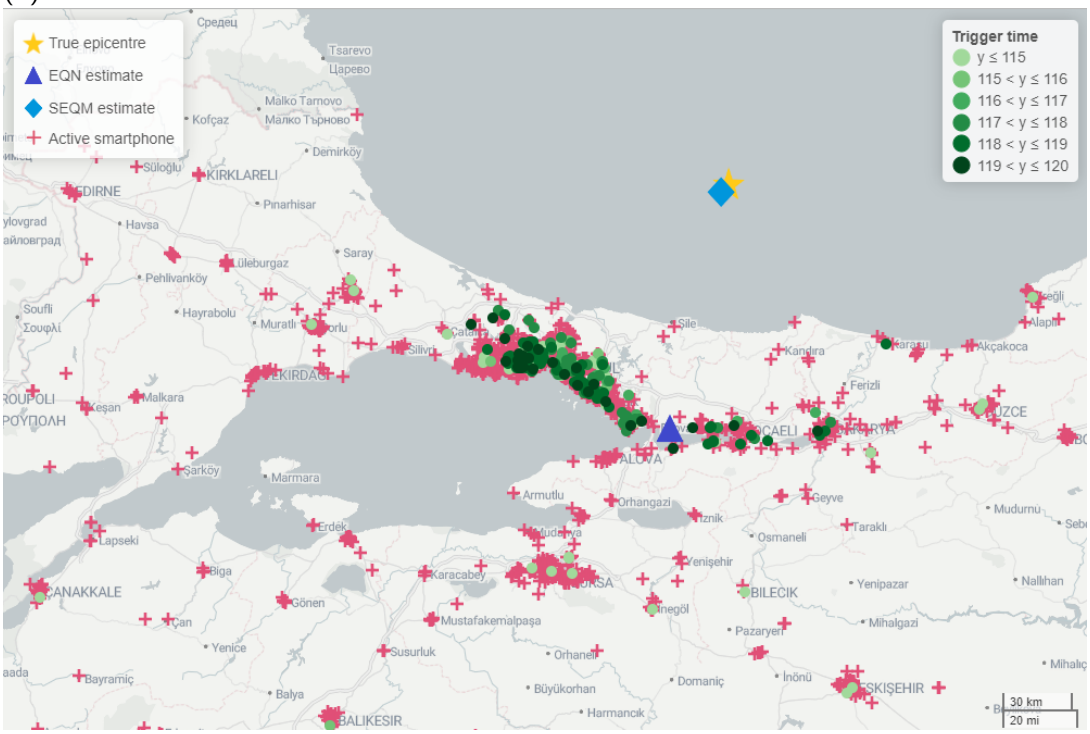


Figure 3.5: Examples of summary map of analysis performed on simulated data with epicenter enclosed in the network (a) and outside of the network (b).

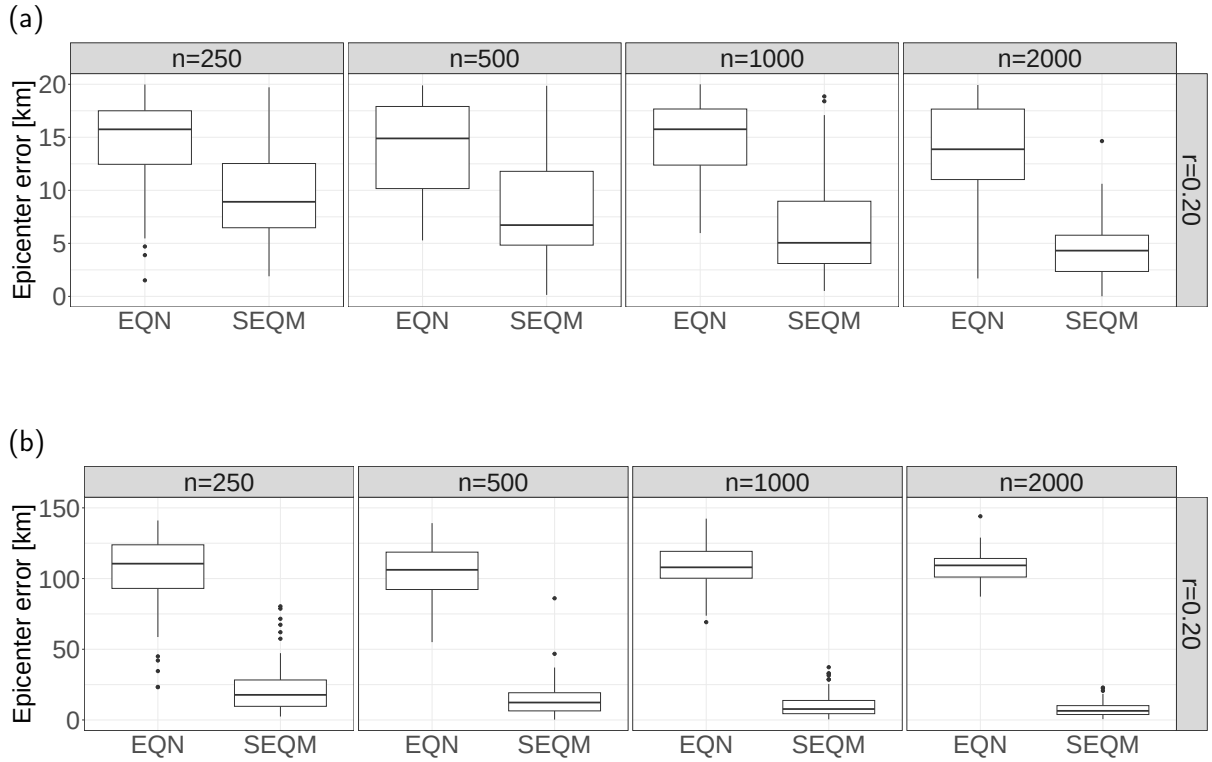


Figure 3.6: Boxplots for the epicenter error for EQN and SEQM in the case of $r = 0.20$, when the real epicenter is inside the network (a) and outside the network (b).

Table 3.3: Median and 95% interval of the origin time error $(t_0 - \hat{t}_0)$ $[s]$ over the 100 simulated datasets for $r = 0.20$.

Epicenter location	$n = 250$	$n = 500$	$n = 1000$	$n = 2000$
Inside the network	1.64	1.32	1.30	1.23
	[-0.37, 6.43]	[-0.05, 2.72]	[0.13, 2.76]	[0.30, 2.69]
Outside the network	2.01	1.90	1.79	1.70
	[-7.84, 7.19]	[-2.20, 7.23]	[-2.35, 5.39]	[-0.59, 4.95]

the true epicenter lies outside the network, the 95% intervals are wider and consistently include zero, but the median origin time error tends to be higher, ranging from 1.70 to 2 seconds. The bias on the origin time results directly from the simulation procedure. In fact, the triggering times are simulated from uniform distributions that induce an average delay of 1.75 seconds with respect to the arrival of the P- or S-wave. Section C.1 of the Supporting Material provides additional results of the simulation study for all scenarios, including the median and 95% interval of the estimated parameters over the simulated datasets, the empirical posterior coverage of the parameters along with their average HPDR length, and the average number of modes for each parameter.

3.8 Analysis of EQN datasets

For each dataset described in Section 3.1.1, earthquake parameters are estimated employing the survival model outlined in Equation (3.11), with the prior defined in Section 3.5 and running the algorithm detailed in Section 3.6 for 100,000 iterations with a 50,000 burn-in period.

Table 3.4: EQN and SEQM earthquake parameter estimates for the three events. For SEQM, also the 95% HPDR is provided.

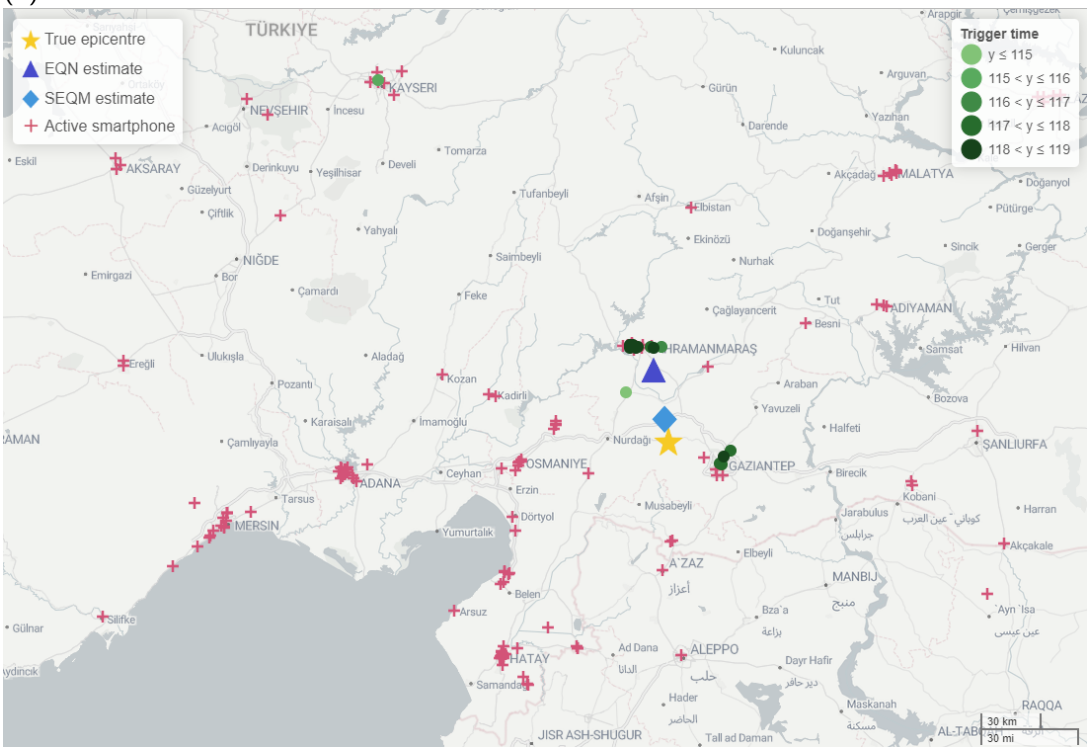
		Pazarcik	Ridgecrest	Mexico
EQN	lat^* [deg]	37.48	34.08	16.47
	lon^* [deg]	37.00	-117.57	-95.05
	t^* [hh:mm:ss]	01:17:46	03:20:33	06:26:17
SEQM	lat_0 [deg]	37.26	35.86	15.16
		[37.18, 37.36]	[35.63, 35.99]	[14.86, 15.32]
	lon_0 [deg]	37.06	-117.52	-95.01
		[36.92, 37.21]	[-117.65, -117.42]	[-95.66, -95.38] [-95.27, -94.83]
	d_0 [km]	12.99	12.83	13.94
		[2.28, 28.08]	[3.13, 39.07]	[2.85, 34.92]
	t_0 [hh:mm:ss]	01:17:36	03:19:48	06:25:46
		[01:17:29, 01:17:38]	[03:19:43, 03:19:52]	[06:25:42,06:25:48]
	α	0.01	0.02	0.34
		[0.00, 0.13] [0.32, 1.00]	[0.00, 0.07] [0.84, 1.00]	[0.22, 0.46]
	π	0.64	0.12	0.78
		[0.09 0.79]	[0.00, 0.29]	[0.72, 0.83]
	v_P [km/s]	6.07	7.73	8.80
		[4.65, 7.64]	[4.69, 5.25] [6.76, 8.66]	[7.95, 10.61]
	τ [s]	0.66	2.26	2.53
		[0.37, 1.95]	[1.77, 3.34]	[2.04, 3.46]
	λ^{-1} [s]	20174.69	1462.04	2030.59
		[7998.62, 77702.18]	[947.66, 2442.56]	[1617.72, 2725.68]
	Runtime [s/iteration]	0.015	0.024	0.057

The MCMC output is provided in Section C.2 of the Supporting material. Here, we summarize the results in Table 3.4 that includes the highest posterior mode and the associated 95% HPDRs of all model parameters for the three case studies. The table also reports the EQN estimates of the earthquake parameters. While EQN only provides point estimates, SEQM also offers uncertainty quantification through the HPDRs. The summary maps for the three earthquakes are presented in the left panel of Figures 3.7 - 3.9, which show the epicenter retrieved by the EMSC and the epicenter estimates provided by EQN and SEQM, where the latter is computed as the highest posterior mode. The right panel of the figures displays the posterior distribution of the epicenter together with its estimated posterior density and the EMSC epicenter along with its 95% confidence ellipse. Notice that, for the Pazarcik and Ridgecrest earthquakes, the marginal posterior distribution of the epicenter encompasses the EMSC ones; see also Table 3.4. In particular, for the Ridgecrest event, the EMSC epicenter is included in the highest bivariate posterior region, as highlighted in Figure 3.8. In the Mexican event, instead, the posterior density exhibits bimodality and is close to the EMSC epicenter, which is far away from the EQN estimate. In the SEQM section of the table, the computational cost for each case is reported, expressed in seconds per MCMC iteration. For instance, in the Pazarcik case, completing 100,000 iterations takes approximately 25 minutes. However, based on our tests, the algorithm demonstrated convergence with as few as 10,000 iterations, requiring only 2.5 minutes.

In Table 3.5, we assess the accuracy of the estimates by measuring the distance between the EMSC and the estimated epicenter, the distance between the EMSC and the estimated depth (only provided by SEQM), and the distance between the EMSC and the estimated origin time. Overall, SEQM is characterized by a lower error, thus improving EQN estimates.

In particular, for the Pazarcik event, epicenter and depth are estimated with an error of around 11 and 7km, respectively. In contrast, the error on the origin time is less than one second. The epicenter

(a)



(b)

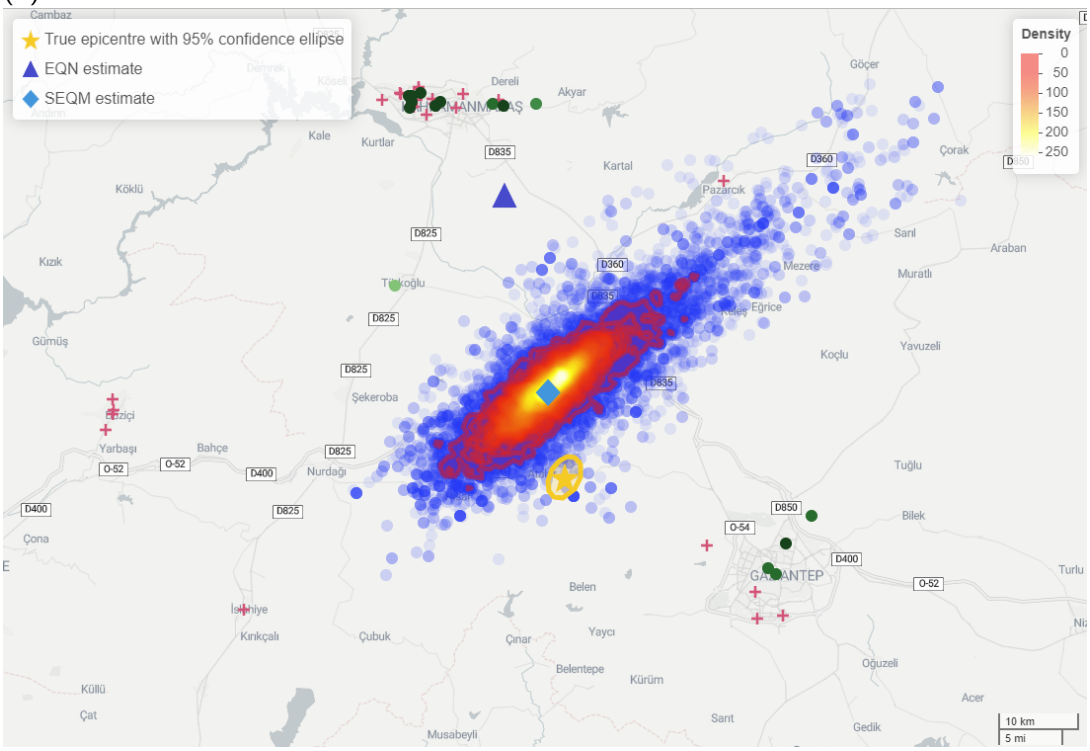
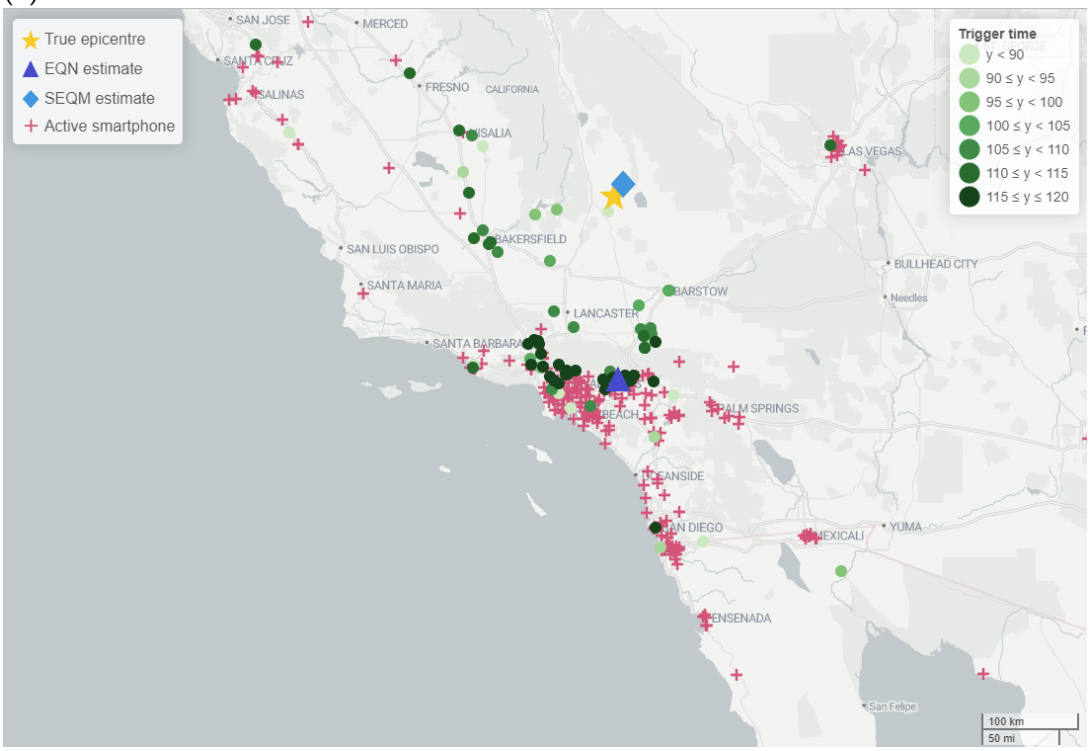


Figure 3.7: Summary map (a) and posterior distribution of the epicenter (blue dots) together with its estimated posterior density (b) for the 2023 Pazarcik earthquake.

(a)



(b)

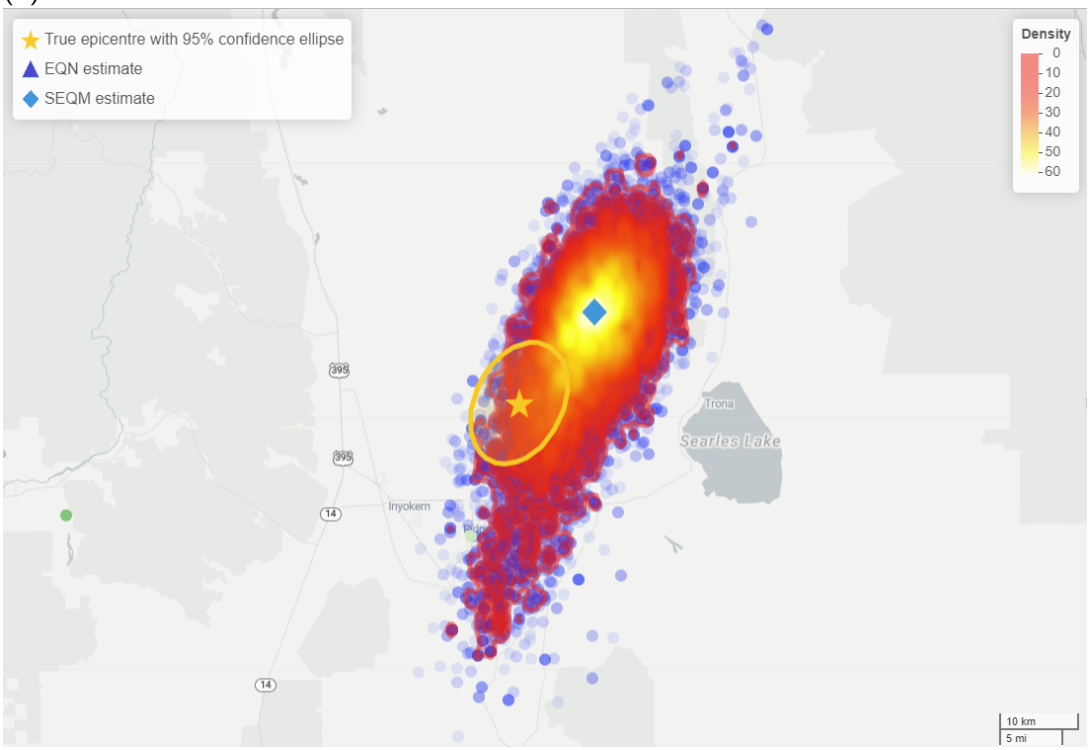
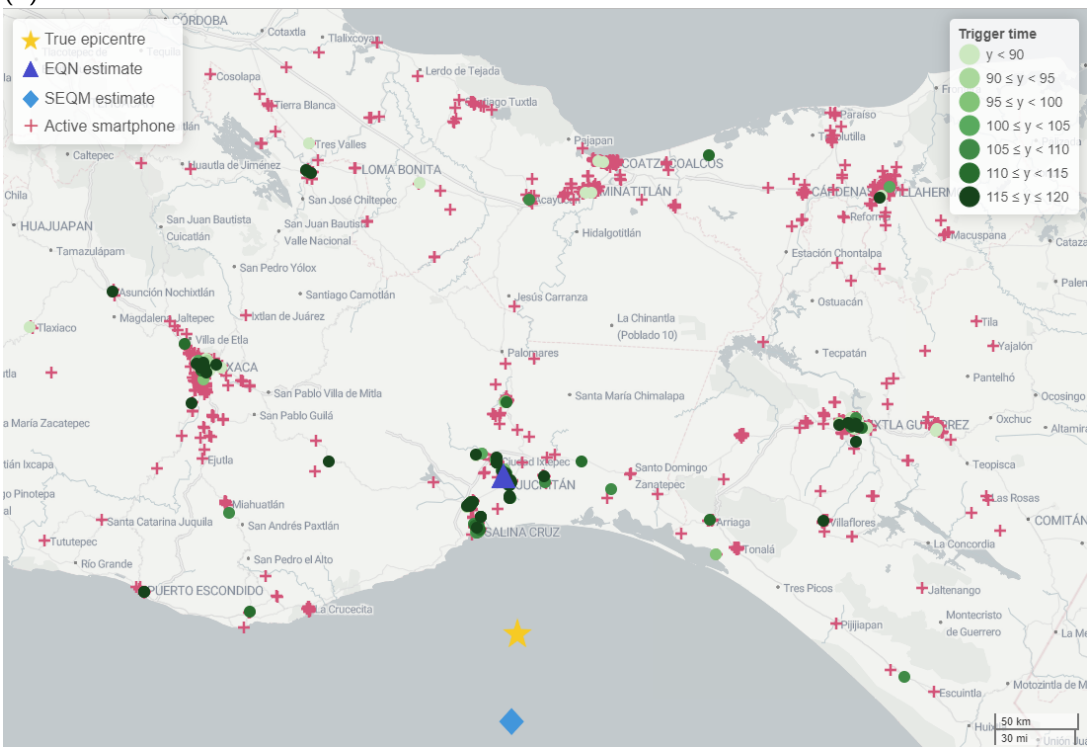


Figure 3.8: Summary map (a) and posterior distribution of the epicenter (blue dots) together with its estimated posterior density (b) for the 2019 Ridgecrest earthquake.

(a)



(b)

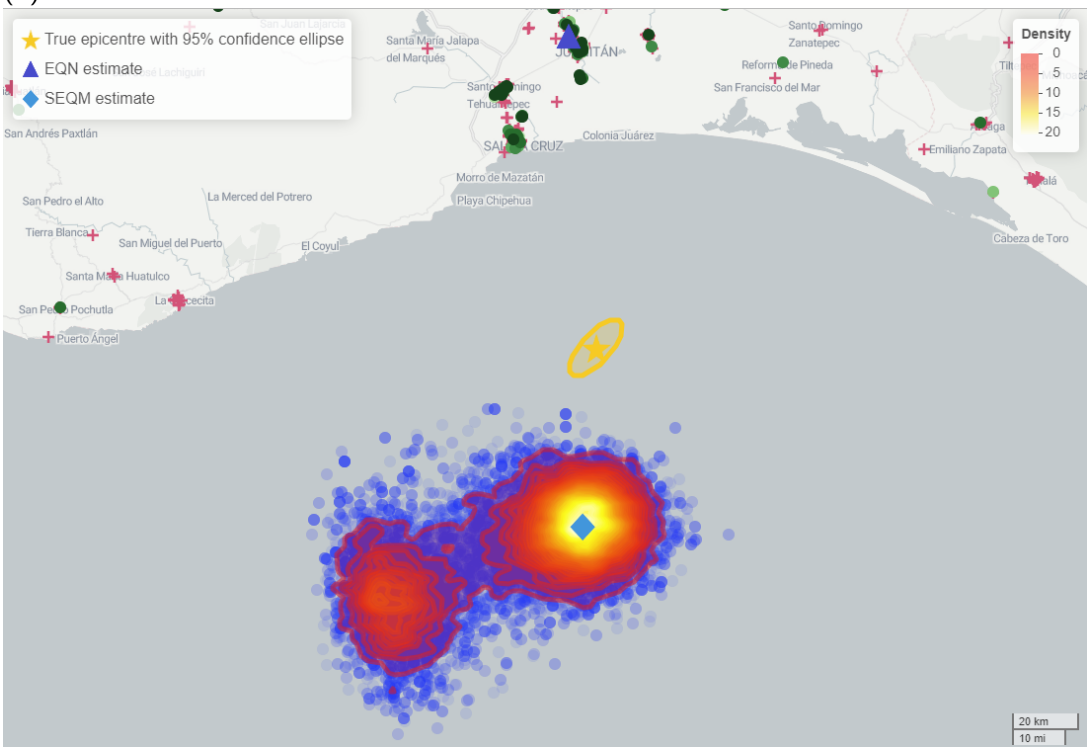


Figure 3.9: Summary map (a) and posterior distribution of the epicenter (blue dots) together with its estimated posterior density (b) for the 2019 Mexican earthquake.

Table 3.5: Comparison between EQN and SEQM earthquake parameter estimates. Errors are computed with respect to the EMSC earthquake parameters in Table 3.1.

		Pazarcik	Ridgecrest	Mexico
EQN error	epicenter [km]	35.40	187.43	92.86
	origin time [s]	-10.00	-41.00	-29.00
SEQM error	epicenter [km]	10.71	14.82	53.43
	depth [km]	7.01	-4.83	13.94
	origin time [s]	-0.15	3.45	1.77

estimation error is around 15 km for the Ridgecrest event, which is notable considering that the EQN detection occurred at a high distance. The error on the depth is around 5 km, while the estimated origin time is around 3 seconds earlier than the EMSC origin time. For the Mexican event, the error on the epicenter location is around 503 km, smaller than the 93 km error given by EQN. The error on depth is around 14 km, while the error on origin time is around 2 seconds.

The model parameters α, π, v_P and λ are more subtle to interpret as their estimation is affected by the relatively short time frame over which the data are observed. For instance, the posterior distribution on α is bi-modal for the Pazarcik event. This happens because all smartphones are triggered on the P-wave. This implies that P- and S-waves can be switched by switching velocities v_P and v_S. Only a longer time frame extended after the EQN detection (the censoring event) could have provided the opportunity to observe triggers caused by the S-wave. Similarly, λ tends to have a large HPDR because non-seismic triggers are rare events, and the time frame (which is only 120 seconds into the past from the EQN detection) is not long enough to obtain an accurate estimate of the parameter. On the other hand, the parameter π (fraction of faulty smartphones) tends to be overestimated if the EQN detection time is close to the origin time. This is because most of the non-triggering smartphones were actually censored, and they (likely) triggered soon after EQN detection. For the Ridgecrest event, π is small because the EQN detection occurred far from the epicenter many seconds after the origin time, and the seismic waves propagated far enough to observe many triggers. This effect can be seen in Table 3.1, where the ratio of triggers to active smartphones is higher for the Ridgecrest event.

Finally, the estimated τ is close to the expected value (0.67) for the Pazarcik event and much larger for the Ridgecrest and Mexican events. Again, this is probably a consequence of the fast EQN detection in the Pazarcik event. In the other two cases, assuming that v_P and v_S are constant across space is less suitable for describing seismic wave propagation at large distances from the epicenter. Consequently, τ increases to accommodate triggering times that are more variable around the expected arrival time of the seismic wave.

3.9 Discussion

Smartphone networks employed in earthquake early warning are complex stochastic objects characterized by a random geometry and by nodes (the smartphones) that exhibit non-predictable behavior during an earthquake. In this paper, we developed a modeling approach for estimating relevant earthquake parameters, uncertainty included, starting from the data provided by the smartphone network when an earthquake is detected. The approach, primarily inspired by survival analysis, allows us

to handle the uncertainty of the smartphone data and fully exploit the information content of the dataset. More precisely, earthquake parameter estimation is not only affected by the trigger spatial locations and times but also by the spatial locations of the non-triggering smartphones. This helps to constrain the epicenter location and origin time, especially when the number of triggers is relatively small and/or all the triggers are clustered within a small area. Such benefit comes from the cure rate model, which finds a novel application in this context.

A simulation study showed that the model can recover the true earthquake parameters even when the epicenter is outside the smartphone network. Notably, our modeling approach significantly improves the EQN estimates for both epicenter and origin time while also offering a measure of uncertainty. This consideration holds also when applied to real earthquake datasets. In particular, the posterior density of the epicenter exhibits spatial variability that covers the EMSC epicenter for the Ridgecrest event and is spatially close to the EMSC epicenter for the Pazarcik and Mexican earthquakes.

Future developments may include implementing a dependent data prior approach. In this work, we defined a diffuse prior on the hypocenter location so that the smartphone data strongly informs its estimate through the likelihood. However, hypocenters are more likely to be located near known faults. Earthquake catalogs may thus be exploited to define more informative priors to better constraint estimates, for instance, following the work of Argiento and Guglielmi (2014) on Bayesian principal curves to specify the prior on the hypocenters around the faults.

As a final remark, we argue that the approach outlined in this paper is not intended for real-time applications. Rather, the approach provides reliable estimates of the earthquake parameters within a few minutes after the EQN detection and the release of the preliminary EQN estimates. In areas well covered by the seismic networks of national and international seismological institutes, a delay of a few minutes may not be competitive. On the other hand, it is a relatively short delay for poorly covered areas (especially in low-income developing countries). Since the computational cost depends mainly on the number of active devices in the EQN network, optimizing the computational efficiency of the algorithm is a crucial future development for the implementation of real-time capabilities.

Conclusions

In conclusion, this thesis has demonstrated the broad applicability and adaptability of Bayesian methods across diverse challenges in public health and environmental risk assessment. By focusing on three distinct yet interconnected areas – Bayesian nonparametric clustering for environmental data, spatial disease boundary detection, and earthquake parameter estimation using smartphone accelerometer data – this work highlights the versatility of Bayesian approaches in managing uncertainty, handling complex datasets, and providing practical insights.

The first project provided a comprehensive review of Bayesian nonparametric clustering methods, specifically within the context of environmental data analysis. Environmental datasets are often high-dimensional, noisy, and sparse, which makes traditional clustering approaches ineffective. By synthesizing recent advancements in Bayesian nonparametrics, this review laid the groundwork for future innovations in spatio-temporal clustering, helping to uncover hidden patterns in environmental factors such as pollution. These methods are essential for policymakers and researchers looking to address environmental risks in a more informed, data-driven manner (Wade, 2023; Piegorsch and Bailer, 2005). This review underscores the need for flexible and adaptive models that can reveal complex structures in environmental data, providing a strong foundation for subsequent research in this critical field.

The second project tackled the challenge of detecting spatial disparities in disease rates, a key concern for public health authorities. Traditional spatial models often fail to capture the intricate relationships between multiple diseases and geographic regions, which limits the ability to inform effective interventions. By introducing a flexible Bayesian nonparametric model, this work advanced the understanding of disease boundaries and their spatial dependencies. The inclusion of graphical models allowed for the explicit modeling of multivariate dependencies, which is particularly useful in public health contexts where diseases may share common risk factors. This application to the SEER cancer database uncovered associations between lung, esophageal, and larynx cancers, suggesting that targeted interventions could yield broader public health benefits (Rao, 2023; Lawson, 2018; Gao et al., 2023). The framework developed here also advances the methodological landscape by offering more rigorous uncertainty quantification in the detection of spatial disparities, opening up new directions for research on health disparities and policy planning.

The third project applied Bayesian survival modeling to the real-time estimation of earthquake parameters using smartphone accelerometer data. This innovative approach offers a novel way to harness crowd-sourced smartphone data for seismic monitoring, providing real-time estimates of earthquake parameters such as the epicenter and origin time. Traditional seismic networks are often limited by geographic coverage, particularly in low-income or rural areas. By integrating smartphone data, the model developed in this project expands the potential for rapid and accurate earthquake detection,

especially in regions that are not well-covered by conventional seismic networks (Finazzi, 2016; Kong et al., 2016; Argiento and Guglielmi, 2014). The Bayesian survival model's ability to incorporate non-triggering smartphones into the analysis enhances the precision of earthquake parameter estimates, even when the epicenter lies outside the smartphone network. This work not only represents a novel application of survival analysis but also provides a path forward for improving early warning systems in earthquake-prone areas.

Together, these projects showcase the critical role of Bayesian methods in addressing the complexities inherent in public health and environmental challenges. The review of Bayesian nonparametric clustering methods highlights their capacity to manage high-dimensional environmental data, while the spatial disease mapping project advances the understanding of geographic health disparities, and the earthquake modeling project demonstrates the power of Bayesian survival models in real-time disaster monitoring. These contributions collectively push the boundaries of Bayesian modeling, demonstrating its effectiveness in managing uncertainty, integrating diverse data sources, and driving actionable insights.

While each project encountered challenges and limitations, they also opened new avenues for further research. For example, the review of Bayesian clustering methods provides a foundation for future work in high-dimensional environmental data analysis, while the spatial disease mapping model could be extended to explore alternative boundary detection techniques such as those proposed by Wu and Banerjee (2024). The earthquake parameter estimation model, on the other hand, could benefit from incorporating more informative priors based on known fault lines, as discussed by Argiento and Guglielmi (2014). These potential future directions promise to advance both methodological development and practical applications in the field of public health and environmental risk management.

In conclusion, this thesis illustrates that Bayesian methods are indispensable for navigating the complexities of modern public health and environmental risk challenges. From improving the detection of disease boundaries to developing more accurate real-time earthquake monitoring systems, Bayesian models provide powerful tools for researchers and policymakers alike. By effectively managing uncertainty and integrating complex datasets, these approaches offer robust frameworks for mitigating risks and informing timely interventions. The innovations and methodologies presented here ensure that data-driven decision-making remains at the forefront of efforts to address evolving public health and environmental crises, providing a foundation for future advancements in these critical areas.

Bibliography

Agrawal, K., Markert, R. J., and Agrawal, S. (2018). Risk factors for adenocarcinoma and squamous cell carcinoma of the esophagus and lung. *Hypertension*, 61(46):0–09. 25

Ahkola, H., Äystö, L., Sikanen, T., Riikonen, S., Pihlaja, T., and Kauppi, S. (2024). Current uncertainties and challenges of publicly available pharmaceutical environmental risk assessment data. *European Journal of Pharmaceutical Sciences*, page 106769. 2

Ahmadi, J. and Nagaraja, H. (2020). Conditional properties of a random sample given an order statistic. *Statistical Papers*, 61(5):1971–1996. 61

Aiello, L. and Banerjee, S. (2023). Detecting spatial health disparities using disease maps. *arXiv preprint arXiv:2309.02086*. 3

Aiello, Argiento, Finazzi and Paci (2023). Survival modelling of smartphone trigger data for earthquake parameter estimation in early warning. with applications to 2023 turkish-syrian and 2019 ridgecrest events. *arXiv preprint arXiv:2303.00806*. 3

Aiello, Legramanti and Paci (2024). A spatial product partition model for pm10 data. In *Book of the Short Papers SIS 2024*. 3

Akhtar, J., Bhargava, R., Shameem, M., Singh, S. K., Baneen, U., Khan, N. A., Hassan, J., and Sharma, P. (2010). Second primary lung cancer with glottic laryngeal cancer as index tumor–a case report. *Case reports in oncology*, 3(1):35–39. 25

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20(3):407–415. 64

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and computing*, 18(4):343–373. 64

Argiento, R. and De Iorio, M. (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics*, 50(5):2641–2663. 8, 9, 10

Argiento, R., Filippi-Mazzola, E., and Paci, L. (2024). Model-based clustering of categorical data based on the Hamming distance. *Journal of the American Statistical Association*, doi:10.1080/01621459.2024.2402568:1–20. 8

Argiento, R. and Guglielmi, A. (2014). Bayesian principal curve clustering by species-sampling mixture models. In *Proceedings of 47th SIS Scientific Meeting of the Italian Statistica Society*, pages 1–6. 77, 80

Arima, S., Cretarola, L., Jona Lasinio, G., and Pollice, A. (2012). Bayesian univariate space-time hierarchical model for mapping pollutant concentrations in the municipal area of Taranto. *Statistical Methods & Applications*, 21:75–91. 6

Arima, S., Datta, G. S., and Liseo, B. (2015). Bayesian estimators for small area models when auxiliary information is measured with error. *Scandinavian Journal of Statistics*, 42(2):518–529. 6

Ascari, R., Di Brisco, A. M., Migliorati, S., and Ongaro, A. (2024). A multivariate mixture regression model for constrained responses. *Bayesian Analysis*, 19(2):377–405. 9

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11(5):815–828. 64, 65

Banerjee, S. (2016). Multivariate spatial models. In Lawson, A. B., Banerjee, S., Haining, R. P., and Ugarte, M. D., editors, *Handbook of Spatial Epidemiology*, pages 375–397. Taylor & Francis/CRC Press, Boca Raton, FL. 33

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling And Analysis For Spatial Data*. CRC press. 2, 3, 5, 6, 26, 34

Bayarri, M. and Berger, J. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80. 2

Beraha, M., Argiento, R., Møller, J., and Guglielmi, A. (2022). MCMC computations for Bayesian mixture models using repulsive point processes. *Journal of Computational and Graphical Statistics*, 31(2):422–435. 9

Bernardo, J. and Girón, F. (1988). A Bayesian analysis of simple mixture problems. *Bayesian Statistics*, 3(3):67–78. 9

Berrocal, V. (2016). Identifying trends in the spatial errors of a regional climate model via clustering. *Environmetrics*, 27(2):90–102. 6, 12

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika*, 65(1):31–38. 11

Bossu, R., Finazzi, F., Steed, R., Fallou, L., and Bondár, I. (2022). "Shaking in 5 seconds!"—performance and user appreciation assessment of the earthquake network smartphone-based public earthquake early warning system. *Seismological Society of America*, 93(1):137–148. 55

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press. 7, 8

Brilleman, S. L., Elci, E. M., Novik, J. B., and Wolfe, R. (2020). Bayesian survival analysis using the rstanarm R package. *arXiv:2002.09633*. 132

Brookmeyer, R. and Stroup, D. F. (2004). *Monitoring the health of populations: statistical principles and methods for public health surveillance*. Oxford University Press. 1

Bucci, A., Ippoliti, L., Valentini, P., and Fontanella, S. (2022). Clustering spatio-temporal series of confirmed COVID-19 deaths in europe. *Spatial Statistics*, 49:100543. 12

Cameletti, M., Ignaccolo, R., and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics*, 22(8):985–996. 6, 15

Carlin, B. P. and Banerjee, S. (2003). Hierarchical multivariate car models for spatio-temporally correlated survival data. *Bayesian Statistics*, 7(7):45–63. 32

Cheam, A., Marbac, M., and McNicholas, P. (2017). Model-based clustering for spatiotemporal data on air quality monitoring. *Environmetrics*, 28(3):e2437. 6

Chiolero, A., Tancredi, S., and Ioannidis, J. P. (2023). Slow data public health. *European journal of epidemiology*, 38(12):1219–1225. 2

Clayton, R., Heaton, T., Chandy, M., Krause, A., Kohler, M., Bunn, J., Guy, R., Olson, M., Faulkner, M., Cheng, M., et al. (2011). Community seismic network. *Annals of Geophysics*, 54(6):738–747. 51

Cochran, E. S., Lawrence, J. F., Christensen, C., and Jakka, R. S. (2009). The quake-catcher network: Citizen science expanding seismic horizons. *Seismological Research Letters*, 80(1):26–30. 51

Connor, R. J. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206. 9

Corpas-Burgos, F. and Martinez-Beneito, M. A. (2020). On the use of adaptive spatial weight matrices from disease mapping multivariate analyses. *Stochastic Environmental Research and Risk Assessment*, 34:531–544. 23

Covello, V. T. and Merkhoher, M. W. (1993). *Risk assessment methods: approaches for assessing health and environmental risks*. Springer Science & Business Media. 1

Cox, D. and Wermuth, N. (1996). *Multivariate Dependencies*. Chapman & Hall/CRC, Boca Raton, FL. 33

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*, volume 21. CRC press. 58

Cox, D. R. and Wermuth, N. (1993). Linear Dependencies Represented by Chain Graphs. *Statistical Science*, 8(3):204 – 218. 33

Cressie, N. and Wikle, C. K. (2015). *Statistics For Spatio-Temporal Data*. John Wiley & Sons. 2, 5

Cutler, S. J. and Axtell, L. M. (1963). Partitioning of a patient population with respect to different mortality risks. *Journal of the American Statistical Association*, 58(303):701–712. 52

Dahl, D. B., Johnson, D. J., and Müller, P. (2022). Search algorithms and loss functions for Bayesian clustering. *Journal of Computational and Graphical Statistics*, 31(4):1189–1201. 11, 18

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American statistical Association*, 93(441):294–302. 6

Datta, A., Banerjee, S., Hodges, J. S., and Gao, L. (2019). Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models. *Bayesian analysis*, 14(4):1221. 24, 30

Davison, A. C. (2003). *Statistical models*, volume 11. Cambridge university press. 60

De Angelis, R., Capocaccia, R., Hakulinen, T., Soderman, B., and Verdecchia, A. (1999). Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine*, 18(4):441–454. 52, 58

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229. 9, 12

De Blasi, P., Lijoi, A., and Prünster, I. (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Statistica Sinica*, 23:1299–1321. 12

Diggle, P. J. and Giorgi, E. (2019). *Model-based geostatistics for global public health: methods and applications*. Chapman and Hall/CRC. 1

Doll, R., Peto, R., Boreham, J., and Sutherland, I. (2005). Mortality from cancer in relation to smoking: 50 years observations on british doctors. *British journal of cancer*, 92(3):426–429. 27, 44

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika*, 94(4):809–825. 11, 12

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063. 36

Escobar, M. D. (1994). Estimating Normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277. 10

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588. 10

European Commission (2008). Directive 2008/50/ec of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. https://eur-lex.europa.eu/eli/dir/2008/50/2015-09-18. 15

European Environment Agency (2022). Air quality in europe 2022. https://www.eea.europa.eu/publications/air-quality-in-europe-2022. 1

Fernández, C. and Green, P. J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):805–826. 11

Finazzi, F. (2016). The earthquake network project: Toward a crowdsourced smartphone-based earthquake early warning system. *Bulletin of the Seismological Society of America*, 106(3):1088–1099. 3, 51, 80

Finazzi, F. (2020). The earthquake network project: A platform for earthquake early warning, rapid impact assessment, and search and rescue. *Frontiers in Earth Science*, 8:243. 51

Finazzi, F., Bondár, I., Bossu, R., and Steed, R. (2022). A probabilistic framework for modeling the detection capability of smartphone networks in earthquake early warning. *Seismological Research Letters*, 93(6):3291–3307. 52

Finazzi, F., Bossu, R., and Cotton, F. (2024). Smartphones enabled up to 58 s strong-shaking warning in the M7.8 Türkiye earthquake. *Scientific Reports*, 14(1):4878. 51

Finazzi, F. and Fassò, A. (2017). A statistical approach to crowdsourced smartphone-based earthquake early warning systems. *Stochastic environmental research and risk assessment*, 31(7):1649–1658. 52, 54, 67

Fitzpatrick, M. C., Preisser, E. L., Porter, A., Elkinton, J., Waller, L. A., Carlin, B. P., and Ellison, A. M. (2010). Ecological boundary detection using bayesian areal wombling. *Ecology*, 91(12):3448–3455. 23

Flegal, J. M., Hughes, J., Vats, D., Dai, N., Gupta, K., and Maji, U. (2021). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA, and Kanpur, India. R package version 1.5-0. 48

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631. 7

Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2013). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1149–1157. 8

Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P. (2019). *Handbook of mixture analysis*. CRC press. 8

Frühwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26(1):78–89. 11

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, 16(4):1279–1307. 10

Frühwirth-Schnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-Normal and skew-t distributions. *Biostatistics*, 11(2):317–336. 8

Fúquene, J., Steel, M., and Rossell, D. (2019). On choosing mixture components via non-local priors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(5):809–837. 9

Gamel, J. W. and Vogel, R. L. (2001). Non-parametric comparison of relative versus cause-specific survival in surveillance, epidemiology and end results (seer) programme breast cancer patients. *Statistical Methods in Medical Research*, 10(5):339–352. 59

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC press. 101

Gao, L., Banerjee, S., and Ritz, B. (2023). Spatial difference boundary detection for multiple outcomes using bayesian disease mapping. *Biostatistics*, 24(4):922–944. 23, 24, 28, 30, 42, 49, 50, 79

Gao, L., Datta, A., and Banerjee, S. (2022). Hierarchical multivariate directed acyclic graph autoregressive models for spatial diseases mapping. *Statistics in Medicine*, 41(16):3057–3075. 33, 34

Gasparini, P., Manfredi, G., Zschau, J., et al. (2007). *Earthquake early warning systems*. Springer. 51

Gelfand, A., Guindani, M., Petrone, S., et al. (2007). Bayesian nonparametric modelling for spatial data using Dirichlet processes. In *Bayesian Statistics 8*, pages 175–200. Oxford University Press. 11

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035. 11, 12

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15. 32

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515 – 534. 64

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press. 1

Ghosal, S. and van der Vaart, A. W. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press. 2

Gianella, M., Beraha, M., and Guglielmi, A. (2023). Bayesian nonparametric boundary detection for income areal data. *arXiv preprint arXiv:2312.13992*. 23

Given, D. D., Cochran, E. S., Heaton, T., Hauksson, E., Allen, R., Hellweg, P., Vidale, J., and Bodin, P. (2014). *Technical implementation plan for the ShakeAlert production system: An earthquake early warning system for the west coast of the United States*. US Department of the Interior, US Geological Survey Reston, VA. 3, 51

Gnedin, A. (2010). A species sampling model with finitely many types. *Electronic Communications in Probability [electronic only]*, 15:79–88. 12

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical sciences*, 138:5674–5685. 12

Gong, L. and Flegal, J. M. (2016). A practical sequential stopping rule for high-dimensional markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 25(3):684–700. 48

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231. 8

Grazian, C. (2023). A review on Bayesian model-based clustering. *arXiv preprint arXiv:2303.17182*. 8

Grazian, C., Villa, C., and Liseo, B. (2020). On a loss-based prior for the number of components in mixture models. *Statistics & Probability Letters*, 158:108656. 9

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732. 10

Grün, B. (2019). Model-based clustering. In *Handbook of mixture analysis*, pages 157–192. Chapman and Hall/CRC. 8

Hamm, N., Finley, A., Schaap, M., and Stein, A. (2015). A spatially varying coefficient model for mapping PM10 air quality at the European scale. *Atmospheric Environment*, 102:393 – 405. 15

Hanson, T., Banerjee, S., Li, P., and McBean, A. (2015). Spatial boundary detection for areal counts. In *Nonparametric Bayesian Inference in Biostatistics*, pages 377–399. Springer. 23

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 38(3):205–228. 6

Hartigan, J. A. (1990). Partition Models. *Communications in statistics-Theory and methods*, 19(8):2745–2756. 12

Hefley, T. J., Hooten, M. B., Hanks, E. M., Russell, R. E., and Walsh, D. P. (2017). Dynamic spatio-temporal models for spatial data. *Spatial statistics*, 20:206–220. 6

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press. 2

Hossain, M. M., Lawson, A., Cai, B., Choi, J., Liu, J., and Kirby, R. S. (2014). Space-time areal mixture model: relabeling algorithm and model selection issues. *Environmetrics*, 25(2):84–96. 11

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2005). *Bayesian Survival Analysis*. John Wiley & Sons, Ltd. 2, 3, 57, 60

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American statistical Association*, 96(453):161–173. 10

Ishwaran, H. and James, L. F. (2003). Some further developments for stick-breaking priors: Finite and infinite clustering and classification. *Sankhya*, 65(3):577–592. 10

Jacquez, G. M. and Greiling, D. A. (2003a). Geographic boundaries in breast, lung and colorectal cancers in relation to exposure to air toxics in long island, new york. *International Journal of Health Geographics*, 2(1):1–22. 23

Jacquez, G. M. and Greiling, D. A. (2003b). Local clustering in breast, lung and colorectal cancer in long island, new york. *International Journal of Health Geographics*, 2(1):1–12. 23

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67. 10

Jin, X., Banerjee, S., and Carlin, B. P. (2007). Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):817–838. 33

Jin, X., Carlin, B. P., and Banerjee, S. (2005). Generalized hierarchical multivariate car models for areal data. *Biometrics*, 61(4):950–961. 25, 32

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and computing*, 21:93–105. 10, 29

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, pages 457–481. 132

Karlis, D. and Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review/Revue Internationale de Statistique*, pages 35–58. 8

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media. 2, 57

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (2016). *Handbook of Survival Analysis*. CRC Press. 57

Knorr-Held, L. and Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in medicine*, 17(18):2045–2060. 6

Koch, T. (2005). *Cartographies of disease: maps, mapping, and medicine*. Esri Press Redlands, CA. 23

Kong, Q., Allen, R. M., and Schreier, L. (2016). Myshake: Initial observations from a global smartphone seismic network. *Geophysical Research Letters*, 43(18):9588–9594. 3, 51, 80

Kottas, A., Duan, J. A., and Gelfand, A. E. (2008). Modeling disease incidence data with spatial and spatio temporal Dirichlet process mixtures. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(1):29–42. 6

Krnjajić, M., Kottas, A., and Draper, D. (2008). Parametric and nonparametric Bayesian model specification: A case study involving models for count data. *Computational Statistics & Data Analysis*, 52(4):2110–2128. 8

Kurishima, K., Miyazaki, K., Watanabe, H., Shiozawa, T., Ishikawa, H., Satoh, H., and Hizawa, N. (2018). Lung cancer patients with synchronous colon cancer. *Molecular and clinical oncology*, 8(1):137–140. 25

Lambert, P. C., Dickman, P. W., Weston, C. L., and Thompson, J. R. (2010). Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(1):35–55. 52, 58, 60

Lambert, P. C., Thompson, J. R., Weston, C. L., and Dickman, P. W. (2007). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8(3):576–594. 52

Laurini, M. P. (2019). A spatio-temporal approach to estimate patterns of climate change. *Environmetrics*, 30(1):e2542. 6

Lauritzen, S. L. (1996). *Graphical Models*, volume 17. Clarendon Press. 3

Lawson, A. B. (2013). *Statistical methods in spatial epidemiology.* John Wiley & Sons. 23

Lawson, A. B. (2018). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC. 3, 79

Lawson, Andrew, B., Banerjee, S., Haining, R., and Ugarte, Maria, D. (2016). *Handbook of Spatial Epidemiology*. CRC press, Boca Raton, FL. 23

Lázaro, E., Armero, C., and Gómez-Rubio, V. (2020). Approximate Bayesian inference for mixture cure models. *TEST*, 29(3):750–767. 52

Lee, D. and Mitchell, R. (2012). Boundary detection in disease mapping studies. *Biostatistics*, 13(3):415–426. 23, 24, 30, 31

Lee, J., James, L. F., and Choi, S. (2016). Finite-dimensional BFRY priors and variational Bayesian inference for power law models. *Advances in Neural Information Processing Systems*, 29. 9

Lee, J., Kamenetsky, M. E., Gangnon, R. E., and Zhu, J. (2021). Clustered spatio-temporal varying coefficient regression model. *Statistics in medicine*, 40(2):465–480. 6

Lee, S. and McLachlan, G. J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24:181–202. 8

Li, B., Zhang, X., and Smerdon, J. E. (2016). Comparison between spatio-temporal random processes and application to climate model data. *Environmetrics*, 27(5):267–279. 6

Li, P., Banerjee, S., Carlin, B. P., and McBean, A. M. (2012). Bayesian areal wombling using false discovery rates. *Statistics and its Interface*, 5(2):149–158. 23, 24

Li, P., Banerjee, S., Hanson, T. A., and McBean, A. M. (2015). Bayesian models for detecting difference boundaries in areal data. *Statistica Sinica*, 25(1):385. 23, 24, 27, 37

Li, P., Banerjee, S., and McBean, A. M. (2011). Mining boundary effects in areally referenced spatial data using the bayesian information criterion. *Geoinformatica*, 15(3):435–454. 23, 31

Li, Y., Tiwari, R. C., and Guha, S. (2007). Mixture cure survival models with dependent censoring. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69(3):285–306. 60

Liang, Y. (2019). Graph-based multivariate conditional autoregressive models. *Statistical Theory and Related Fields*, 3(2):158–169. 23

Lijoi, A., Mena, R. H., and Prünster, I. (2007a). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786. 12

Lijoi, A., Mena, R. H., and Prünster, I. (2007b). Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):715–740. 12

Lijoi, A., Prünster, I., and Rigon, T. (2020). The Pitman–Yor multinomial process for mixture modelling. *Biometrika*, 107(4):891–906. 9

Lijoi, A., Prunster, I., and Walker, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Annals of Applied Probability*, 18(4):1519–1547. 12

Lindley, D. V. (2013). *Understanding uncertainty*. John Wiley & Sons. 1

Liu, J., Wade, S., and Bochkina, N. (2024). Shared differential clustering across single-cell rna sequencing datasets with the hierarchical dirichlet process. *Econometrics and Statistics*. 8

Lu, H. and Carlin, B. P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, 37(3):265–285. 23

Lu, H., Reilly, C. S., Banerjee, S., and Carlin, B. P. (2007). Bayesian areal wombling via adjacency modeling. *Environmental and ecological statistics*, 14:433–452. 23, 31

Ma, H. and Carlin, B. P. (2007). Bayesian multivariate areal wombling for multiple disease boundary analysis. *Bayesian Analysis*, 2(2):281–302. 23

Ma, H., Carlin, B. P., and Banerjee, S. (2010). Hierarchical and joint site-edge methods for medicare hospice service region boundary analysis. *Biometrics*, 66(2):355–364. 23, 31

MacEachern, S. N. (1994). Estimating Normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741. 10

MacEachern, S. N. (1998). Computational methods for mixture of Dirichlet process models. In *Practical nonparametric and semiparametric Bayesian statistics*, pages 23–43. Springer. 10

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238. 10

MacNab, Y. C. (2018). Some recent work on multivariate Gaussian Markov random fields (with discussion). *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 27(3):497–541. 33

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*. 7

Mardia, K. (1988). Multi-dimensional multivariate gaussian markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2):265–284. 32, 34

Massoda Tchoussi, F. Y. and Finazzi, F. (2023). A statistical methodology for classifying earthquake detections and for earthquake parameter estimation in smartphone-based earthquake early warning systems. *Frontiers in Applied Mathematics and Statistics*, 9. 52

Mastrantonio, G., Grazian, C., Mancinelli, S., and Bibbona, E. (2019). New formulation of the logistic-gaussian process to analyze trajectory tracking data. *The Annals of Applied Statistics*, 13(4):2483–2508. 12

Mastrantonio, G., Jona Lasinio, G., Pollice, A., Teodonio, L., and Capotorti, G. (2022). A dirichlet process model for change-point detection with multivariate bioclimatic data. *Environmetrics*, 33(1):e2699. 12

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6(1):355–378. 8

Megan Othus, Y. L. and Tiwari, R. C. (2009). A class of semiparametric mixture cure survival models with dependent censoring. *Journal of the American Statistical Association*, 104(487):1241–1250. 60

Meilă, M. (2007). Comparing clusterings–an information based distance. *Journal of multivariate analysis*, 98(5):873–895. 11

Miasojedow, B., Moulines, E., and Vihola, M. (2013). An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics*, 22(3):649–664. 64, 65, 101

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356. 9, 10, 12

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99(468):990–1001. 24, 37

Müller, P., Quintana, F., and Rosner, G. L. (2011). A Product Partition Model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278. 13, 14

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer. 29

Musau, V. M., Gaetan, C., and Girardi, P. (2022). Clustering of bivariate satellite time series: A quantile approach. *Environmetrics*, 33(7):e2755. 12

National Cancer Institute (2019). Seer*stat software. 25

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265. 19

Neelon, B., Gelfand, A. E., and Miranda, M. L. (2014). A multivariate spatial mixture model for areal data: examining regional differences in standardized test scores. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(5):737–761. 11

Nguyen, X. and Gelfand, A. E. (2011). The Dirichlet labeling process for clustering functional data. *Statistica Sinica*, pages 1249–1289. 11, 12

Nieto-Barajas, L. E. and Contreras-Cristán, A. (2014). A Bayesian nonparametric approach for time series clustering. *Bayesian Analysis*, 9(1):147–170. 11, 12, 17

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *Annals of Statistics*, 32:2044–2073. 8, 9

Olkin, I. and Sampson, A. R. (1972). Jacobians of matrix transformations and induced functional equations. *Linear Algebra and its applications*, 5(3):257–276. 36

Ongaro, A., Migliorati, S., and Ascari, R. (2020). A new mixture model on the simplex. *Statistics and Computing*, 30:749–770. 9

O'Hagan, A., Murphy, T. B., Gormley, I. C., McNicholas, P. D., and Karlis, D. (2016). Clustering with the multivariate Normal inverse Gaussian distribution. *Computational Statistics & Data Analysis*, 93:18–30. 8

Paci, L. and Finazzi, F. (2018). Dynamic model-based clustering for spatio-temporal data. *Statistics and Computing*, 28:359–374. 11

Paci, L., Gelfand, A. E., and Holland, D. M. (2013). Spatio-temporal modeling for real-time ozone forecasting. *Spatial Statistics*, 4:79–93. 6

Page, G. L. and Quintana, F. A. (2016). Spatial Product Partition Models. *Bayesian Analysis*, 11(1):265–298. 13, 14, 17

Page, G. L., Quintana, F. A., and Dahl, D. B. (2022). Dependent modeling of temporal sequences of random partitions. *Journal of Computational and Graphical Statistics*, 31(2):614–627. 12

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1):169–186. 10

Paton, F. and McNicholas, P. D. (2020). Detecting British Columbia coastal rainfall patterns by clustering Gaussian processes. *Environmetrics*, 31(8):e2631. 6

Pavani, J. and Quintana, F. A. (2024). A bayesian multivariate model with temporal dependence on random partition of areal data. *arXiv preprint arXiv:2401.08303*. 23

Peluso, S., Mira, A., Rue, H., Tierney, N. J., Benvenuti, C., Cianella, R., Caputo, M. L., and Auricchio, A. (2020). A Bayesian spatiotemporal statistical analysis of out-of-hospital cardiac arrests. *Biometrical Journal*, 62(4):1105–1119. 6

Peng, Y. and Taylor, J. M. (2014). Cure models. *Handbook of survival analysis*, 34:113–134. 52, 58, 59, 125

Perman, M., Pitman, J., and Yor, M. (1992). Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39. 9

Petralia, F., Rao, V., and Dunson, D. (2012). Repulsive mixtures. *Advances in neural information processing systems*, 25. 9

Petrone, S., Guindani, M., and Gelfand, A. E. (2009). Hybrid dirichlet mixture models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):755–782. 28

Piegorsch, W. W. and Bailer, A. J. (2005). *Analyzing environmental data*. John Wiley & Sons. 1, 79

Pietrogrande, M. C., Demaria, G., Colombi, C., Cuccia, E., and Dal Santo, U. (2022). Seasonal and spatial variations of PM10 and PM2.5 oxidative potential in five urban and rural sites across Lombardia Region, Italy. *International Journal of Environmental Research and Public Health*, 19(13):7778. 15

Pitman, J. (1996). Some developments of the Blackwell-Macqueen URN scheme. *Lecture Notes-Monograph Series*, 30:245–267. 9

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900. 9

Quintana, F. A. (2006). A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, 136(8):2407–2429. 7

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and Product Partition Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):557–574. 12

Ranalli, M. and Maruotti, A. (2020). Model-based clustering for noisy longitudinal circular data, with application to animal movement. *Environmetrics*, 31(2):e2572. 6

Rao, J. S. (2023). *Statistical Methods in Health Disparity Research*. Chapman & Hall/CRC, Boca Raton, FL. 3, 23, 49, 79

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press. 6

Reich, B. J. and Fuentes, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *The Annals of Applied Statistics*, 1(1):249–264. 11

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(4):731–792. 10

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367. 64

Rodrıguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1):145–178. 11

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2010). Latent stick-breaking processes. *Journal of the American Statistical Association*, 105(490):647–659. 28

Rousseau, J., Grazian, C., and Lee, J. E. (2019). Bayesian mixture models: Theory and methods. In *Handbook of mixture analysis*, pages 53–72. Chapman and Hall/CRC. 9

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in over-fitted mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(5):689–710. 8

Sahu, S. K., Gelfand, A. E., and M, D. (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics*, 11:61–86. 15

Sarang, P. (2023). *Centroid-Based Clustering*, pages 171–183. Springer International Publishing, Cham. 7

Shi, W.-X. and Chen, S.-Q. (2004). Frequencies of poor metabolizers of cytochrome p450 2c19 in esophagus cancer, stomach cancer, lung cancer and bladder cancer in chinese population. *World journal of gastroenterology: WJG*, 10(13):1961. 25

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components–an alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74. 9

Stroud, J. R., Müller, P., and Sansó, B. (2001). Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):673–689. 6

Sun, W., Reich, B. J., Tony Cai, T., Guindani, M., and Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1):59–83. 37

Suslick, K. S. (2001). Encyclopedia of physical science and technology. *Sonoluminescence and sonochemistry, 3rd edn. Elsevier Science Ltd, Massachusetts*, pages 1–20. 62

Suter II, G. W. (2016). *Ecological risk assessment*. CRC press. 1

Torabi, M. (2014). Spatiotemporal modeling of odds of disease. *Environmetrics*, 25(5):341–350. 6

United Nations Office for Disaster Risk Reduction (2024). Global assessment report on disaster risk reduction 2024. https://www.undrr.org/gar. 1

U.S. Environmental Protection Agency (2024). Inventory of U.S. greenhouse gas emissions and sinks: 1990-2022. https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks-1990-2022. 1

Vanhatalo, J., Foster, S. D., and Hosack, G. R. (2021). Spatiotemporal clustering using Gaussian processes embedded in a mixture model. *Environmetrics*, 32(7):e2681. 6

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for markov chain monte carlo. *Biometrika*, 106(2):321–337. 48

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2024). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.8.0. 41

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27:1413–1432. 41

Viroli, C. (2011). Model based clustering for three-way data structures. *Bayesian Analysis*, 6(4):573–602. 11

Wade, S. (2023). Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A*, 381(2247):20220149. 3, 7, 8, 79

Wade, S. and Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626. 11

Waller, L. and Carlin, B. (2010). Disease mapping. In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M., editors, *Handbook Of Spatial Statistics*, page 217–243. CRC Press, Boca Raton, FL. 23

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical association*, 92(438):607–617. 6

Waller, L. A. and Gotway, C. A. (2004). *Applied spatial statistics for public health data*, volume 368. John Wiley & Sons. 1, 23

Wang, F., Duan, C., Li, Y., Huang, H., and Shia, B.-C. (2024). Spatiotemporal varying coefficient model for respiratory disease mapping in Taiwan. *Biostatistics*, 25(1):40–56. 6

Watanabe, S. and Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12). 41

West, M. and Harrison, J. (2006). *Bayesian forecasting and dynamic models*. Springer Science & Business Media. 6

Womble, W. H. (1951). Differential systematics. *Science*, 114(2961):315–322. 23

Woodard, D. B., Schmidler, S. C., and Huber, M. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 19(2):617–640. 64

World Health Assembly (2015). Health and the environment: addressing the health impact of air pollution. https://iris.who.int/handle/10665/253237. 1

Wu, K. L. and Banerjee, S. (2024). Assessing spatial disparities: A bayesian linear regression approach. arXiv:2407.19171. 50, 80

Wu, Q. and Luo, X. (2022). Nonparametric Bayesian two-level clustering for subject-level single-cell expression data. *Statistica Sinica*, 32(4):1835–1856. 8

Xie, F. and Xu, Y. (2020). Bayesian repulsive Gaussian mixture model. *Journal of the American Statistical Association*, 115(529):187–203. 9

Xu, Y., Müller, P., and Telesca, D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, 72(3):955–964. 9

Yu, B. and Tiwari, R. C. (2012). A Bayesian approach to mixture cure models with spatial frailties for population-based cancer relative survival data. *The Canadian Journal of Statistics*, 40(1):40–54. 52

Zhang, Y., Hodges, J. S., and Banerjee, S. (2009). Smoothed anova with spatial effects as a competitor to mcar in multivariate spatial smoothing. *The Annals of Applied Statistics*, 3(4):1805. 33

Appendix A

Chapter 1 supplementary materials

A.1 Similarity function

Here we provide the computations through which we obtained the closed form of the similarity function $g_3(\cdot)$:

$$
\begin{aligned}
&\prod_{i \in S_k} q\left(\mathbf{s}_i \mid \boldsymbol{\xi}_k\right) q\left(\boldsymbol{\xi}_k\right) = \prod_{i \in S_k} \mathcal{N}_2\left(\mathbf{s}_i \mid \mathbf{m}_k, V_k\right) NIW\left(\mathbf{m}_k, V_k \mid \boldsymbol{\mu}_0, k_0, \nu_0, \Lambda_0\right) \\
&= \prod_{i \in S_k} (2\pi)^{-\frac{D}{2}} \left|V_k\right|^{-1/2} \exp\left(-\frac{1}{2}\left(\mathbf{s}_i - \mathbf{m}_k\right)^\top V_k^{-1}\left(\mathbf{s}_i - \mathbf{m}_k\right)\right) \\
&\quad \times \frac{k_0^{\frac{D}{2}} \left|\Lambda_0\right|^{\frac{\nu_0}{2}} \left|V_k\right|^{-\frac{\nu_0+D+2}{2}}}{(2\pi)^{\frac{D}{2}} 2^{\frac{\nu_0 D}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right)} \exp\left(-\frac{k_0}{2}\left(\mathbf{m}_k - \boldsymbol{\mu}_0\right)^\top V_k^{-1}\left(\mathbf{m}_k - \boldsymbol{\mu}_0\right) - \frac{1}{2}\operatorname{tr}\left(\Lambda_0 V_k^{-1}\right)\right) \\
&= \frac{k_0^{\frac{D}{2}} \left|\Lambda_0\right|^{\frac{\nu_0}{2}} \left|V_k\right|^{-\frac{(\nu_0+|S_k|)+D+2}{2}}}{(2\pi)^{(|S_k|+1)\frac{D}{2}} 2^{\frac{\nu_0 D}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right)} \exp\Bigg(- \frac{1}{2}\sum_{i \in S_k}\left(\mathbf{s}_i - \mathbf{m}_k\right)^\top V_k^{-1}\left(\mathbf{s}_i - \mathbf{m}_k\right) \\
&\qquad\qquad - \frac{k_0}{2}\left(\mathbf{m}_k - \boldsymbol{\mu}_0\right)^\top V_k^{-1}\left(\mathbf{m}_k - \boldsymbol{\mu}_0\right) - \frac{1}{2}\operatorname{tr}\left(\Lambda_0 V_k^{-1}\right)\Bigg) \\
&= \frac{k_0^{\frac{D}{2}} \left|\Lambda_0\right|^{\frac{\nu_0}{2}} \left|V_k\right|^{-\frac{(\nu_0+|S_k|)+D+2}{2}}}{(2\pi)^{(|S_k|+1)\frac{D}{2}} 2^{\frac{\nu_0 D}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right)} \exp\Bigg(- \frac{|S_k|}{2}\left(\mathbf{m}_k - \bar{\mathbf{s}}_k\right)^\top V_k^{-1}\left(\mathbf{m}_k - \bar{\mathbf{s}}_k\right) - \frac{k_0}{2}\left(\mathbf{m}_k - \boldsymbol{\mu}_0\right)^\top V_k^{-1}\left(\mathbf{m}_k - \boldsymbol{\mu}_0\right) \\
&\qquad\qquad - \frac{1}{2}\operatorname{tr}\left(\Lambda_0 V_k^{-1}\right) - \frac{1}{2}\operatorname{tr}\left(\Delta_k V_k^{-1}\right)\Bigg) \\
&= \frac{k_0^{\frac{D}{2}} \left|\Lambda_0\right|^{\frac{\nu_0}{2}} \left|V_k\right|^{-\frac{(\nu_0+|S_k|)+D+2}{2}}}{(2\pi)^{(|S_k|+1)\frac{D}{2}} 2^{\frac{\nu_0 D}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right)} \exp\Bigg(- \frac{1}{2}\left(\mathbf{m}_k - \frac{|S_k|\,\bar{\mathbf{s}}_k + k_0\boldsymbol{\mu}_0}{|S_k| + k_0}\right)^\top \left(|S_k| + k_0\right) V_k^{-1}\left(\mathbf{m}_k - \frac{|S_k|\,\bar{\mathbf{s}}_k + k_0\boldsymbol{\mu}_0}{|S_k| + k_0}\right) \\
&\qquad\qquad - \frac{1}{2}\left(\boldsymbol{\mu}_0 - \bar{\mathbf{s}}_k\right)^\top \frac{k_0\,|S_k|}{|S_k| + k_0} V_k^{-1}\left(\boldsymbol{\mu}_0 - \bar{\mathbf{s}}_k\right) - \frac{1}{2}\operatorname{tr}\left(\Lambda_0 V_k^{-1}\right) - \frac{1}{2}\operatorname{tr}\left(\Delta_k V_k^{-1}\right)\Bigg) \\
&= \frac{k_0^{\frac{D}{2}} \left|\Lambda_0\right|^{\frac{\nu_0}{2}} \left|V_k\right|^{-\frac{(\nu_0+|S_k|)+D+2}{2}}}{(2\pi)^{(|S_k|+1)\frac{D}{2}} 2^{\frac{\nu_0 D}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right)} \exp\Bigg(- \frac{1}{2}\left(\mathbf{m}_k - \frac{|S_k|\,\bar{\mathbf{s}}_k + k_0\boldsymbol{\mu}_0}{|S_k| + k_0}\right)^\top \left(|S_k| + k_0\right) V_k^{-1}\left(\mathbf{m}_k - \frac{|S_k|\,\bar{\mathbf{s}}_k + k_0\boldsymbol{\mu}_0}{|S_k| + k_0}\right) \\
&\qquad\qquad - \frac{1}{2}\operatorname{tr}\left(\left(\Lambda_0 + \Delta_k + \frac{k_0\,|S_k|}{|S_k| + k_0}\left(\bar{\mathbf{s}}_k - \boldsymbol{\mu}_0\right)\left(\bar{\mathbf{s}}_k - \boldsymbol{\mu}_0\right)^\top\right) V_k^{-1}\right)\Bigg)
\end{aligned}
$$

Simplifying the notation:

$$\prod_{i\in S_k} q\left(\mathbf{s}_i \mid \boldsymbol{\xi}_k\right) q\left(\boldsymbol{\xi}_k\right) = \frac{k_0^{\frac{D}{2}} \left|\Lambda_0\right|^{\frac{\nu_0}{2}} \left|V_k\right|^{-\frac{\nu_k+D+2}{2}}}{(2\pi)^{(|S_k|+1)\frac{D}{2}} 2^{\frac{\nu_0 D}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right)} \exp\left(-\frac{k_k}{2}\left(\mathbf{m}_k-\boldsymbol{\mu}_k\right)^\top V_k^{-1}\left(\mathbf{m}_k-\boldsymbol{\mu}_k\right) - \frac{1}{2}\operatorname{tr}\left(\Lambda_k V_k^{-1}\right)\right)$$

where $\boldsymbol{\mu}_k = \frac{|S_k|\bar{\mathbf{s}}_k + k_0\boldsymbol{\mu}_0}{|S_k|+k_0}$.

The similarity function then becomes:

$$\begin{aligned} C\left(S_k^{(-i)} \cup \{i\}\right) g\left(s_k^{*(-i)} \cup \mathbf{s}_i\right) =& M \times \Gamma\left(|S_k|\right) \times \int \prod_{i\in S_k} q\left(\mathbf{s}_i \mid \boldsymbol{\xi}_k\right) q\left(\boldsymbol{\xi}_k\right) d\boldsymbol{\xi}_k \\ =& M \times \Gamma\left(|S_k|\right) \times \frac{1}{(2\pi)^{|S_k|\frac{D}{2}}} \frac{k_0^{\frac{D}{2}} \left|\Lambda_0\right|^{\frac{\nu_0}{2}} 2^{\frac{\nu_k D}{2}} \Gamma_D\left(\frac{\nu_k}{2}\right)}{k_k^{\frac{D}{2}} \left|\Lambda_k\right|^{\frac{\nu_k}{2}} 2^{\frac{\nu_0 D}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right)} \end{aligned}$$

so that

$$\frac{C\left(S_k^{(-i)} \cup \{i\}\right) g\left(s_k^{*(-i)} \cup \mathbf{s}_i\right)}{C\left(S_k^{(-i)}\right) g\left(s_k^{*(-i)}\right)} = \frac{\Gamma\left(|S_k|\right)}{\Gamma\left(\left|S_k^{(-i)}\right|\right)} \times \frac{(2\pi)^{\left|s_k^{(-i)}\right|\frac{D}{2}} k_k^{(-i)\frac{D}{2}} \left|\Lambda_k^{(-i)}\right|^{\frac{\nu_k^{(-i)}}{2}} 2^{\frac{\nu_k D}{2}} \Gamma_D\left(\frac{\nu_k}{2}\right)}{(2\pi)^{|S_k|\frac{D}{2}} k_k^{\frac{D}{2}} \left|\Lambda_k\right|^{\frac{\nu_k}{2}} 2^{\frac{\nu_k^{(-i)} D}{2}} \Gamma_D\left(\frac{\nu_k^{(-i)}}{2}\right)}$$

while

$$C\left(\{i\}\right) g\left(\mathbf{s}_i\right) = M \times \Gamma\left(|\{i\}|\right) \times \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{k_0^{\frac{D}{2}} \left|\Lambda_0\right|^{\frac{\nu_0}{2}} 2^{\frac{(\nu_0+1)D}{2}} \Gamma_D\left(\frac{\nu_0+1}{2}\right)}{(k_0+1)^{\frac{D}{2}} \left|\Lambda_0 + \frac{k_0}{1+k_0}\left(\mathbf{s}_i-\boldsymbol{\mu}_0\right)\left(\mathbf{s}_i-\boldsymbol{\mu}_0\right)^\top\right|^{\frac{\nu_0+1}{2}} 2^{\frac{\nu_0 D}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right)}$$

where $\bar{\mathbf{s}}_k^{(-i)}$ is the mean of the elements belonging to $S_k^{(-i)}$ and

$$\begin{aligned} \Delta_k^{(-i)} &= \sum_{j\in S_k^{(-i)}} \left(\mathbf{s}_j - \bar{\mathbf{s}}_k^{(-i)}\right)\left(\mathbf{s}_j - \bar{\mathbf{s}}_k^{(-i)}\right)^\top \\ k_k^{(-i)} &= k_0 + \left|S_k^{(-i)}\right| \\ \nu_k^{(-i)} &= \nu_0 + \left|S_k^{(-i)}\right| \\ \Lambda_k^{(-i)} &= \left(\Lambda_0 + \Delta_k^{(-i)} + \frac{k_0\left|S_k^{(-i)}\right|}{\left|S_k^{(-i)}\right| + k_0}\left(\bar{\mathbf{s}}_k^{(-i)} - \boldsymbol{\mu}_0\right)\left(\bar{\mathbf{s}}_k^{(-i)} - \boldsymbol{\mu}_0\right)^\top\right) \end{aligned}$$

A.2 Additional details on data

Figure A.1 illustrates the spatial distribution of the estimated AR(1) model parameters. Panel (a) presents the mean PM10 concentrations recorded across different stations, showing the highest values – often exceeding 40 $\mu g/m^3$ – in urban and industrial areas like Turin, Milan, Brescia, Verona, and Venice. In contrast, lower concentrations (below 20 $\mu g/m^3$) are found in northern and coastal regions, such as Bolzano and Genoa. The Po Valley exhibits moderate concentrations, generally between 20 and 40 $\mu g/m^3$. Panel (b) displays the estimated autocorrelation, which reveals strong

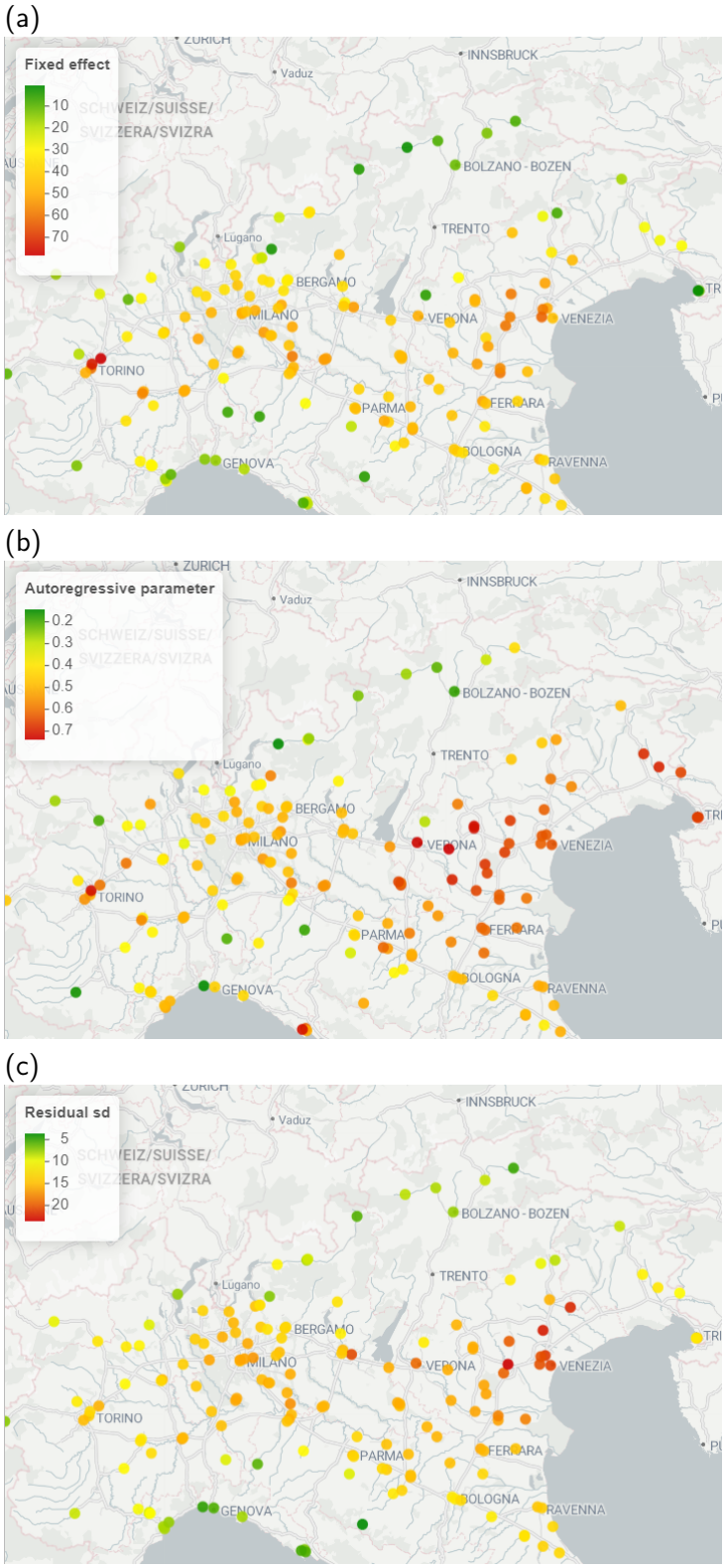


Figure A.1: Maps with mean (a) autocorrelation (b) and standard deviation (c) of AR(1) model applied to each PM10 concentration time series.

persistence in eastern Po Valley, the Turin area, and southern Liguria, while moderate values are observed around Milan and lower values appear in mountainous regions. Panel (c) shows the residual standard deviation, highlighting how much PM10 levels fluctuate over time. Regions with higher mean concentrations, particularly urban and industrial areas, tend to experience greater fluctuations, while areas with lower mean concentrations and persistence show less variability. These spatial patterns motivate our decision to focus the clustering on persistence and residual variability. Panels (b) and (c) suggest that stations tend to group spatially based on autocorrelation and variability, with a similar grouping also reflected in the mean levels.

A.3 Additional results

In the spatial clustering obtained using the Binder loss, shown in panel (a) Figure A.2, four distinct clusters are identified. The blue cluster has a median autocorrelation of 0.16 and a variance of 0.76, while the green cluster shows higher values, with an autocorrelation of 0.46 and variance of 1.00. The red cluster has even stronger autocorrelation (0.61) and higher variance (1.57), and the black cluster displays intermediate values, with an autocorrelation of 0.87 and a variance of 1.23. Unlike the clustering derived from the VI loss, the characteristics of the clusters here are less clearly interpretable.

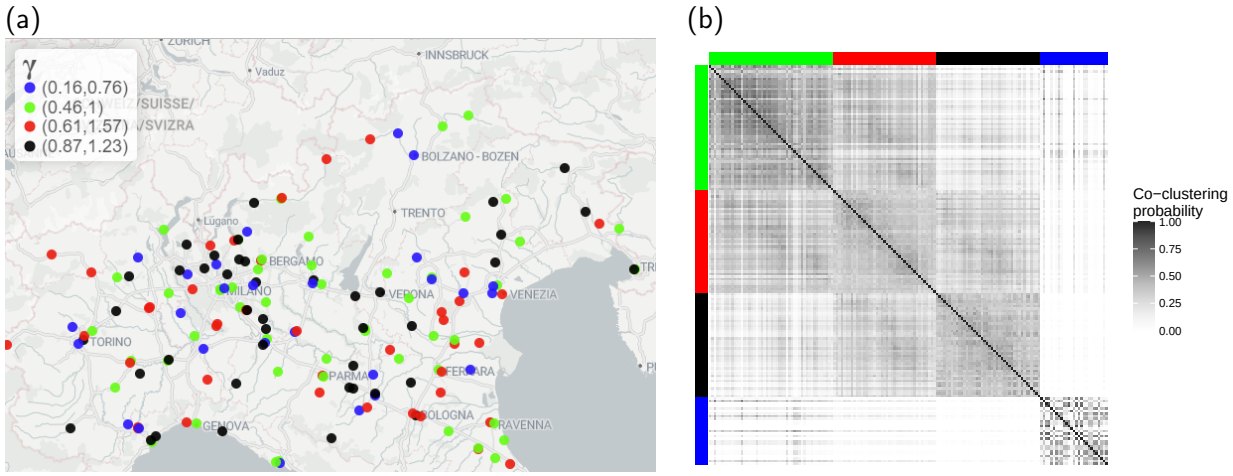


Figure A.2: Estimated partition of the monitoring stations (a); posterior co-clustering matrix (b).

Panel (b) shows a heatmap of the co-clustering matrix, where each entry represents the proportion of MCMC iterations in which two individuals (monitoring stations) were clustered together based on posterior samples. The stations are ordered according to the point estimate of the partition obtained using the Binder loss function. Unlike the heatmap obtained with the VI loss function, shown in Figure 1.4 this one does not reveal a clear block diagonal structure, indicating that the four identified clusters do not align well with the co-clustering probabilities.

Appendix B

Chapter 2 supplementary materials

B.1 MCMC algorithm

Our models rely on posterior inference, which is achieved through MCMC posterior simulations. To accomplish this, we employ the blocked Gibbs Sampler with some adaptive and some classic Metropolis-Hastings steps, based on the work by Gamerman and Lopes (2006) and Miasojedow et al. (2013). This approach allows us to achieve the desired acceptance rates while updating all random parameters in our model. It is important to note that ϕ_i corresponds to θ_{u_i}, and we update ϕ_i by updating both $\boldsymbol{\theta}$ and u_i, and, ultimately, that $\boldsymbol{\eta}$ is constrained to be between 0 and a positive constant M. The MCMC algorithm proceeds through the following steps, with a noteworthy observation that step 8 is further subdivided into three parts, depending on the selected model used to represent disease dependence. Here, we present the algorithm used for the data analysis described in Section 2.9. In the simulation setting, there would be an additional step to calculate the error variances, along with the closed-form full conditionals for the error variances, as well as for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

B.2 Further results on simulations

Table B.1: Root median squared error for all the model parameters under all the simulation settings.

True graph	Unstructured			Directed			Undirected		
Fitted graph	Unstructured	Directed	Undirected	Unstructured	Directed	Undirected	Unstructured	Directed	Undirected
$\boldsymbol{\beta}$	1.585	1.475	1.107	0.993	0.869	0.793	0.642	0.548	0.466
ϕ_1	7.364	7.286	6.541	15.345	6.862	15.915	16.559	16.478	6.052
ϕ_2	7.937	7.088	7.033	14.943	6.491	16.339	16.226	16.431	6.447
ϕ_3	7.971	7.126	6.832	15.386	6.875	15.549	16.440	15.226	6.265
ϕ_4	7.908	6.924	6.923	15.254	6.805	15.488	16.566	15.638	6.345
γ_1	8.160	8.215	7.995	8.727	8.623	8.342	7.995	8.004	7.487
γ_2	12.448	14.472	4.419	14.008	16.880	5.140	10.869	13.610	4.031
γ_3	14.770	29.796	7.681	16.414	33.943	7.792	13.457	29.285	6.770
γ_4	18.784	45.308	9.233	23.094	72.170	9.337	15.280	44.172	6.264
$\boldsymbol{\theta}$	7.783	7.579	7.892	8.675	8.196	9.208	8.782	8.418	8.479
τ_{dis}	7.869	7.864	8.931	15.854	12.490	21.795	21.822	17.967	26.407
$\mathbf{V}$	1.739	1.716	1.762	1.888	1.830	1.882	1.851	1.865	1.876
τ	0.311	0.348	0.253	0.305	0.404	0.261	0.387	0.512	0.351
$\boldsymbol{\rho}$	0.763	0.784	0.798	0.817	0.750	0.707	0.726	0.695	0.666
$\boldsymbol{\eta}$	0.332	0.359	0.360	0.368	0.329	0.344	0.332	0.340	0.301
$\mathbf{A}$	7.625	-	-	-	-	-	-	-	-
$\boldsymbol{\alpha}$	-	-	-	-	7.789	-	-	-	-
ρ_{dis}	-	-	-	-	-	-	-	-	0.250

Algorithm 3: Posterior inference of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, τ_s, $\mathbf{V}$, $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$, $\boldsymbol{\eta}$ and $\mathbf{A}$ or $\boldsymbol{\alpha}$ or ρ_{dis} based on our models

1: update $\boldsymbol{\beta} \mid \boldsymbol{\theta}$

(a) sample candidate $\boldsymbol{\beta}^*$ from $\mathcal{N}(\boldsymbol{\beta}, \xi_{\boldsymbol{\beta}})$

(b) accept $\boldsymbol{\beta}^*$ with probability

$$\min\left\{1, \frac{\exp\left(\sum_{k=1}^{K}\sum_{i,d:\phi_{id}=\theta_k} y_{id}(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d^* + \phi_{id}) - E_{id}\exp\left(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d^* + \phi_{id}\right) - \boldsymbol{\beta}_d^{*\top}\boldsymbol{\beta}_d^*/2\sigma_\beta^2\right)}{\exp\left(\sum_{k=1}^{K}\sum_{i,d:\phi_{id}=\theta_k^*} y_{id}(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \phi_{id}) - E_{id}\exp\left(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \phi_{id}\right) - \boldsymbol{\beta}_d^\top\boldsymbol{\beta}_d/2\sigma_\beta^2\right)}\right\}$$

(c) adapt $\xi_{\boldsymbol{\beta}}$

2: update $\boldsymbol{\theta} \mid \boldsymbol{\beta}, \tau_s$

(a) sample candidate $\boldsymbol{\theta}^*$ from $\mathcal{N}(\boldsymbol{\theta}, \xi_{\boldsymbol{\theta}})$

(b) accept $\boldsymbol{\theta}^*$ with probability

$$\min\left\{1, \frac{\exp\left(\sum_{k=1}^{K}\sum_{i,d:\phi_{id}=\theta_k^*} y_{id}(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \phi_{id}) - E_{id}\exp\left(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \phi_{id}\right) - \theta_k^{*2}\tau_s/2\right)}{\exp\left(\sum_{k=1}^{K}\sum_{i,d:\phi_{id}=\theta_k} y_{id}(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \phi_{id}) - E_{id}\exp\left(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \phi_{id}\right) - \theta_k^2\tau_s/2\right)}\right\}$$

(c) adapt $\xi_{\boldsymbol{\theta}}$

3: update $\gamma_{id} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\eta}, \mathbf{A}$, $i = 1, \ldots, n$, $d = 1, \ldots, q$

(a) sample candidate γ_{id}^* from $\mathcal{N}(\gamma_{id}, \xi_{\gamma_{id}})$

(b) compute the corresponding candidate $u_{id}^* = \sum_{k=1}^{K} kI\left(\sum_{t=1}^{k-1} p_t < F^{(id)}(\gamma_{id}^*) < \sum_{t=1}^{k} p_t\right)$

(c) accept γ_{id}^* with probability

$$\min\left\{1, \frac{\exp\left(y_{id}(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \theta_{u_{id}^*}) - E_i\exp\left(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \theta_{u_{id}^*}\right) - \boldsymbol{\gamma}^{*T}\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma}^*/2\right)}{\exp\left(y_{id}(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \theta_{u_{id}}) - E_{id}\exp\left(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \theta_{u_{id}}\right) - \boldsymbol{\gamma}^T\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma}/2\right)}\right\}$$

(d) adapt $\xi_{\gamma_{id}}$

4: update $\mathbf{V} \mid \boldsymbol{\beta}, \boldsymbol{\theta}$

(a) let $\widetilde{\mathbf{V}} = \text{logit}(\mathbf{V})$ and sample $\widetilde{\mathbf{V}}^*$ from $\mathcal{N}(\widetilde{\mathbf{V}}, \xi_{\widetilde{\mathbf{V}}})$ then $\mathbf{V}^* = \exp\left(\widetilde{\mathbf{V}}^*\right)/\left(1 + \exp\left(\widetilde{\mathbf{V}}^*\right)\right)$

(b) compute the corresponding candidate $\mathbf{p}^* = (p_1^*, \ldots, p_K^*)$ and $\mathbf{u} = \{u_{id}^*\}$

(c) accept $\mathbf{V}^*$ with probability

$$\min\left\{1, \frac{\exp\left(\sum_{i,d} y_{id}(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \theta_{u_{id}^*}) - E_{id}\exp\left(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \theta_{u_{id}^*}\right)\right)\prod_{k=1}^{K} V_k^*(1-V_k^*)^\alpha}{\exp\left(\sum_{i,d} y_{id}(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \theta_{u_{id}}) - E_{id}\exp\left(\mathbf{x}_{id}^\top\boldsymbol{\beta}_d + \theta_{u_{id}}\right)\right)\prod_{k=1}^{K} V_k(1-V_k)^\alpha}\right\}$$

(d) adapt $\xi_{\widetilde{\mathbf{V}}}$

5: update $\boldsymbol{\rho} \mid \boldsymbol{\gamma}, \boldsymbol{\eta}, \mathbf{A}$

(a) let $\widetilde{\boldsymbol{\rho}} = \text{logit}(\boldsymbol{\rho})$ and sample $\widetilde{\boldsymbol{\rho}}^*$ from $\mathcal{N}(\widetilde{\boldsymbol{\rho}}, \xi_{\widetilde{\boldsymbol{\rho}}})$ then $\boldsymbol{\rho}^* = \exp\left(\widetilde{\boldsymbol{\rho}}^*\right)/\left(1 + \exp\left(\widetilde{\boldsymbol{\rho}}^*\right)\right)$

(b) update $\boldsymbol{\Sigma}_\gamma^*$ with the proposed $\boldsymbol{\rho}^*$

(c) accept $\widetilde{\boldsymbol{\rho}}^*$ with probability

$$\min\left\{1, \frac{|\boldsymbol{\Sigma}_\gamma^*|^{-\frac{N}{2}}\exp\left(-\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_\gamma^{*-1}\boldsymbol{\gamma}/2\right)\prod_{d=1}^{q}\rho_d^*(1-\rho_d^*)}{|\boldsymbol{\Sigma}_\gamma|^{-\frac{N}{2}}\exp\left(-\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma}/2\right)\prod_{d=1}^{q}\rho_d(1-\rho_d)}\right\}$$

(d) adapt $\xi_{\widetilde{\boldsymbol{\rho}}}$

6: sample $\tau_s \,|\, \boldsymbol{\theta}, \beta_0$ from

$$\Gamma\left(a_s + \frac{K}{2}, b_s + \frac{\sum_{k=1}^{K} \theta_k^2}{2}\right)$$

7: update $\boldsymbol{\eta} \,|\, \boldsymbol{\gamma}, \boldsymbol{\rho}, \mathbf{A}$

(a) let $\widetilde{\boldsymbol{\eta}} = \log\left(\frac{\boldsymbol{\eta}}{M-\boldsymbol{\eta}}\right)$ and sample $\widetilde{\boldsymbol{\eta}}^*$ from $\mathcal{N}(\widetilde{\boldsymbol{\eta}}, \xi_{\widetilde{\boldsymbol{\eta}}})$ then $\boldsymbol{\eta}^* = \frac{M \exp(\widetilde{\boldsymbol{\eta}}^*)}{1+\exp(\widetilde{\boldsymbol{\eta}}^*)}$

(b) update $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^*$ with the proposed $\boldsymbol{\eta}^*$

(c) accept $\widetilde{\boldsymbol{\eta}}^*$ with probability

$$\min\left\{1, \frac{|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^*|^{-\frac{N}{2}} \exp\left(-\frac{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{*-1} \boldsymbol{\gamma}}{2}\right) \prod_{d=1}^{q} \frac{\exp(\widetilde{\eta}_d^*)}{(1+\exp(\widetilde{\eta}_d^*))^2}}{|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-\frac{N}{2}} \exp\left(-\frac{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\gamma}}{2}\right) \prod_{d=1}^{q} \frac{\exp(\widetilde{\eta}_d)}{(1+\exp(\widetilde{\eta}_d))^2}}\right\}$$

(d) adapt $\xi_{\widetilde{\boldsymbol{\eta}}}$

8: **Unstructured** update $\mathbf{A} \,|\, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\eta}$

(a) let diagonal elements $\widetilde{a}_{dd} = \log(a_{dd})$ and off-diagonal elements $\widetilde{a}_{dh} = a_{dh}$ for $h < d$. Then, sample candidates $\widetilde{\mathbf{a}}^*$ (lower triangular elements of $\mathbf{A}$) from $\mathcal{N}(\widetilde{\mathbf{a}}, \xi_{\widetilde{\mathbf{a}}})$ and set $a_{dd}^* = \exp(\widetilde{a}_{dd}^*)$ and for $h < d$, $a_{dh}^* = \widetilde{a}_{dh}^*$

(b) construct the proposed $\mathbf{A}^*$ and update $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^*$ accordingly

(c) accept $\mathbf{A}^*$ with probability

$$\min\left\{1, \frac{|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^*|^{-\frac{N}{2}} \exp\left(-\frac{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{*-1} \boldsymbol{\gamma}}{2}\right) p(\mathbf{A}^*) \prod_{d=1}^{q} a_{dd}^*}{|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-\frac{N}{2}} \exp\left(-\frac{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\gamma}}{2}\right) p(\mathbf{A}) \prod_{d=1}^{q} a_{dd}}\right\}$$

(d) adapt $\xi_{\widetilde{\mathbf{a}}}$

Directed sample $\boldsymbol{\alpha}_d \,|\, \boldsymbol{\gamma}, \boldsymbol{\rho}_d$ from

$$\mathcal{N}(\mathbf{H}_d h_d, H_d)$$

where $\mathbf{H}_d = \left(\boldsymbol{\delta}_d^\top Q_d \boldsymbol{\delta}_d + \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}\right)^{-1}$ and $\mathbf{h}_d = \boldsymbol{\delta}_d^\top Q_d \boldsymbol{\gamma}_d + \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\alpha}}$. Here $\boldsymbol{\delta}_d = (F_h)_{h \in N_d}$, $F_h = (\boldsymbol{\gamma}_h, \boldsymbol{\zeta}_h)$, $N_d = \{h : h \sim d, h < d\}$ and $\boldsymbol{\zeta}_h = \left(\sum_{j \sim 1} \boldsymbol{\gamma}_{h,j}, \ldots, \sum_{j \sim n} \boldsymbol{\gamma}_{h,j}\right)$. Then update $\mathbf{A}_{d,h}$ matrices and consequently $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$.

Undirected update $\rho_{dis} \,|\, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\eta}$

(a) let $\widetilde{\rho}_{dis} = \log\left(\frac{\rho_{dis}+1}{1-\rho_{dis}}\right)$ and sample $\widetilde{\rho}_{dis}^*$ from $\mathcal{N}(\widetilde{\rho}_{dis}, \xi_{\widetilde{\rho}_{dis}})$ then $\rho_{dis}^* = \frac{\exp(\widetilde{\rho}_{dis}^*)-1}{1+\exp(\widetilde{\rho}_{dis}^*)}$

(b) update $\boldsymbol{\Lambda}_{dis}^*$ with the proposed ρ_{dis}^* and then $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^*$

(c) accept $\widetilde{\rho}_{dis}^*$ with probability

$$\min\left\{1, \frac{|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^*|^{-\frac{N}{2}} \exp\left(-\frac{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{*-1} \boldsymbol{\gamma}}{2}\right) \frac{\exp(\widetilde{\rho}_{dis}^*)}{(1+\exp(\widetilde{\rho}_{dis}^*))^2}}{|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-\frac{N}{2}} \exp\left(-\frac{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\gamma}}{2}\right) \frac{\exp(\widetilde{\rho}_{dis})}{(1+\exp(\widetilde{\rho}_{dis}))^2}}\right\}$$

(d) adapt $\xi_{\widetilde{\rho}_{dis}}$

Table B.2: Boundary detection results in the simulation study with DAGAR spatial dependence for the three disease graph models with data generated under the corresponding true model.

		Disease 1		Disease 2		Disease 3		Disease 4	
Disease graph	T	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
Unstructured	85	0.553	0.868	0.513	0.891	0.538	0.885	0.544	0.880
	90	0.503	0.882	0.454	0.906	0.482	0.900	0.490	0.893
	95	0.433	0.906	0.402	0.921	0.428	0.913	0.434	0.904
	100	0.383	0.921	0.342	0.932	0.359	0.930	0.379	0.918
	105	0.329	0.931	0.297	0.940	0.296	0.942	0.330	0.929
Directed	85	0.563	0.822	0.527	0.867	0.491	0.877	0.480	0.872
	90	0.528	0.838	0.482	0.881	0.450	0.893	0.437	0.881
	95	0.487	0.854	0.436	0.896	0.405	0.903	0.395	0.892
	100	0.441	0.871	0.386	0.906	0.359	0.915	0.351	0.903
	105	0.390	0.890	0.333	0.921	0.310	0.924	0.305	0.914
Undirected	85	0.645	0.838	0.516	0.864	0.589	0.861	0.596	0.857
	90	0.593	0.856	0.470	0.881	0.539	0.875	0.548	0.876
	95	0.534	0.871	0.424	0.892	0.487	0.888	0.494	0.889
	100	0.476	0.887	0.377	0.906	0.437	0.904	0.434	0.902
	105	0.419	0.905	0.330	0.917	0.379	0.918	0.381	0.916

Table B.3: Boundary detection results in the simulation study with CAR spatial dependence for the three disease graph models with data generated under the corresponding true model.

		Disease 1		Disease 2		Disease 3		Disease 4	
Disease graph	T	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
Unstructured	85	0.652	0.859	0.630	0.869	0.650	0.873	0.655	0.871
	90	0.601	0.875	0.581	0.883	0.599	0.886	0.604	0.888
	95	0.544	0.890	0.527	0.897	0.542	0.903	0.545	0.900
	100	0.480	0.904	0.463	0.913	0.485	0.918	0.487	0.913
	105	0.418	0.920	0.408	0.926	0.425	0.932	0.426	0.926
Directed	85	0.669	0.860	0.637	0.895	0.551	0.896	0.537	0.896
	90	0.618	0.877	0.584	0.909	0.501	0.907	0.485	0.907
	95	0.561	0.893	0.527	0.919	0.452	0.919	0.437	0.918
	100	0.503	0.908	0.469	0.929	0.400	0.930	0.387	0.928
	105	0.445	0.924	0.410	0.938	0.350	0.941	0.323	0.941
Undirected	85	0.666	0.871	0.618	0.885	0.643	0.871	0.648	0.879
	90	0.611	0.887	0.567	0.900	0.589	0.883	0.596	0.894
	95	0.553	0.899	0.514	0.915	0.535	0.898	0.538	0.907
	100	0.498	0.915	0.459	0.926	0.480	0.911	0.480	0.921
	105	0.440	0.929	0.405	0.939	0.423	0.924	0.422	0.933

Table B.4: Adjacencies detection with DAGAR spatial dependence for the three disease graph models with data generated under the corresponding true model.

	Disease 1		Disease 2		Disease 3		Disease 4	
Disease graph	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
Unstructured	0.386	0.986	0.598	0.928	0.600	0.959	0.261	1.000
Directed	0.406	0.985	0.786	0.939	0.668	0.954	0.213	1.000
Undirected	0.415	0.981	0.753	0.927	0.571	0.969	0.225	0.990

Table B.5: Adjacencies detection with CAR spatial dependence for the three disease graph models with data generated under the corresponding true model.

	Disease 1		Disease 2		Disease 3		Disease 4	
Disease graph	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
Unstructured	0.084	1.000	0.305	0.998	0.142	1.000	0.045	1.000
Directed	0.160	0.995	0.445	0.963	0.306	0.973	0.211	1.000
Undirected	0.846	0.776	0.834	0.632	0.888	0.672	0.848	0.947

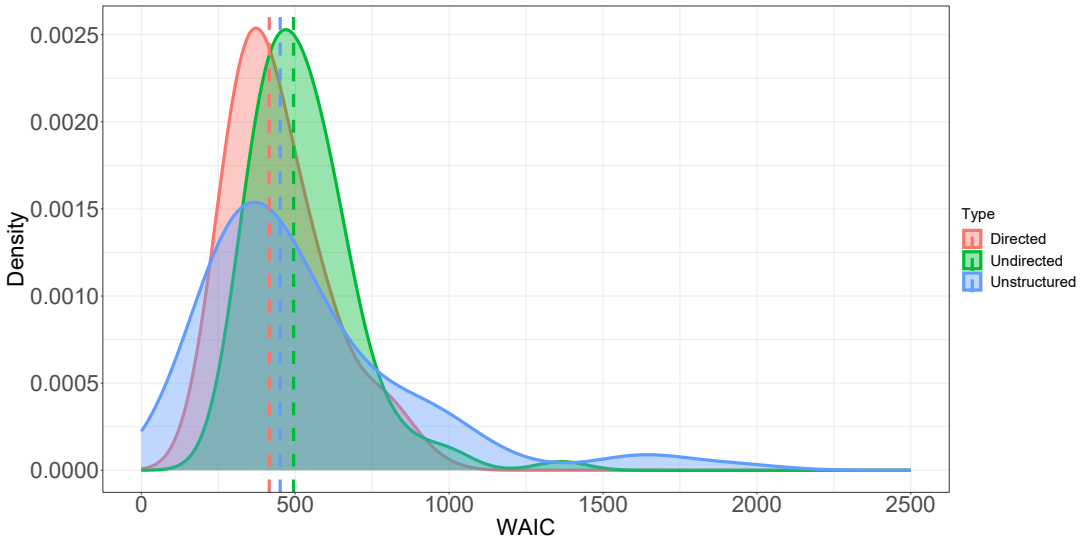


Figure B.1: WAIC densities along with the median values (dotted lines) in the DAGAR spatial dependence case for different choices of the disease graph with data generated under the corresponding true model.

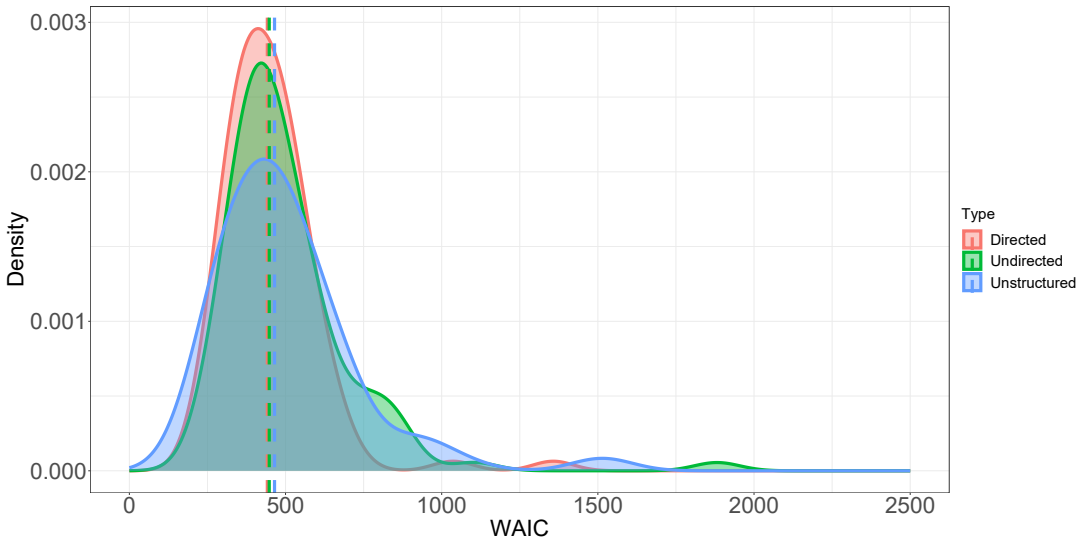


Figure B.2: WAIC densities along with the median values (dotted lines) in the CAR spatial dependence case for different choices of the disease graph with data generated under the corresponding true model.

B.3 Posterior summaries for SEER data analysis

We present further analysis for the model used in Section 2.9. In this section the regression model $\mathbf{x}_{ij}^{\top}\boldsymbol{\beta}_j$ consists of only cancer-specific intercepts while $\mathbf{z}_{ij}$ in the adjacency model is built from the absolute differences between geographic neighbors for the three covariates described in Section 2.1. We present credible intervals in Figure B.3, B.4, B.5 and B.6.

Regarding Figure B.3 We see β_1 (lung) and β_2 (esophageal) having positive posterior medians with substantial mass greater than 0, while β_3 (larynx) and β_4 (colorectal) having negative posterior medians with substantial mass less than 0. The 90% interval for colorectal cancer is almost entirely to the left of 0. Turning to the right panel, we see about 4 atoms (θ_2, θ_{12}, θ_3 and θ_8) to have their HPD intervals almost entirely to the left of 0, while 5 atoms (θ_1, θ_{13}, θ_{10}, θ_9 and θ_7) have intervals situated entirely to the right of 0. The remaining 6 atoms (out of a total of $K = 15$) are not significantly different from 0. The variation in the posterior distribution of these atoms reflect substantial posterior learning and information from the data that propagates to the spatial random effects.

It is expected that the intervals for the γ random effects, shown in Figure B.4, will mostly not be significantly different from 0, but they reveal substantial variation indicating enough information in the data to discern spatial patterns. It is worth pointing out that the dispersion varies substantially among the cancers with the intervals for lung and colorectal spanning a range much smaller than esophageal and larynx.

The variations among the posterior intervals in Figure B.5 indicate substantial learning from the data that informs the joint distribution of the spatial effects through the stick-breaking probabilities and the baseline MDAGAR distribution in the DP.

Coming to Figure B.6, the 12 coefficients in $\boldsymbol{\eta}$ include 3 slopes for each of the 4 cancer types corresponding to the covariates shown in Figure 2.3. For each of the 4 cancers, the parameters in the left panel appear in (2.5) as coefficients of the 3×1 vector $\mathbf{z}_{ij}$, where the elements of $\mathbf{z}_{ij}$ are $|z_{ki} - z_{kj}|$ for $k = 1, 2, 3$ corresponding to the 3 covariates. The four slopes corresponding to smoking are denoted as η_1, η_4, η_7, and η_{10}. For the variable population over 65, the corresponding slopes are η_2, η_5, η_8, and η_{11}. The slopes related to "poverty"" levels are η_3, η_6, η_9, and η_{12}. For all cancers considered, these coefficients correspond respectively to lung, esophageal, laryngeal, and colorectal cancers. Here, the greater the η's, the more likely two regions are not considered to be adjacent by the model. For instance, the smoking covariate plays a significant role in determining adjacencies for larynx cancer (η_7), while the over 65 covariate has a similar impact for both esophageal and larynx cancers (η_5 and η_8). Additionally, the below poverty covariate appears to have a notable effect on the adjacencies related to esophageal cancer (η_6). In the right panel, the HPD intervals for the elements of $\mathbf{A}\mathbf{A}^{\top}$ are shown, representing the non-spatial covariance among the cancers. Lung and larynx cancers exhibit the highest variances $(\mathbf{A}\mathbf{A}^{\top})_{11}$ and $(\mathbf{A}\mathbf{A}^{\top})_{44}$, with larynx showing the widest interval. The non-spatial association between lung and larynx captured by $(\mathbf{A}\mathbf{A}^{\top})_{41}$ and that between lung and colorectal captured by $(\mathbf{A}\mathbf{A}^{\top})_{31}$ are negative with the 95% intervals situated entirely below 0. It is important to interpret the non-spatial component of these associations with caution. What can, perhaps, be gleaned from these negative non-spatial associations is a phenomenon of spatial confounding at the residual scale, where much of the positive associations among these aggregated data are absorbed by the spatial associations arising from geographic proximity.

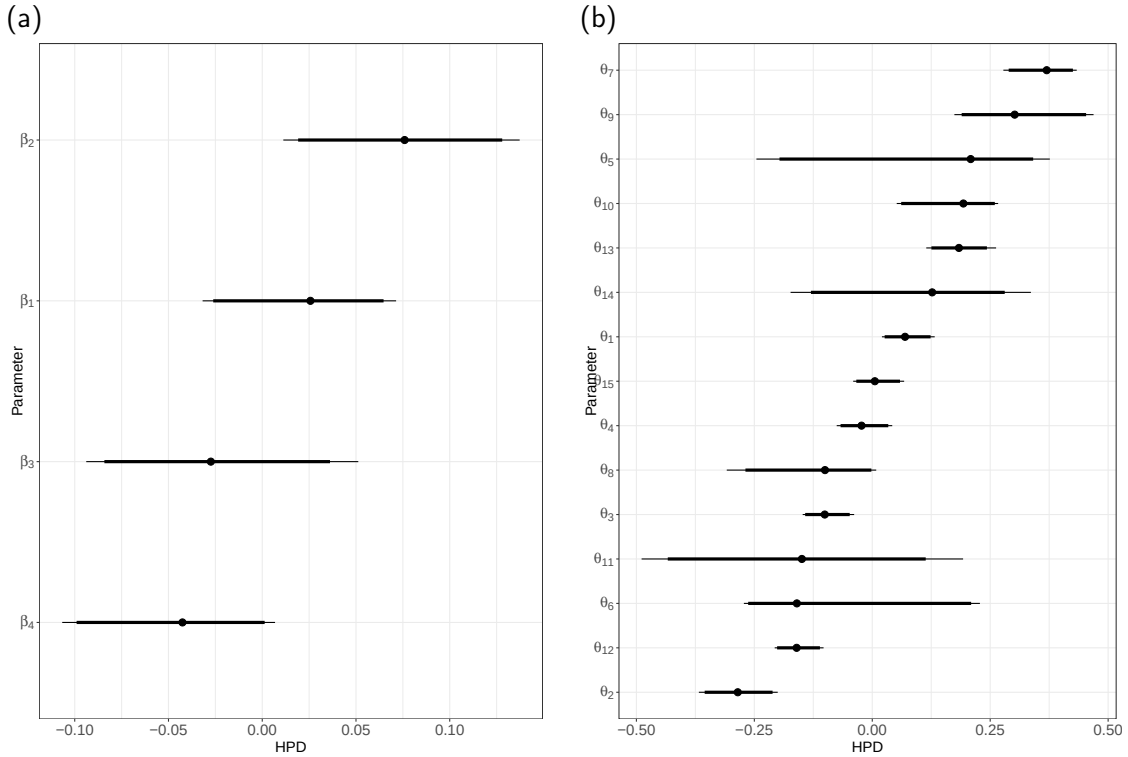


Figure B.3: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for $\boldsymbol{\beta}$ (a) and $\boldsymbol{\theta}$ (b).

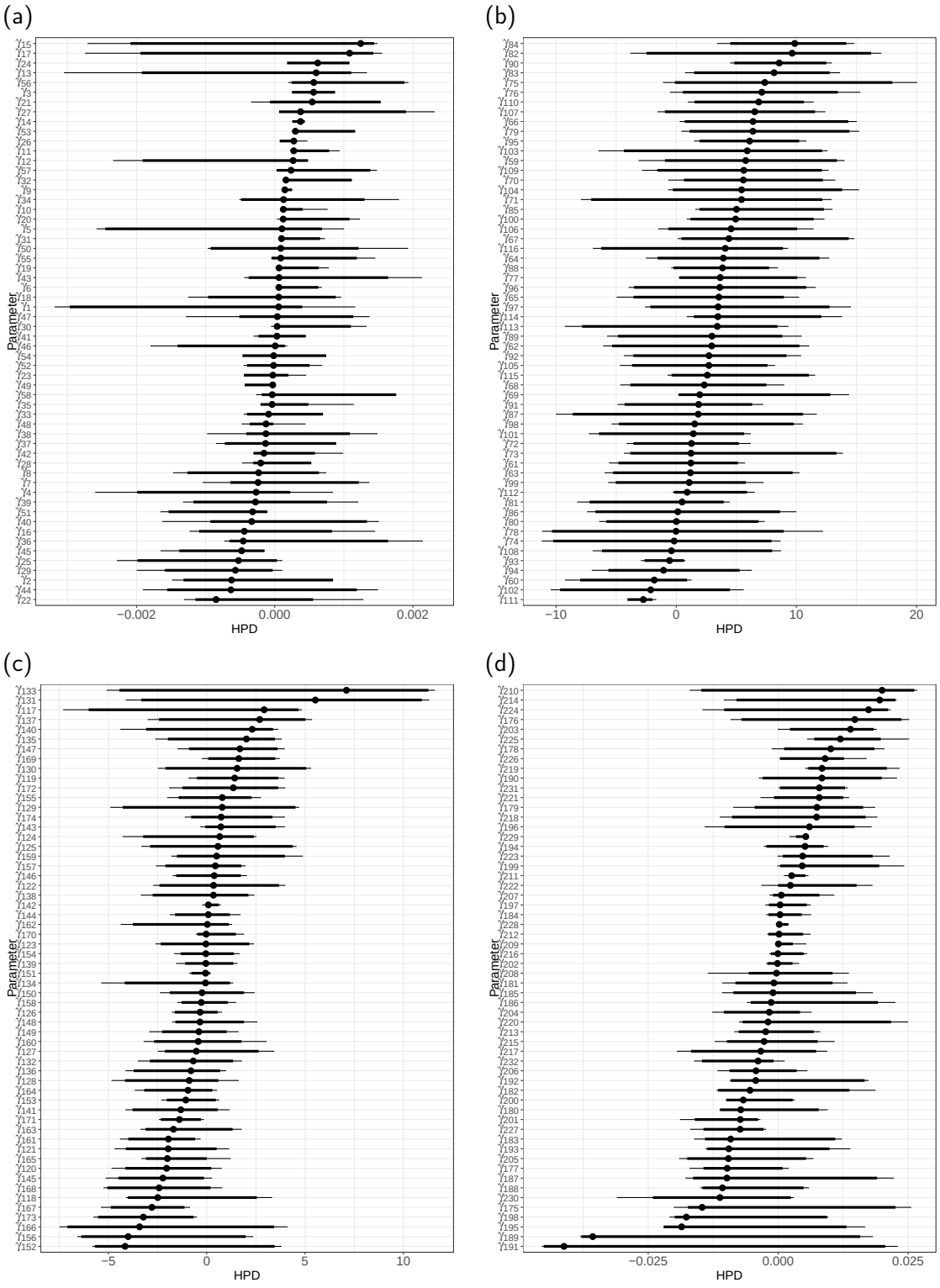


Figure B.4: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for the elements of the disease-specific spatial effects in γ using the MDAGAR specification in Section 2.4.1 and the precision matrix given in Equation 2.6. The four panels correspond to the 4 cancers: lung (a), esophageal (b), larynx (c) and colorectal (d).

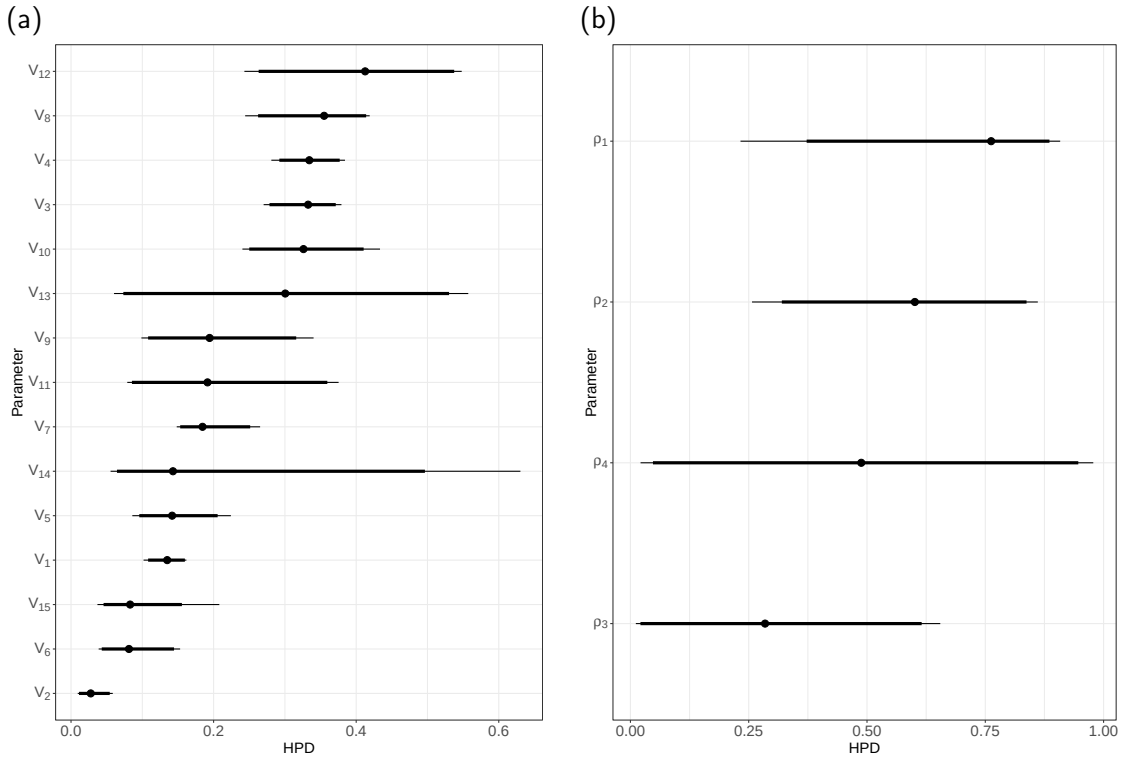


Figure B.5: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for $\mathbf{V}$ (a), which consists of the 15 stick-breaking weights used in Equation (2.2), and for $\boldsymbol{\rho}$ (b), which consists of the 4 spatial autocorrelation parameters (one for each disease) in the MDAGAR model.

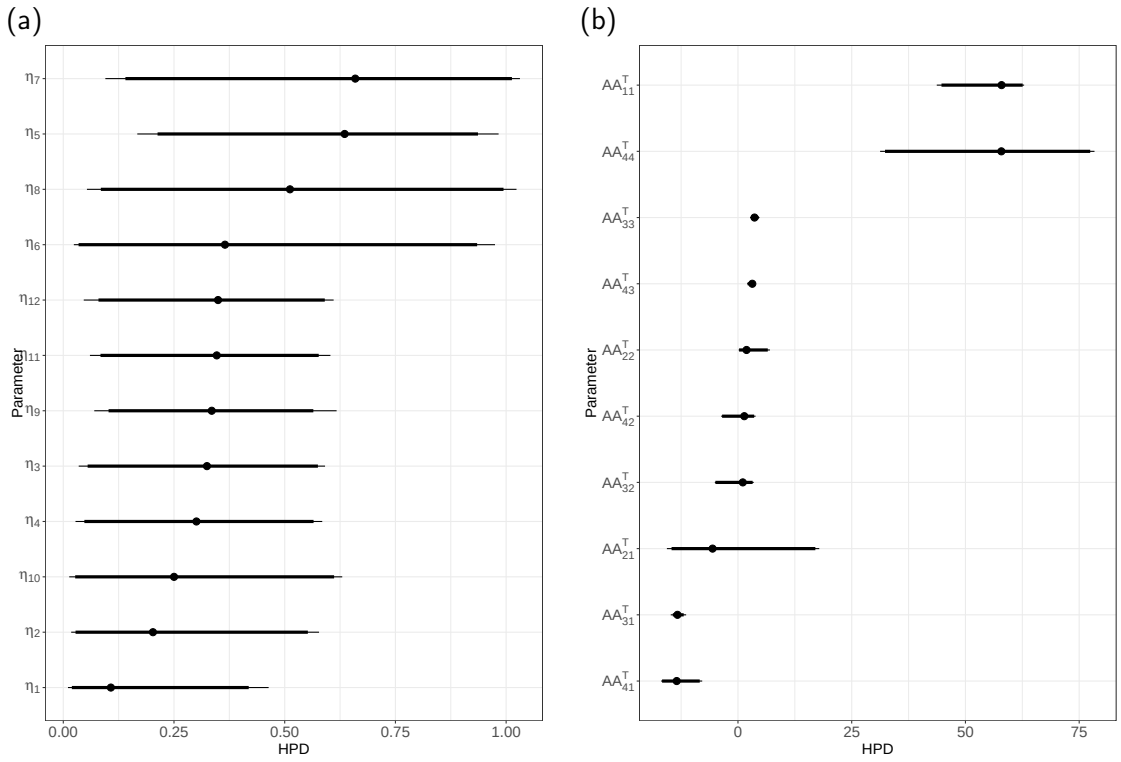


Figure B.6: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for $\boldsymbol{\eta}$ (a) and $\mathbf{A}$ (b).

B.4 Additional figures and tables

Figure B.7: California map with county names. This will be helpful in detecting the boundaries.

B.4.1 Analysis with covariates in the mean and adjacency model

We present inference from the model used in Section 2.9, where we include cancer-specific intercepts and slopes for the three covariates described in Section 2.1 and also include the covariates in the adjacency model. Table B.6 presents the estimates of the model described in Section 2.9 including only cancer-specific intercepts in the regression model and the 3 covariates described in Section 2.1 in the adjacency model. The last column of Table B.6 presents Monte Carlo standard errors corresponding to the parameter estimates.

Table B.6: Posterior estimates, standard deviations and their Monte Carlo standard errors for the regression coefficients, atoms and their precision in the hierarchical model in (2.11) using the Poisson regression model described in Section 2.9.

Variable	Estimate	SD	MC SE
β_1	-0.068	0.037	0.003
β_2	0.029	0.002	$< 10^{-3}$
β_3	-0.011	0.002	$< 10^{-3}$
β_4	-0.009	0.002	$< 10^{-3}$
β_5	-0.361	0.109	0.009
β_6	0.034	0.007	$< 10^{-3}$
β_7	0.010	0.007	$< 10^{-3}$
β_8	-0.010	0.006	$< 10^{-3}$
β_9	-0.394	0.132	0.008
β_{10}	0.032	0.008	0.001
β_{11}	0.000	0.009	$< 10^{-3}$
β_{12}	0.004	0.007	$< 10^{-3}$
β_{13}	0.156	0.052	0.007
β_{14}	0.013	0.002	$< 10^{-3}$
β_{15}	-0.018	0.004	$< 10^{-3}$
β_{16}	-0.006	0.002	$< 10^{-3}$
θ_1	0.021	0.024	0.005
θ_2	0.190	0.077	0.011
θ_3	-0.223	0.060	0.010
θ_4	0.168	0.198	0.029
θ_5	-0.082	0.234	0.030
θ_6	0.209	0.136	0.018
θ_7	-0.081	0.290	0.030
θ_8	0.259	0.139	0.015
θ_9	0.061	0.307	0.036
θ_{10}	-0.071	0.241	0.034
θ_{11}	0.115	0.034	0.007
θ_{12}	-0.144	0.029	0.007
θ_{13}	0.175	0.048	0.008
θ_{14}	0.132	0.145	0.018
θ_{15}	-0.082	0.025	0.005
τ	6.979	2.336	0.026

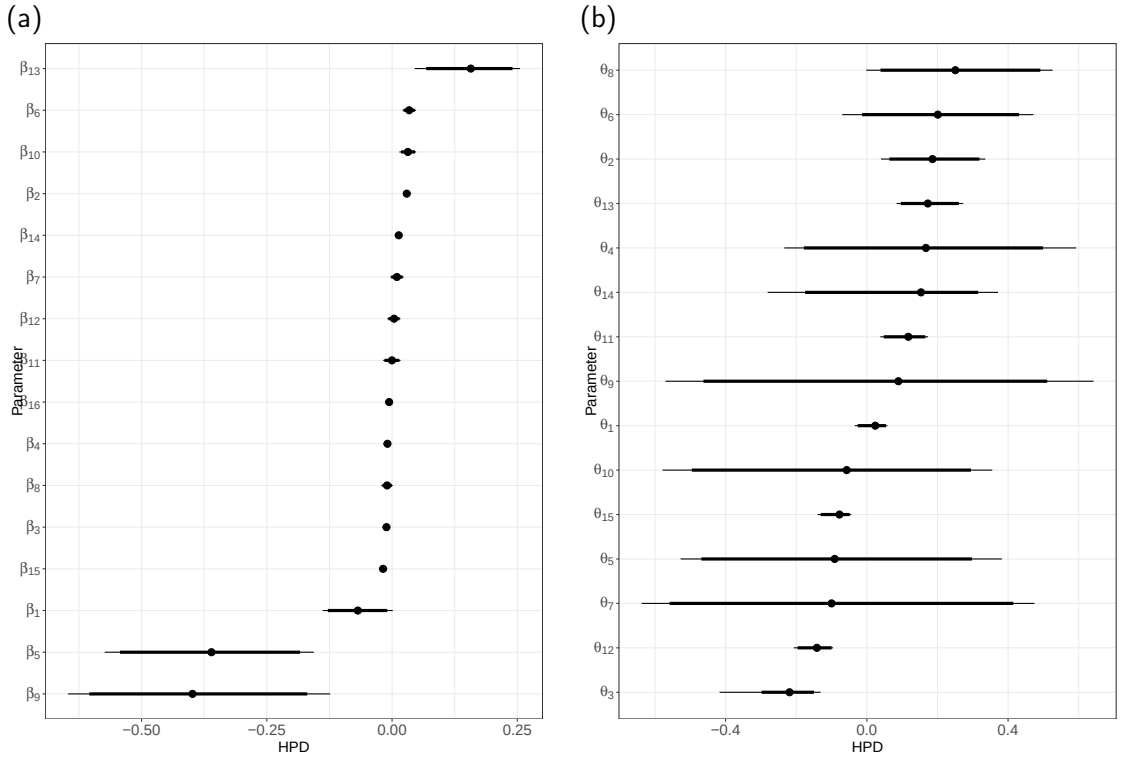


Figure B.8: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for $\boldsymbol{\beta}$ (a) and $\boldsymbol{\theta}$ (b).

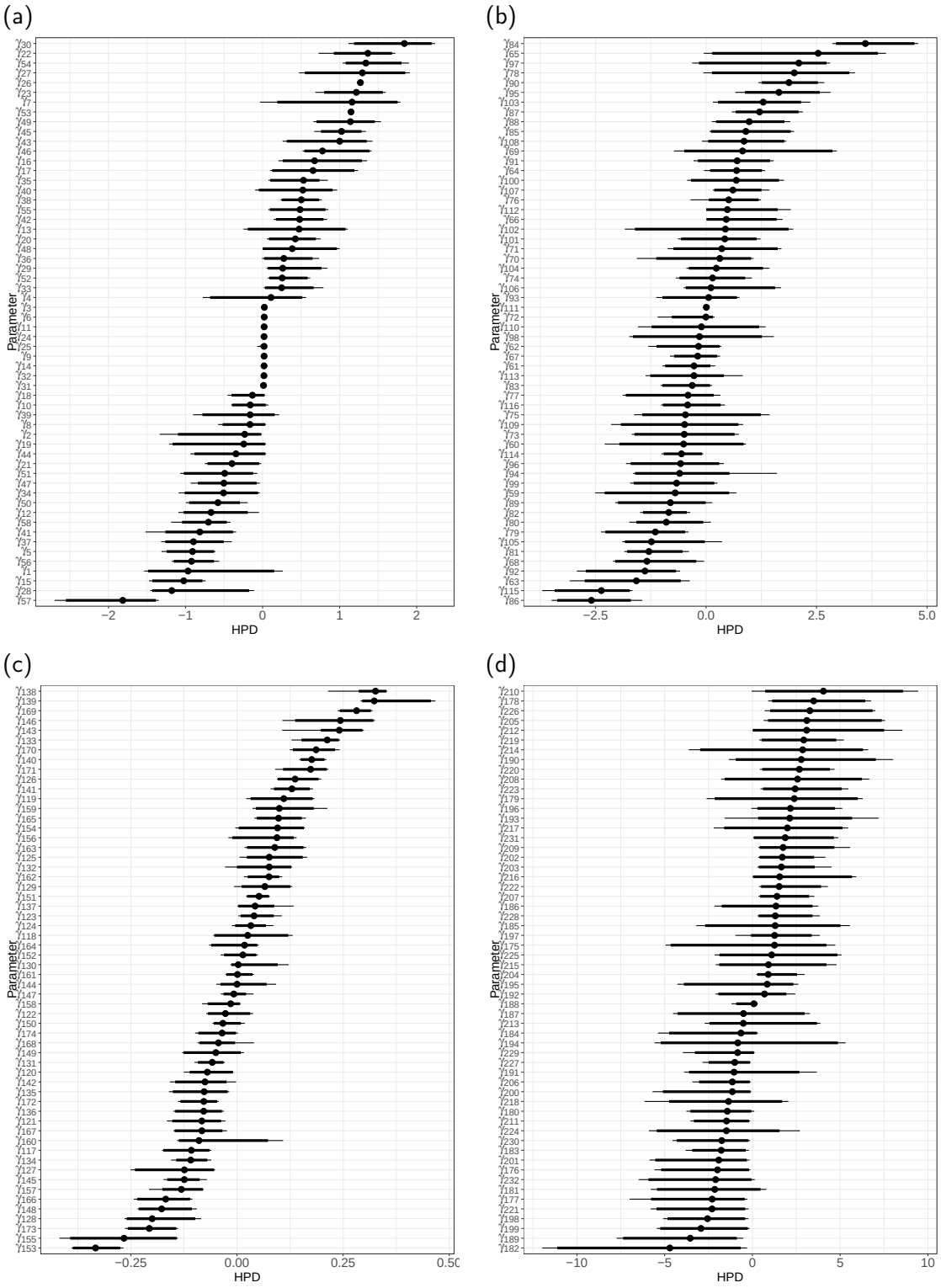


Figure B.9: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for the elements of the disease-specific spatial effects in γ using the MDAGAR specification in Section 2.4.1 and the precision matrix given in Equation 2.6. The four panels correspond to the 4 cancers: lung (a), esophageal (b), larynx (c) and colorectal (d).

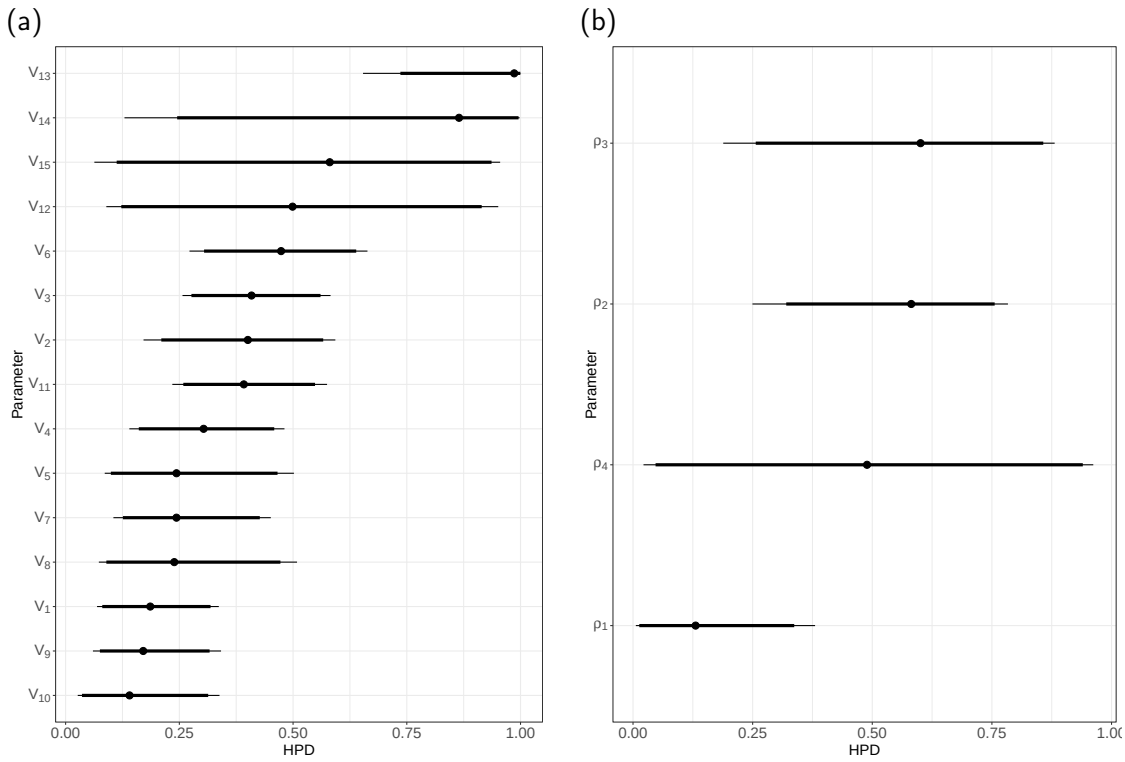


Figure B.10: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for $\mathbf{V}$ (a), which consists of the 15 stick-breaking weights used in Equation (2.2), and for $\boldsymbol{\rho}$ (b), which consists of the 4 spatial autocorrelation parameters (one for each disease) in the MDAGAR model.

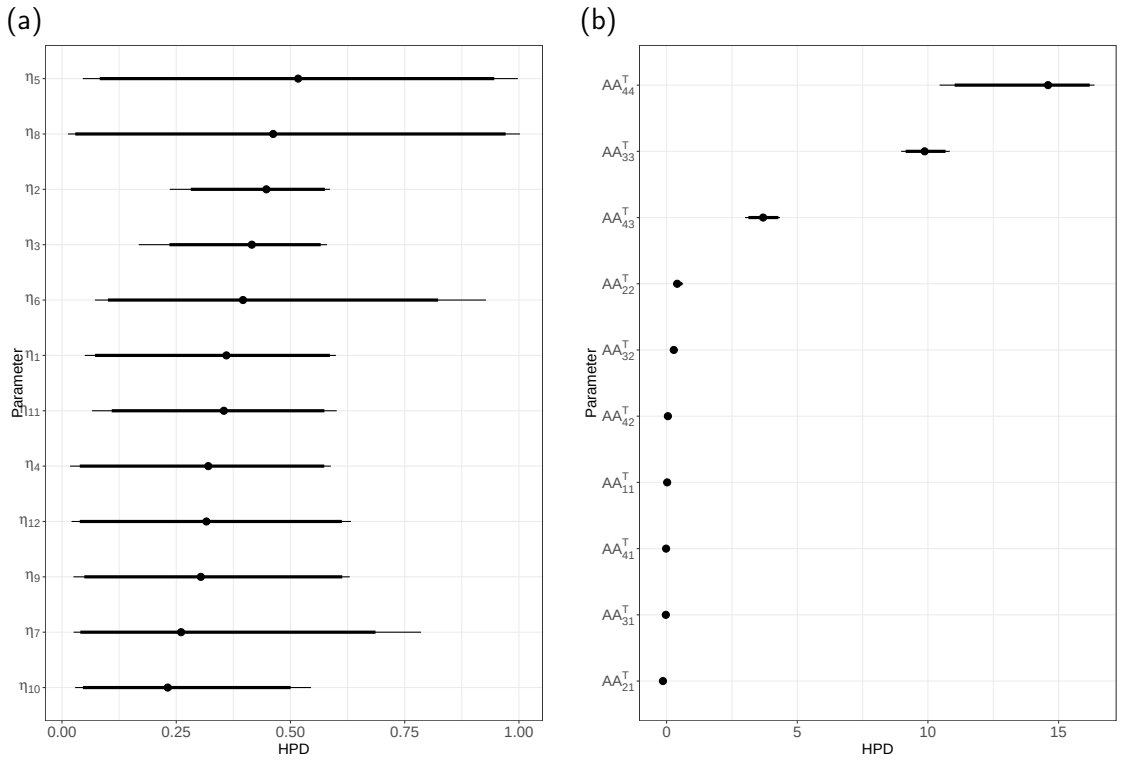


Figure B.11: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for $\boldsymbol{\eta}$ (a) and $\mathbf{A}$ (b).

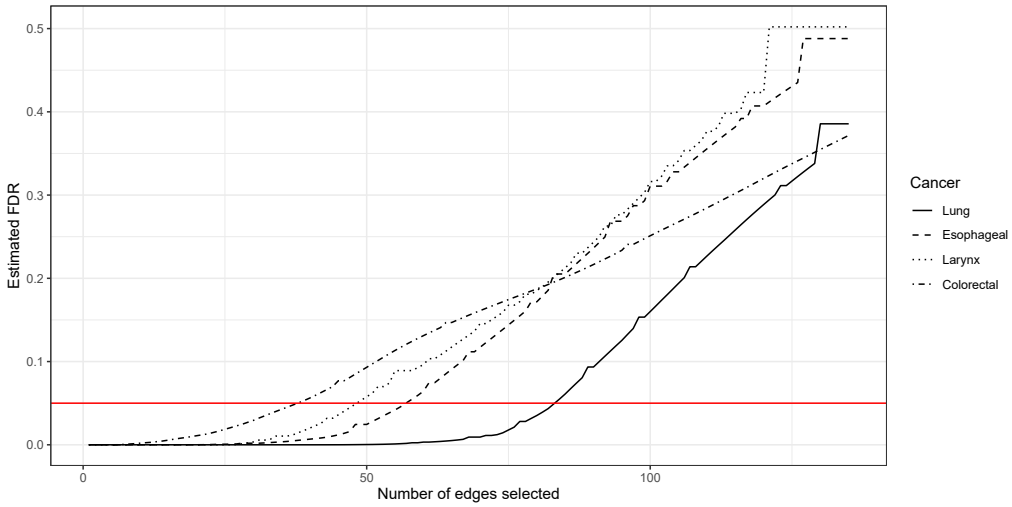


Figure B.12: Estimated FDR curves plotted against the number of selected difference boundaries for the four cancers

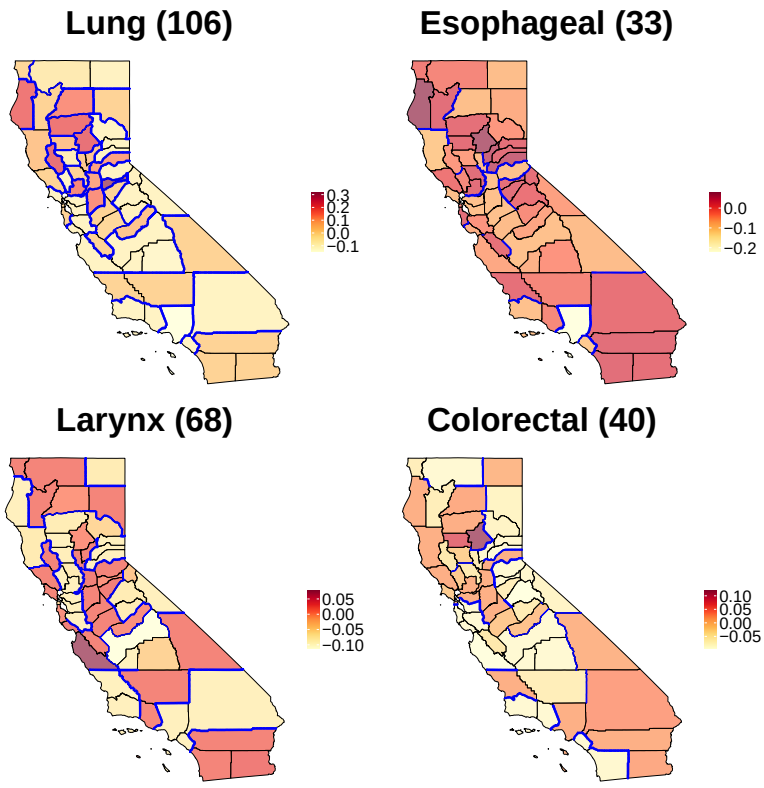


Figure B.13: Difference boundaries (highlighted in red) detected by the model in the California map, colored according to the posterior mean of the corresponding ϕ_{id}, for four cancers individually when $\zeta = 0.05$. The values in brackets are the number of difference boundaries detected.

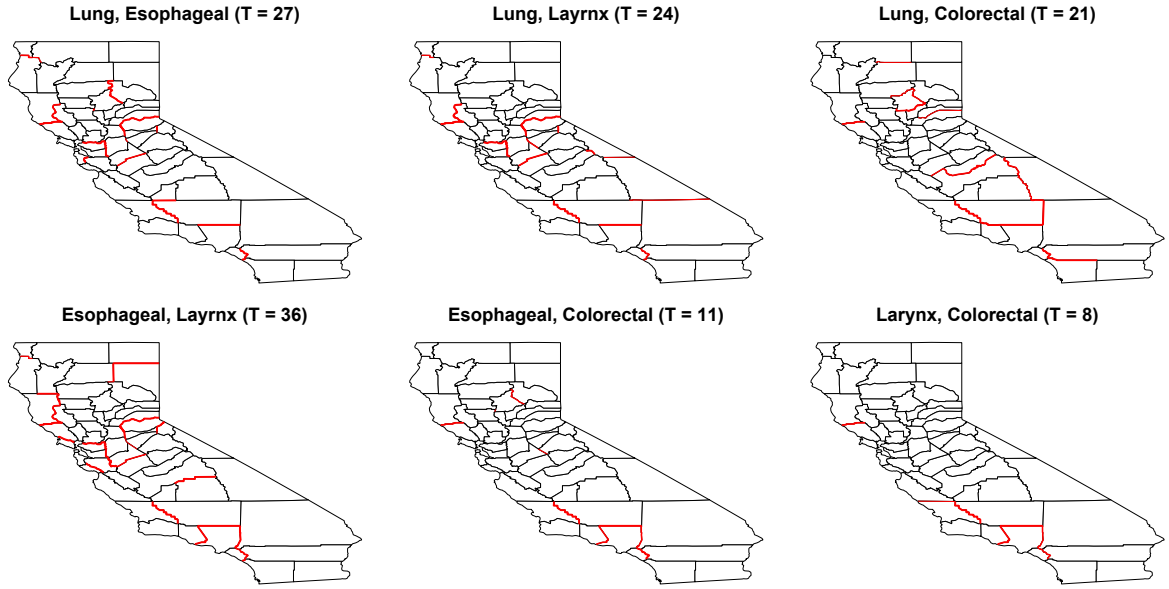


Figure B.14: Shared difference boundaries (highlighted in red) detected by the model for each pair of cancers in California map when $\zeta = 0.05$. The values in brackets are the number of difference boundaries detected.

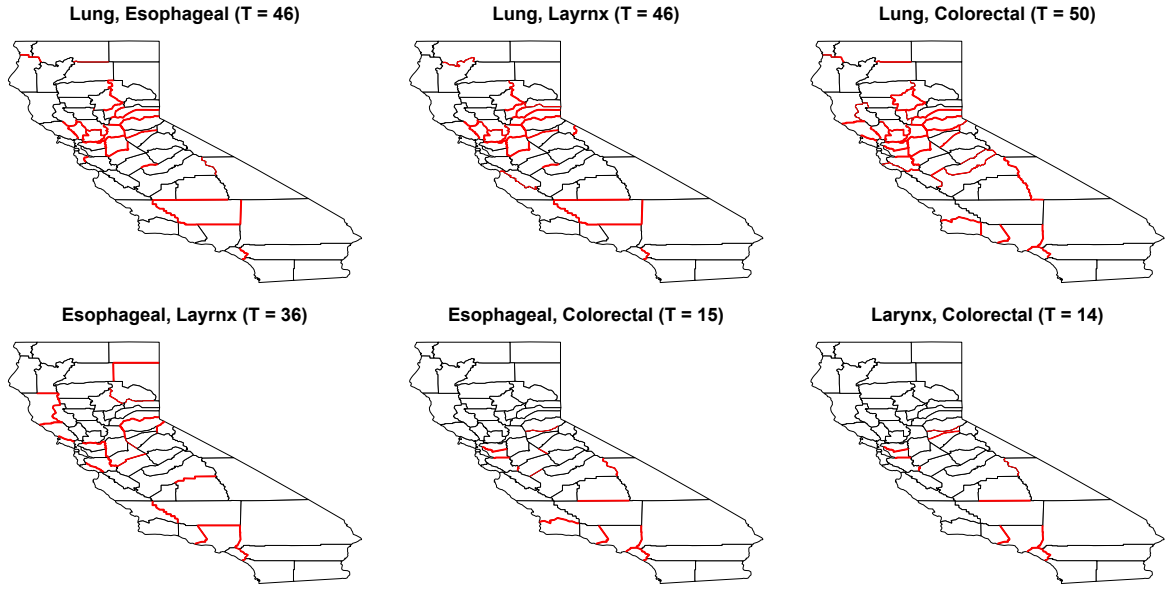


Figure B.15: Mutual cross-difference boundaries (highlighted in red) detected by the model for each pair of cancers in California map when $\zeta = 0.05$. The values in brackets are the number of difference boundaries detected.

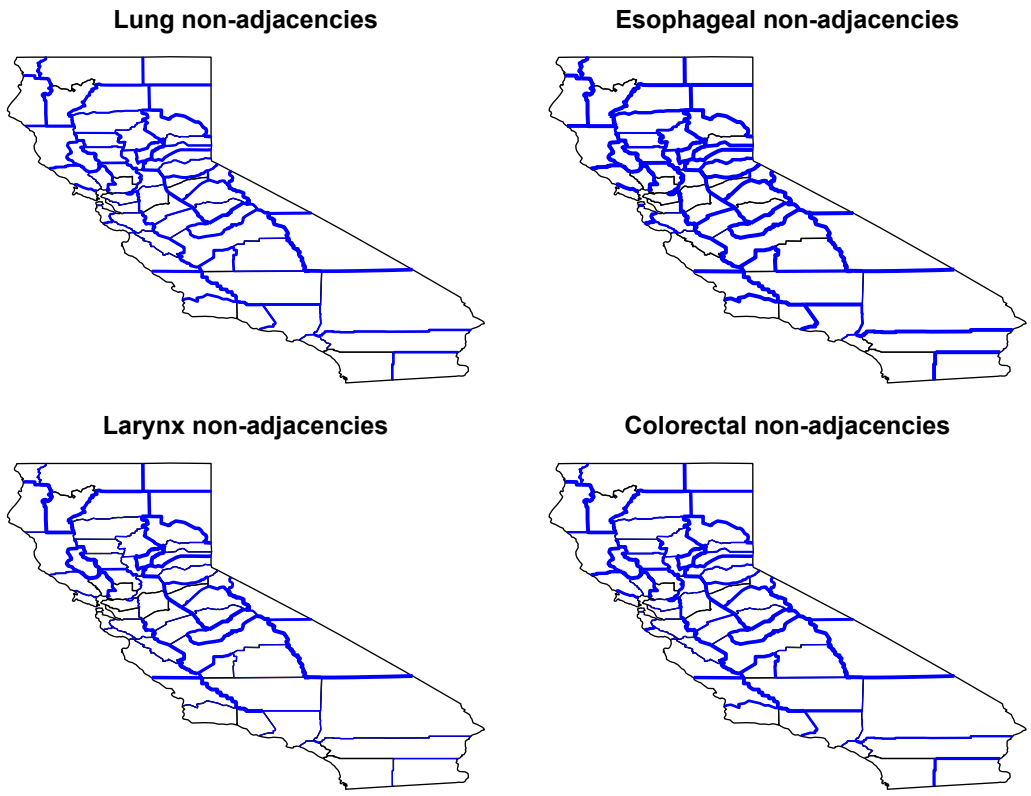


Figure B.16: Non-adjacencies (shown in blue) over the California map. The thickness of the lines is proportional to the probability of being considered as a non-adjacency

B.4.2 Analysis with covariates in the mean with fixed adjacencies

We present inference from the model used in Section 2.9 of the main article, where we include cancer-specific intercepts and slopes for the three covariates described in Section 2.1 of the main article in the regression $\mathbf{x}_{ij}^{\top}\boldsymbol{\beta}_j$, but exclude modeling adjacencies and keep them fixed. Table B.7 presents the estimates of the model described in Section 2.9 including only cancer-specific intercepts in the regression model and the 3 covariates described in Section 2.1 of the main article in the adjacency model. The last column of Table B.7 presents Monte Carlo standard errors corresponding to the parameter estimates.

Table B.7: Posterior estimates and standard deviations for the regression coefficients, atoms and their precision in the hierarchical model in (2.11) using the Poisson regression model described in Section 2.9.

Variable	Estimate	SD	MC SE
β_1	-0.060	0.040	0.005
β_2	0.031	0.003	$< 10^{-3}$
β_3	-0.009	0.003	$< 10^{-3}$
β_4	-0.010	0.003	$< 10^{-3}$
β_5	-0.304	0.104	0.009
β_6	0.028	0.010	0.001
β_7	0.014	0.008	0.001
β_8	-0.009	0.008	0.001
β_9	-0.407	0.150	0.011
β_{10}	0.026	0.010	0.001
β_{11}	0.005	0.010	0.001
β_{12}	0.008	0.008	0.001
β_{13}	0.236	0.061	0.009
β_{14}	0.011	0.004	$< 10^{-3}$
β_{15}	-0.019	0.004	$< 10^{-3}$
β_{16}	-0.005	0.002	$< 10^{-3}$
θ_1	-0.122	0.038	0.014
θ_2	-0.019	0.045	0.013
θ_3	0.001	0.036	0.009
θ_4	-0.016	0.067	0.012
θ_5	-0.033	0.061	0.013
θ_6	-0.026	0.046	0.016
θ_7	-0.089	0.294	0.046
θ_8	-0.087	0.120	0.029
θ_9	-0.231	0.104	0.023
θ_{10}	0.038	0.137	0.054
θ_{11}	0.054	0.125	0.036
θ_{12}	-0.181	0.043	0.017
θ_{13}	-0.034	0.095	0.030
θ_{14}	-0.042	0.241	0.082
θ_{15}	0.082	0.052	0.021
τ	8.042	2.676	0.028

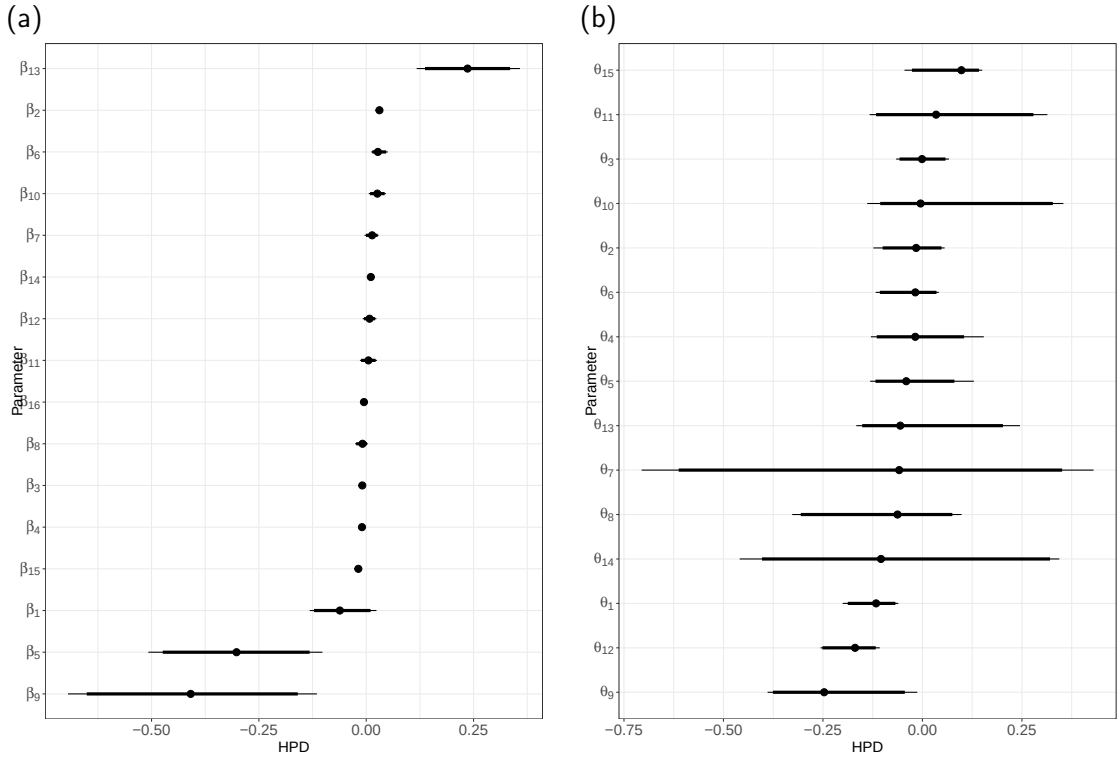


Figure B.17: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for $\boldsymbol{\beta}$ (a) and $\boldsymbol{\theta}$ (b).

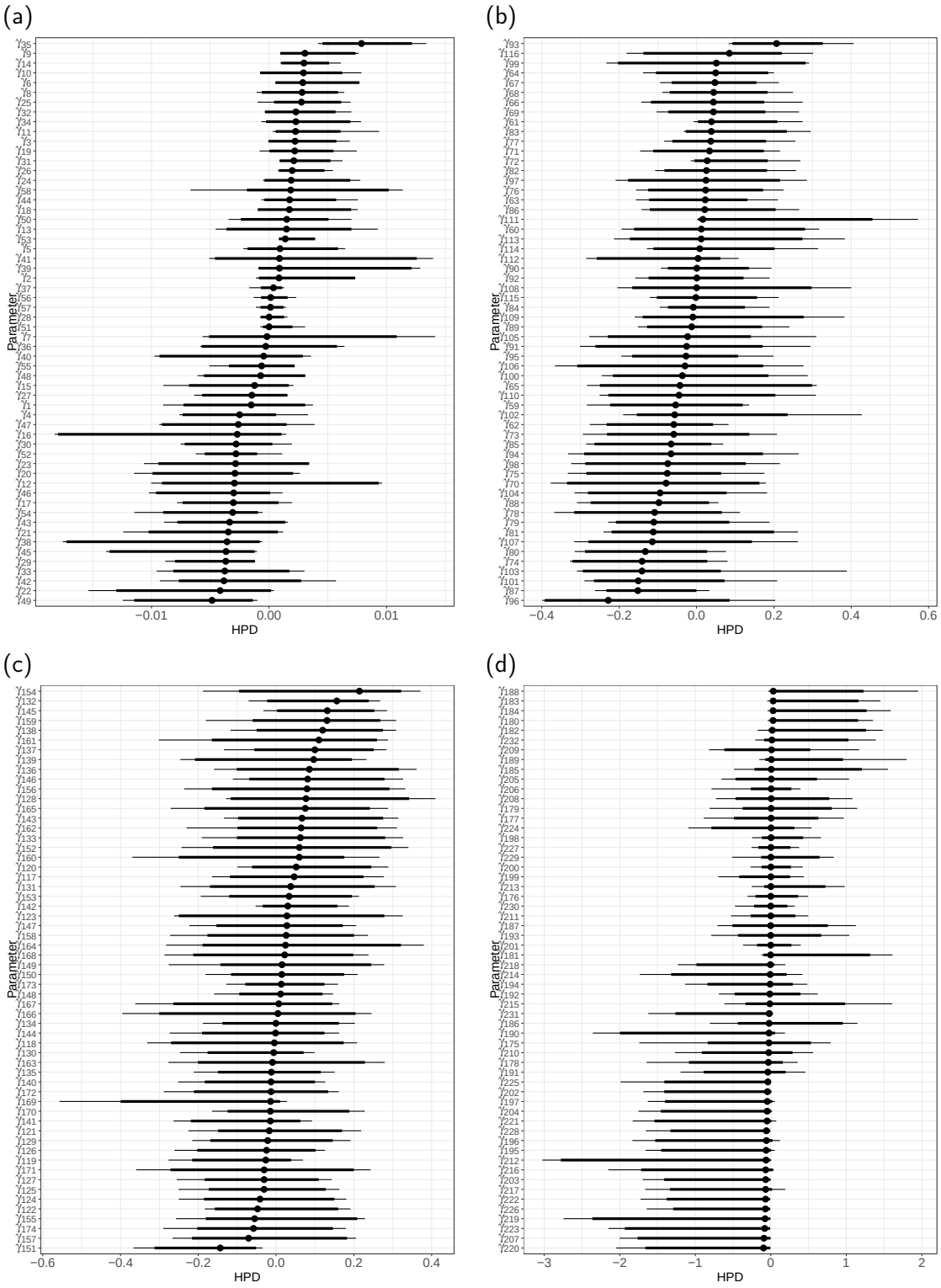


Figure B.18: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for the elements of the disease-specific spatial effects in γ using the MDAGAR specification in Section 2.4.1 and the precision matrix given in Equation 2.6. The four panels correspond to the 4 cancers: lung (a), esophageal (b), larynx (c) and colorectal (d).

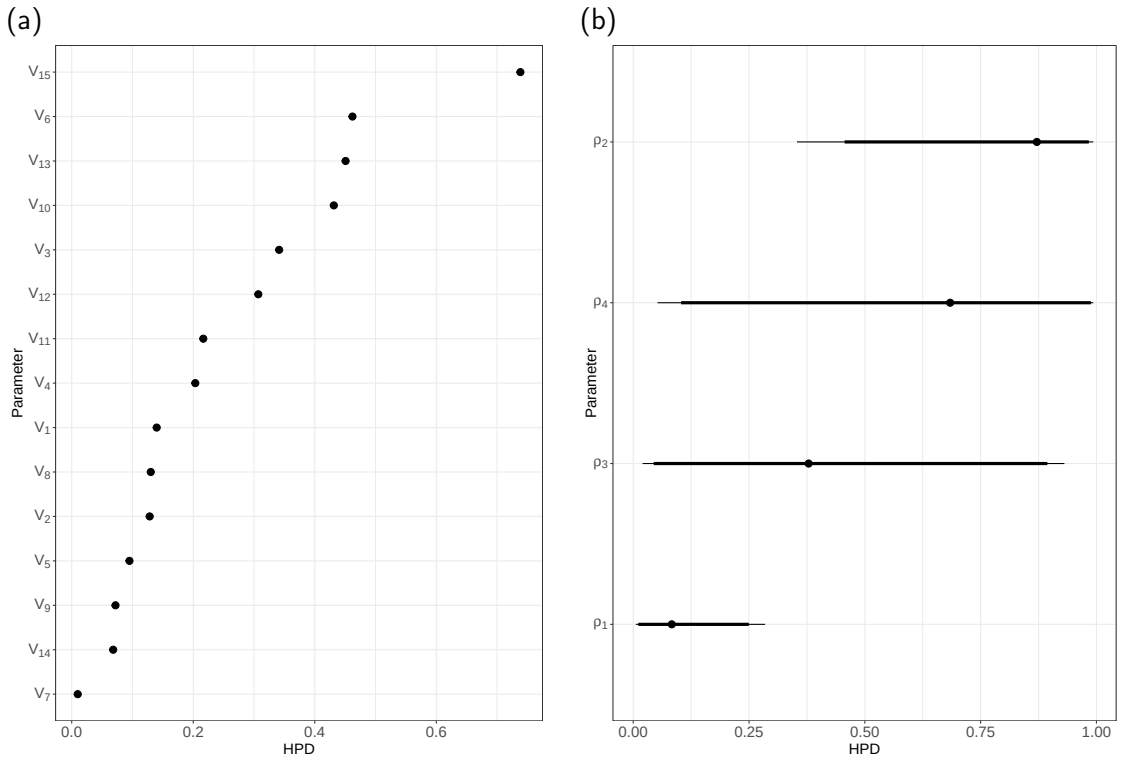


Figure B.19: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for $\mathbf{V}$ (a), which consists of the 15 stick-breaking weights used in Equation (2.2), and for $\boldsymbol{\rho}$ (b), which consists of the 4 spatial autocorrelation parameters (one for each disease) in the MDAGAR model.

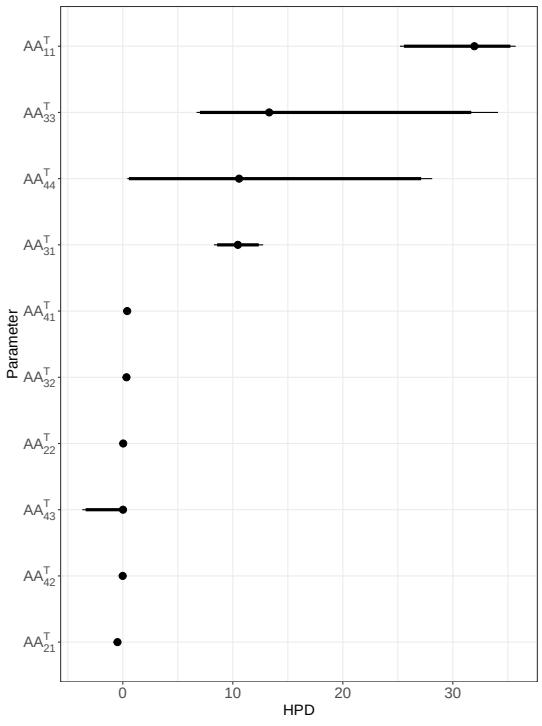


Figure B.20: Highest posterior density (HPD) intervals, ordered according to their posterior medians, at 90% (thick lines) and 95% (thin lines) levels for $\mathbf{A}$.

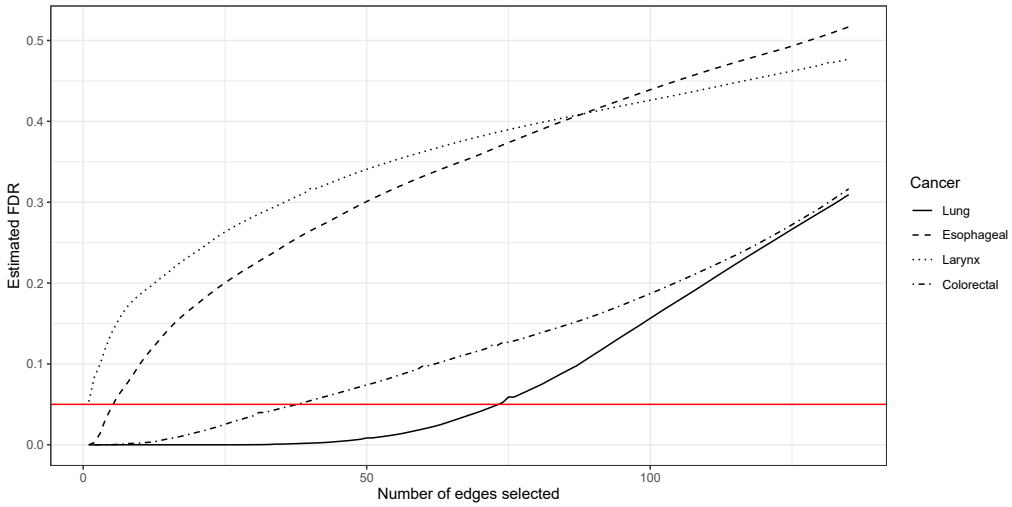


Figure B.21: Estimated FDR curves plotted against the number of selected difference boundaries for four cancers

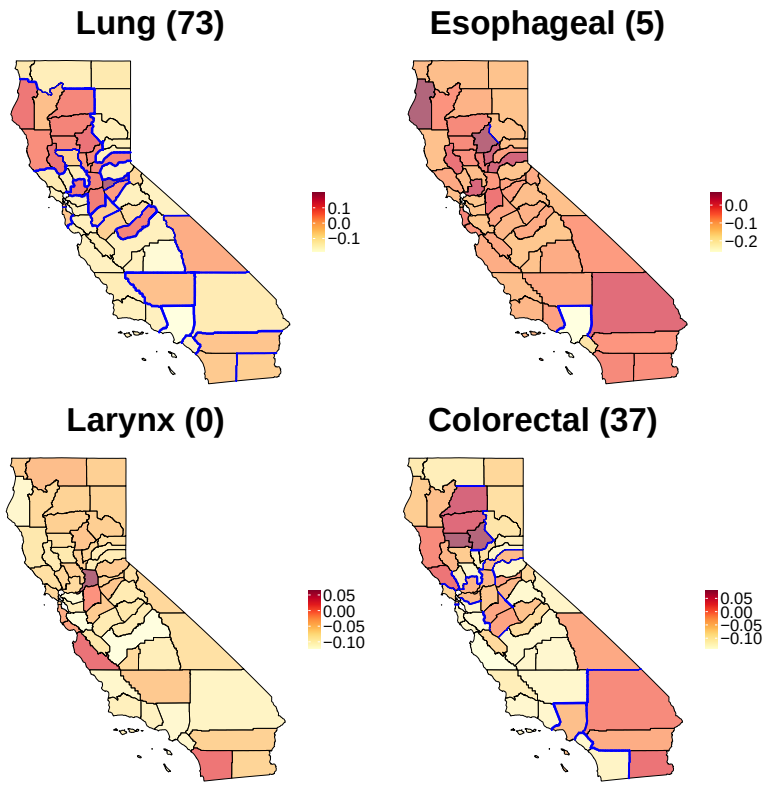


Figure B.22: Difference boundaries (highlighted in red) detected by the model in the California map, colored according to the posterior mean of the corresponding ϕ_{id}, for four cancers individually when $\zeta = 0.05$. The values in brackets are the number of difference boundaries detected.

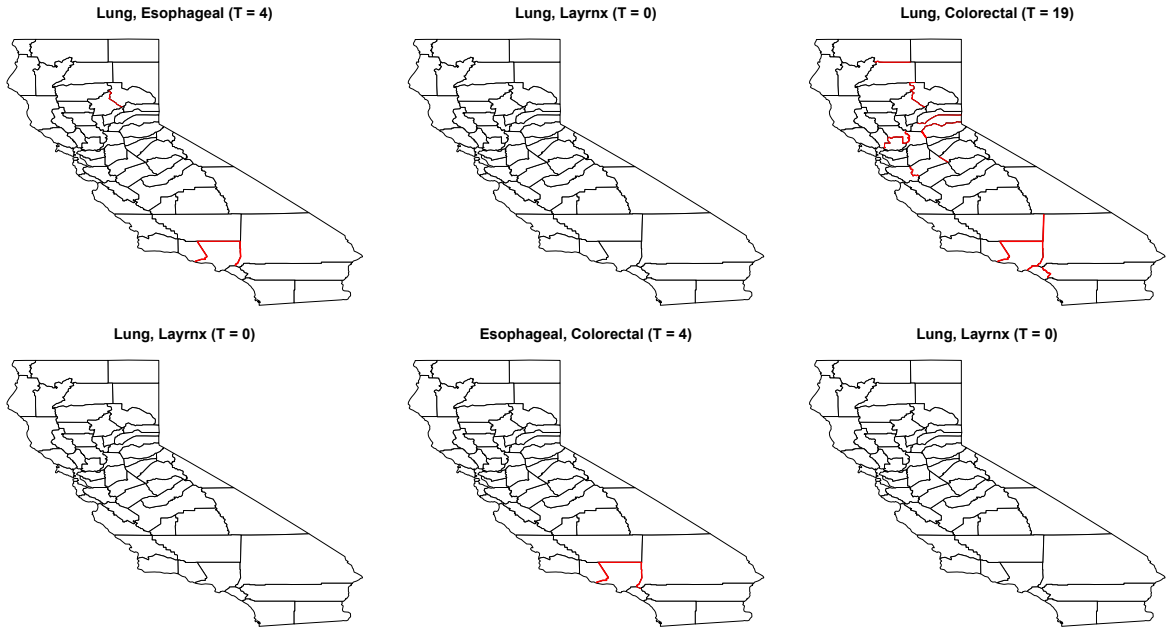


Figure B.23: Shared difference boundaries (highlighted in red) detected by the model for each pair of cancers in California map when $\zeta = 0.05$. The values in brackets are the number of difference boundaries detected.

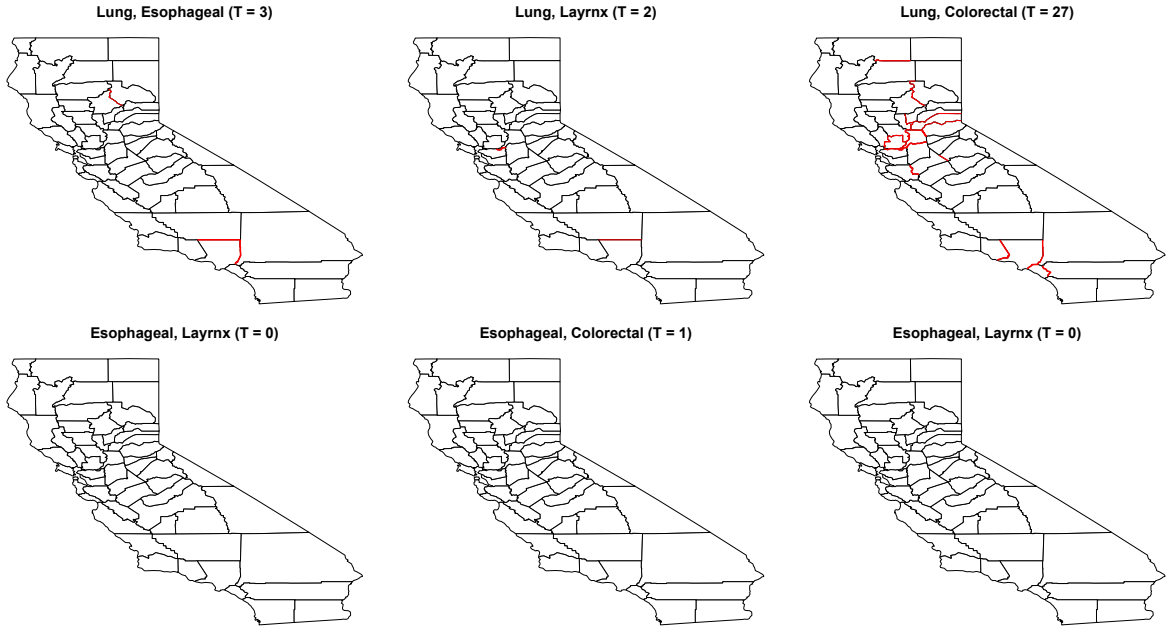


Figure B.24: Mutual cross-difference boundaries (highlighted in red) detected by the model for each pair of cancers in California map when $\zeta = 0.05$. The values in brackets are the number of difference boundaries detected.

Appendix C

Chapter 3 supplementary materials

C.1 Additional results of the simulation study

Figures C.1 and C.2 show a comparison of the accuracy of the epicentre estimation performed by EQN and SEQM when the real epicentre is located within and outside the network, respectively. In particular, the figures display the distribution of the geodetic distance (error) between the estimated and simulated epicentre for EQN and SEQM for different values of n under all scenarios. Overall, SEQM is characterized by a lower error, thus improving EQN estimates. In particular, the error decreases as the number of active smartphones increases, with small differences for varying EQN thresholds. Rather, when the epicentre is located far from the network, the improvement of SEQM over EQN is much more evident.

In Tables C.1 and C.2, we present the median point estimates and 95% interval for each parameter over the 100 simulation datasets, where the point estimates are computed as the highest posterior modes. Notably, for the earthquake parameters, the true value consistently falls within the 95% interval. Since the true origin time differs over the 100 simulated datasets, we present the time origin error in Tables C.1 and C.2, i.e., the difference between the true and the estimated origin time. The 95% interval of such error always includes zero when the epicentre is located outside the network while it remains close to zero for the scenarios with the epicentre located inside the network. The median estimate approaches 1.76; this is not surprising since the triggering times are simulated from uniform distributions that induce, on average, a delay of 1.75 seconds with respect to the arrival of the P- or S-wave. For parameters α and π, the true values are always within the 95% interval, albeit with relatively large interval lengths. Similar conclusions can be drawn for the seismic velocity parameter v_P when the epicentre is distant from the network. Estimating the cure fraction is challenging when the plateau phase has not yet been reached, e.g., in the presence of administrative censoring, like in our setting (Peng and Taylor, 2014). Hence, the high posterior value of π is likely due to the fact that the cure fraction encodes both the faulty smartphones and the late triggers (censored uncured smartphones), and these two possibilities cannot be, eventually, distinguished. However, the proportion of faulty smartphones is not relevant in studying earthquake dynamics. Finally, for parameter λ^{-1}, the 95% intervals always include the true value, with increasing precision as n and r increase.

To clarify the uncertainty quantification of the approach, Tables C.3 and C.5 present the empirical coverages for the parameters across all the simulation scenarios, i.e., the percentage of times that

each parameter belongs to its 95% HPDRs over the 100 simulated datasets. The average lengths of the 95% HPDRs are reported in Tables C.4 and C.6. The empirical coverages of the earthquake parameters are close to the nominal level of 95%, showing that the SEQM is able to recover the true parameter, except for the origin time t_0. Again, triggering times are simulated using uniform distributions, leading to an average delay of 1.75 seconds compared to the arrival of either the P- or S-wave. Likely, the consequence is that the empirical coverage of v_p also tends to decrease with respect to its nominal value as n increases.

Finally, Tables C.7 and C.8 display the average number of posterior modes for each parameter over the 100 generated datasets. Notably, the number of modes when the real epicentre lies within the network is somehow greater compared to the alternative scenario.

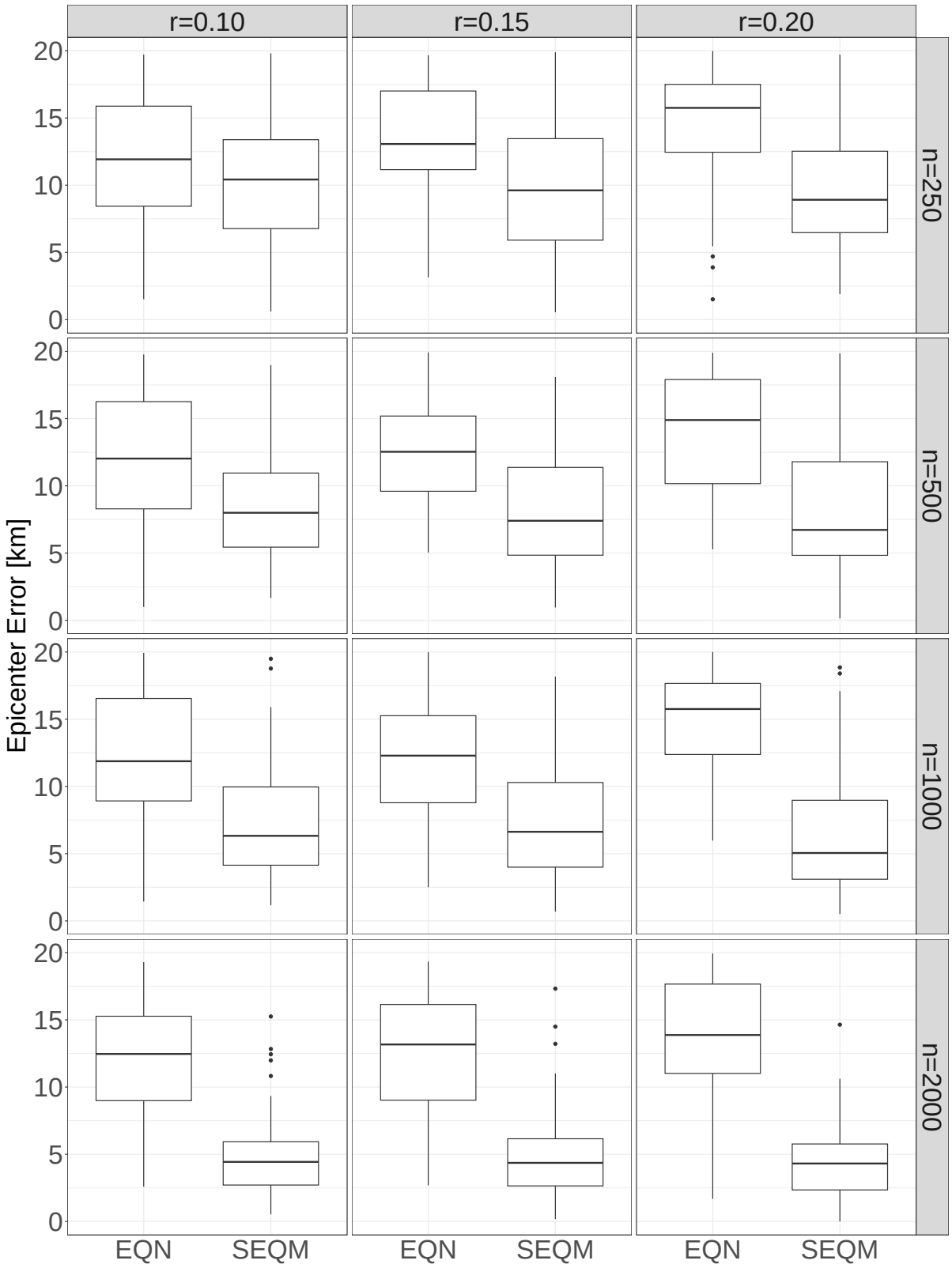


Figure C.1: Boxplots of the epicentre estimation error for EQN and SEQM in all the simulation scenarios in the case of the real epicentre located within the network.

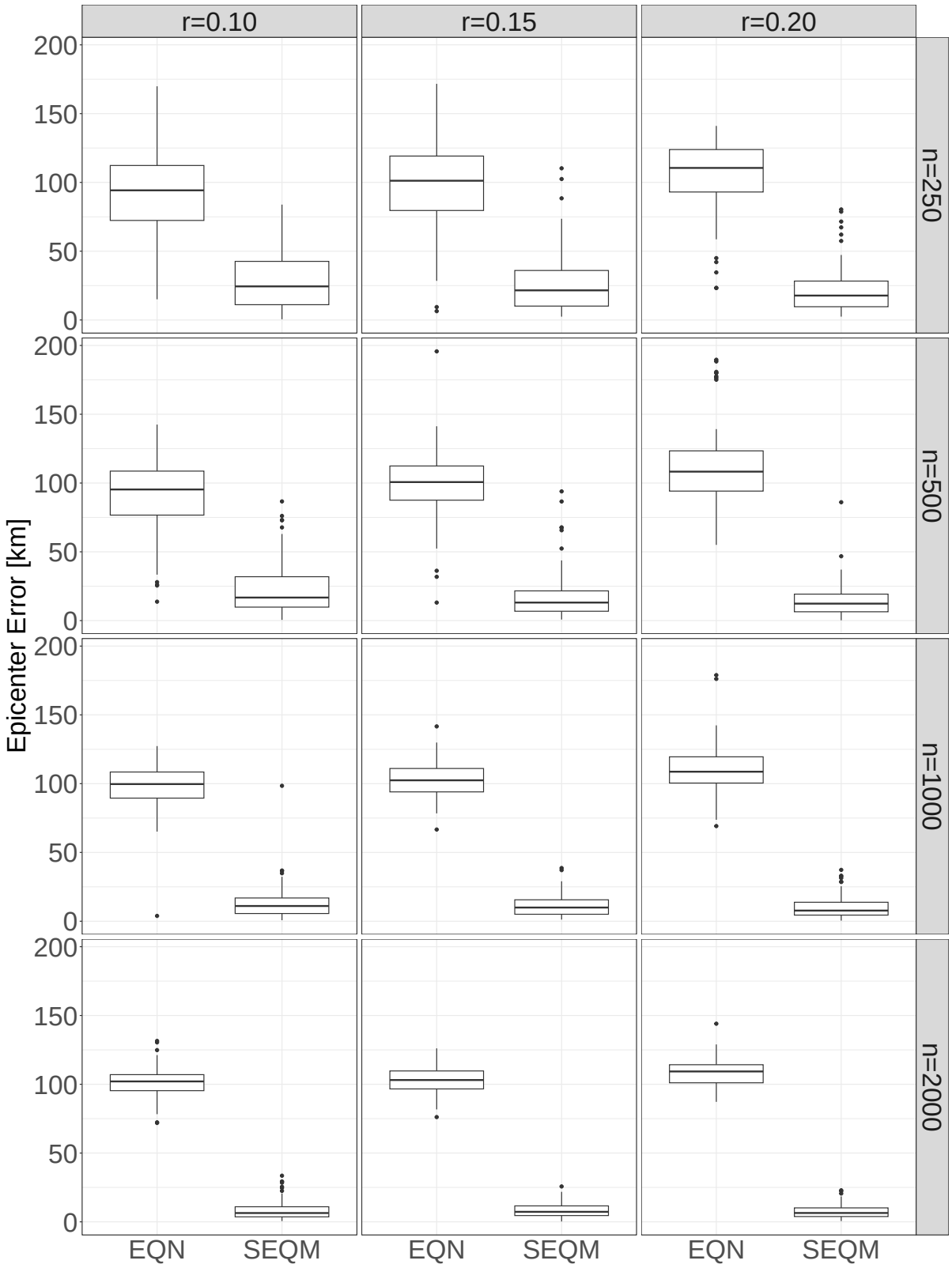


Figure C.2: Boxplots of the epicentre estimation error for EQN and SEQM in all the simulation scenarios in the case of the real epicentre located outside the network.

Table C.1: Median and 95% interval of the estimated parameters over the 100 simulated datasets for all scenarios with the epicentre inside the network. Bold values on the left are the true quantities used in the data-generating process.

	r=0.10				r=0.15				r=0.20			
	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000
lat_0	41.11	41.10	41.10	41.12	41.11	41.10	41.10	41.12	41.11	41.13	41.11	41.12
41.12	[40.79, 41.47]	[40.87, 41.33]	[40.93, 41.20]	[41.05, 41.19]	[40.82, 41.85]	[40.92, 41.27]	[40.86, 41.23]	[41.02, 41.17]	[40.84, 41.51]	[40.94, 41.27]	[40.93, 41.21]	[41.05, 41.18]
lon_0	28.59	28.59	28.58	28.57	28.58	28.57	28.57	28.57	28.58	28.57	28.56	28.56
28.55	[28.42, 28.75]	[28.38, 28.72]	[28.44, 28.69]	[28.51, 28.65]	[28.17, 28.76]	[28.46, 28.70]	[28.48, 28.68]	[28.50, 28.65]	[28.42, 28.73]	[28.47, 28.69]	[28.47, 28.65]	[28.51, 28.62]
d_0	13.04	14.49	15.76	17.72	13.94	15.06	16.46	18.81	13.12	14.44	16.38	15.96
15.00	[9.61, 17.45]	[9.74, 23.61]	[8.58, 25.96]	[10.60, 35.50]	[9.72, 20.08]	[9.32, 21.85]	[10.14, 27.01]	[10.22, 33.33]	[10.10, 19.72]	[9.48, 25.02]	[10.30, 26.68]	[9.09, 29.74]
$t_0 - \hat{t}_0$	1.48	1.46	1.56	1.76	1.57	1.48	1.45	1.60	1.64	1.32	1.30	1.23
	[-0.66, 4.60]	[-0.27, 4.11]	[0.06, 2.92]	[0.69, 3.02]	[-0.54, 11.15]	[0.01, 2.97]	[-0.40, 2.93]	[0.40, 2.86]	[-0.37, 6.43]	[-0.05,2.72]	[0.13,2.76]	[0.30,2.69]
α	0.96	0.96	0.96	0.55	0.96	0.96	0.96	0.37	0.96	0.96	0.96	0.61
0.80	[0.03, 0.97]	[0.03, 0.97]	[0.03, 0.97]	[0.11, 0.98]	[0.03, 0.97]	[0.12, 0.98]	[0.17, 0.98]	[0.14, 0.98]	[0.03, 0.99]	[0.22, 0.99]	[0.30, 0.99]	[0.20, 0.99]
π	0.46	0.44	0.67	0.70	0.63	0.64	0.67	0.65	0.70	0.70	0.70	0.70
0.30	[0.07, 0.82]	[0.06, 0.81]	[0.07, 0.86]	[0.06, 0.86]	[0.07, 0.81]	[0.07, 0.79]	[0.05, 0.82]	[0.05, 0.81]	[0.07, 0.79]	[0.08, 0.79]	[0.13, 0.78]	[0.08, 0.78]
v_P	6.36	6.44	6.45	6.62	6.40	6.42	6.51	6.63	6.43	6.51	6.62	6.71
7.80	[6.09, 6.79]	[6.04, 6.93]	[6.05, 7.04]	[6.00, 7.33]	[6.03, 6.82]	[6.07, 6.85]	[6.07, 6.96]	[6.06, 7.33]	[6.03, 7.17]	[6.15, 7.35]	[6.06, 7.65]	[6.00, 7.67]
τ	1.32	1.02	0.90	0.80	1.26	1.17	1.02	0.84	1.22	1.12	1.13	0.99
	[0.41, 2.66]	[0.30, 2.04]	[0.47, 1.91]	[0.44, 1.24]	[0.56, 2.13]	[0.41, 1.83]	[0.59, 1.65]	[0.53, 1.39]	[0.65, 2.39]	[0.66, 1.67]	[0.67, 1.58]	[0.67, 1.34]
λ^{-1}	5061.82	4945.44	4755.45	4838.34	4913.27	5019.23	4552.22	4711.01	5030.40	4639.21	4840.36	4652.11
4800.00	[2377.16, 16641.22]	[2764.15, 14521.27]	[3232.28, 7179.21]	[3756.53, 6267.75]	[2504.67, 26242.66]	[3103.49, 9582.80]	[3296.61, 6844.14]	[3664.52, 6163.91]	[2542.97, 15693.00]	[2990.68, 8158.50]	[3350.32, 7329.16]	[3628.87, 6348.10]
seconds/iteration	0.0153	0.0264	0.0462	0.0820	0.0158	0.0271	0.0484	0.0833	0.0163	0.0288	0.0509	0.0898

Table C.2: Median and 95% interval of the estimated parameters over the 100 simulated datasets for all scenarios with the epicentre outside the network. Bold values on the left are the true quantities used in the data-generating process.

	r=0.10				r=0.15				r=0.20			
	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000
lat_0	41.59	41.65	41.69	41.71	41.67	41.69	41.71	41.70	41.68	41.70	41.72	41.72
41.70	[41.16, 42.13]	[41.17, 41.99]	[41.41, 41.89]	[41.51, 41.91]	[41.06, 42.13]	[41.21, 42.02]	[41.52, 41.94]	[41.56, 41.84]	[41.18, 42.02]	[41.45, 41.95]	[41.48, 41.92]	[41.60, 41.87]
lon_0	29.88	29.87	29.89	29.91	29.87	29.90	29.89	29.90	29.87	29.88	29.90	29.91
29.90	[29.42, 30.11]	[29.39, 30.04]	[29.73, 29.99]	[29.79, 29.98]	[29.31, 30.02]	[29.57, 30.02]	[29.76, 30.00]	[29.81, 29.97]	[29.45, 30.03]	[29.69, 30.01]	[29.73, 29.99]	[29.84, 29.96]
d_0	12.78	12.83	13.02	13.72	12.73	13.18	13.15	13.27	12.92	12.80	13.11	13.39
15.00	[11.01, 15.02]	[10.92, 16.99]	[10.95, 15.54]	[11.28, 17.80]	[11.16, 15.32]	[11.32, 15.81]	[11.23, 15.45]	[11.51, 15.62]	[11.17, 16.08]	[10.96, 15.80]	[11.09, 16.06]	[11.37, 17.42]
$t_0 - \hat{t}_0$	1.26	1.80	1.83	1.39	1.70	2.17	2.01	1.58	2.01	1.90	1.79	1.70
	[-7.25, 9.44]	[-7.56, 7.27]	[-2.84, 6.25]	[-1.50, 4.53]	[-7.57, 8.00]	[-6.56, 6.47]	[-1.37, 6.08]	[-1.40, 4.26]	[-7.84, 7.19]	[-2.20, 7.23]	[-2.35, 5.39]	[-0.59, 4.95]
α	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
0.80	[0.12, 0.98]	[0.08, 0.97]	[0.18, 0.98]	[0.19, 0.98]	[0.23, 0.98]	[0.23, 0.98]	[0.22, 0.97]	[0.23, 0.97]	[0.23, 0.98]	[0.25, 0.97]	[0.26, 0.97]	[0.25, 0.97]
π	0.64	0.62	0.68	0.70	0.66	0.69	0.73	0.70	0.71	0.71	0.74	0.74
0.30	[0.08, 0.79]	[0.09, 0.82]	[0.08, 0.85]	[0.16, 0.85]	[0.08, 0.79]	[0.29, 0.80]	[0.40, 0.81]	[0.11, 0.82]	[0.07, 0.79]	[0.16, 0.79]	[0.13, 0.79]	[0.25, 0.80]
v_P	6.52	6.54	6.73	7.41	6.39	6.87	6.84	6.98	6.67	6.82	7.22	7.26
7.80	[6.12, 8.12]	[5.99, 8.23]	[6.19, 8.33]	[6.46, 8.44]	[5.95, 8.18]	[6.13, 8.33]	[6.38, 8.47]	[6.64, 8.63]	[6.20, 8.28]	[6.34, 8.47]	[6.48, 8.44]	[6.73, 8.55]
τ	1.20	1.03	0.97	0.85	1.21	1.16	1.12	1.08	1.20	1.15	1.12	1.14
	[0.63, 2.84]	[0.57, 1.87]	[0.66, 1.54]	[0.63, 1.16]	[0.75, 2.30]	[0.73, 1.63]	[0.79, 1.43]	[0.89, 1.34]	[0.85, 2.63]	[0.95, 1.65]	[0.96, 1.45]	[0.94, 1.41]
λ^{-1}	4974.19	4875.66	4775.89	4819.85	4861.18	5054.31	4682.67	4916.37	4885.66	4650.62	4975.41	4723.76
4800.00	[2437.38, 16633.83]	[2882.09, 12616.28]	[3310.62, 6921.19]	[3709.72, 6293.17]	[2323.98, 30033.98]	[2838.54, 8228.65]	[3122.52, 7149.67]	[3780.09, 5904.47]	[2727.43, 15875.64]	[3172.08, 8127.89]	[3304.38, 7351.58]	[3618.18, 6341.63]
seconds/iteration	0.0148	0.0246	0.0450	0.0784	0.0152	0.0250	0.0485	0.0871	0.0157	0.0279	0.0492	0.0883

Table C.3: Empirical coverage of the true parameter being in the 95% HPDR over the 100 datasets, for all simulation scenarios with the epicentre located within the network.

	r=0.10				r=0.15				r=0.20			
	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000
lat_0	0.93	0.91	0.92	0.96	0.94	0.94	0.90	0.93	0.96	0.95	0.89	0.96
lon_0	0.98	0.90	0.93	0.92	0.96	0.96	0.91	0.87	0.98	0.96	0.93	0.91
d_0	1.00	0.98	1.00	0.90	1.00	1.00	1.00	0.87	1.00	1.00	1.00	0.90
t_0	0.98	0.93	0.82	0.57	0.94	0.90	0.84	0.51	0.94	0.83	0.76	0.56
α	0.99	0.97	0.96	0.85	0.98	0.98	0.96	0.68	0.98	0.95	0.91	0.75
π	1.00	1.00	0.90	0.87	0.87	0.86	0.75	0.84	0.62	0.51	0.44	0.38
v_P	0.84	0.82	0.79	0.73	0.83	0.79	0.68	0.59	0.75	0.76	0.77	0.69
λ^{-1}	0.96	0.93	0.97	0.98	0.94	0.99	0.97	0.95	0.96	0.99	0.97	0.94

Table C.4: Average 95% HPDR length for each parameter over the 100 datasets, for each simulation scenario in the case of the real epicentre located within the network.

	r=0.10				r=0.15				r=0.20			
	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000
lat_0	0.78	0.40	0.27	0.16	0.78	0.41	0.25	0.14	0.67	0.34	0.24	0.21
lon_0	0.50	0.29	0.21	0.14	0.47	0.28	0.19	0.13	0.43	0.23	0.16	0.17
d_0	28.18	27.83	26.78	26.23	28.92	27.57	26.74	20.82	28.07	27.41	26.38	22.51
t_0	12.26	7.04	5.71	4.60	11.64	6.60	5.12	3.61	9.91	5.34	4.43	5.06
α	1.01	0.94	0.91	0.80	0.92	0.89	0.81	0.69	0.82	0.69	0.63	0.56
π	0.77	0.76	0.74	0.69	0.69	0.67	0.61	0.61	0.55	0.45	0.41	0.37
v_P	2.95	2.90	2.85	2.75	2.86	2.82	2.75	2.53	2.77	2.66	2.56	2.41
τ	2.25	1.72	1.35	0.82	2.00	1.66	1.17	0.77	1.77	1.17	0.84	0.60
λ^{-1}	21152.38	10417.93	4421.46	2938.45	22252.14	8687.46	4271.39	2880.40	20406.07	7917.56	4685.36	2852.58

Table C.5: Empirical coverage of the true parameter being in the 95% HPDR over the 100 datasets, for all simulation scenarios with the epicentre located outside the network.

	r=0.10				r=0.15				r=0.20			
	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000
lat_0	0.94	0.93	0.91	0.91	0.98	0.93	0.96	0.96	0.94	0.96	0.92	0.93
lon_0	0.92	0.88	0.97	0.89	0.92	0.95	0.97	0.96	0.93	0.92	0.92	0.92
d_0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t_0	0.95	0.95	0.86	0.85	0.95	0.95	0.91	0.91	0.95	0.83	0.85	0.83
α	1.00	1.00	1.00	0.98	1.00	0.97	0.99	0.95	0.99	0.96	0.96	0.95
π	1.00	1.00	1.00	0.99	0.99	1.00	0.99	0.97	0.97	0.92	0.92	0.84
v_P	0.90	0.75	0.71	0.90	0.74	0.83	0.83	0.84	0.85	0.75	0.84	0.78
λ^{-1}	0.94	0.96	0.97	0.96	0.93	0.97	0.95	0.98	0.98	0.99	0.97	0.92

Table C.6: Average 95% HPDR length for each parameter over the 100 datasets, for each simulation scenario in the case of the real epicentre outside the network.

	r=0.10				r=0.15				r=0.20			
	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000
lat_0	1.01	0.73	0.47	0.31	0.97	0.66	0.43	0.30	0.89	0.54	0.37	0.28
lon_0	0.61	0.45	0.26	0.17	0.54	0.38	0.23	0.16	0.51	0.28	0.19	0.14
d_0	30.09	30.34	31.37	33.06	30.35	30.56	31.67	32.63	30.69	30.48	31.84	37.36
t_0	19.75	14.83	9.88	6.60	18.70	13.33	9.09	6.73	17.19	11.02	8.00	7.53
α	0.87	0.84	0.80	0.80	0.84	0.80	0.79	0.76	0.82	0.75	0.75	0.75
π	0.78	0.78	0.76	0.75	0.76	0.75	0.74	0.71	0.74	0.69	0.68	0.64
v_P	2.84	2.75	2.56	2.29	2.79	2.65	2.41	2.13	2.67	2.32	2.11	1.92
τ	1.92	1.36	0.85	0.49	1.65	1.18	0.70	0.48	1.31	0.83	0.51	0.36
λ^{-1}	21103.66	10978.15	4549.48	3015.55	21738.94	8803.28	4323.14	2943.74	20731.20	8048.77	4768.53	2924.67

Table C.7: Average number of modes for each parameter over the 100 generated datasets with the real epicentre located within the network.

	r=0.10				r=0.15				r=0.20			
	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000
lat_0	1.21	1.09	1.07	1.00	1.19	1.06	1.02	1.00	1.22	1.03	1.02	1.28
lon_0	1.14	1.06	1.04	1.03	1.13	1.03	1.01	1.03	1.19	1.02	1.00	1.23
d_0	1.04	1.05	1.01	1.06	1.00	1.00	1.00	1.08	1.00	1.01	1.03	1.22
t_0	1.32	1.06	1.12	1.08	1.22	1.06	1.03	1.14	1.26	1.08	1.08	1.36
α	1.57	1.39	1.38	1.39	1.51	1.36	1.32	1.52	1.48	1.25	1.24	1.48
π	1.02	1.02	1.19	1.18	1.06	1.10	1.17	1.12	1.20	1.15	1.25	1.48
v_P	1.00	1.00	1.00	1.01	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.04
τ	1.06	1.07	1.05	1.20	1.03	1.04	1.06	1.17	1.05	1.03	1.02	1.04
λ^{-1}	1.10	1.03	1.00	1.00	1.09	1.01	1.01	1.00	1.03	1.00	1.00	1.00

Table C.8: Average number of modes for each parameter over the 100 generated datasets with the real epicentre located outside the network.

	r=0.10				r=0.15				r=0.20			
	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000	n=250	n=500	n=1000	n=2000
lat_0	1.03	1.05	1.05	1.02	1.07	1.06	1.00	1.01	1.04	1.04	1.01	1.03
lon_0	1.10	1.08	1.05	1.01	1.05	1.08	1.00	1.01	1.03	1.01	1.00	1.03
d_0	1.00	1.01	1.01	1.04	1.01	1.01	1.02	1.02	1.02	1.00	1.02	1.10
t_0	1.09	1.05	1.07	1.00	1.06	1.03	1.01	1.01	1.15	1.02	1.03	1.08
α	1.22	1.17	1.13	1.11	1.26	1.12	1.05	1.08	1.19	1.08	1.06	1.06
π	1.00	1.02	1.05	1.05	1.02	1.02	1.03	1.07	1.02	1.01	1.07	1.13
v_P	1.01	1.01	1.00	1.04	1.00	1.01	1.01	1.03	1.01	1.01	1.04	1.07
τ	1.00	1.04	1.01	1.00	1.03	1.00	1.00	1.00	1.06	1.00	1.00	1.00
λ^{-1}	1.06	1.02	1.00	1.00	1.06	1.01	1.00	1.00	1.06	1.00	1.00	1.00

C.2 Additional empirical results

Figures C.3, C.4 and C.5 present diagnostic plots for the real data analysis. In each case, there are no apparent convergence issues observed in the chains of every parameter. Additionally, the marginal posterior densities reveal that the EMCS earthquake parameters consistently fall within the posterior intervals, with the exception of the latitude parameter in the Mexican event.

In Figure C.6, for each data analysis, both the Kaplan-Meier estimator (Kaplan and Meier, 1958) and the posterior standardized survival function (Brilleman et al., 2020) are illustrated. Specifically, each plot showcases the average posterior survival function across the considered smartphones, accompanied by their 95% credible bands. Notably, following the earthquakes (highlighted in green), the trajectories of all curves closely emulate those of the empirical Kaplan-Meier curves. This indicates that our model captures and describes well the data at hand.

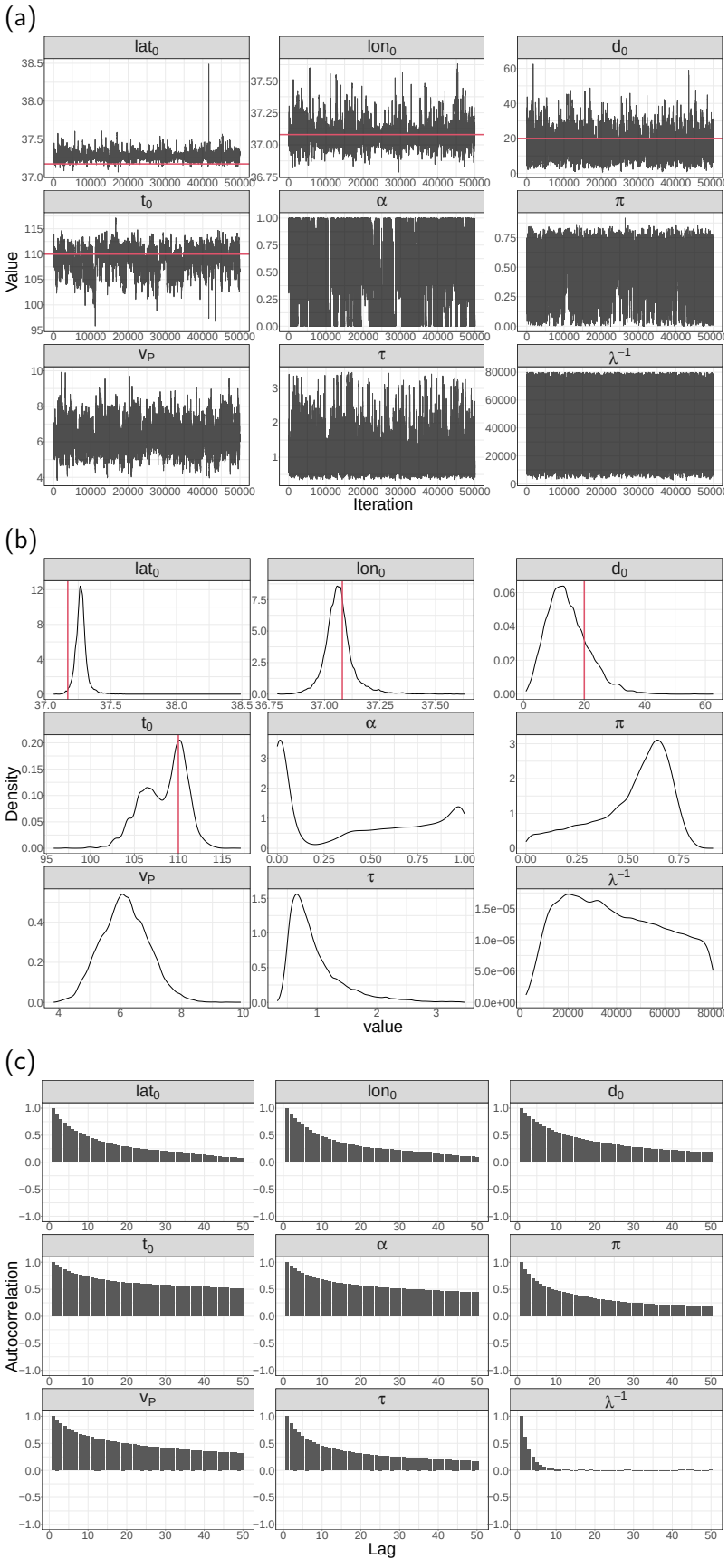


Figure C.3: Diagnostics plot for the Pazarcik case. Traceplots (a), densities (b) and autocorrelations (c). Red lines represent the EMSC earthquake parameters.

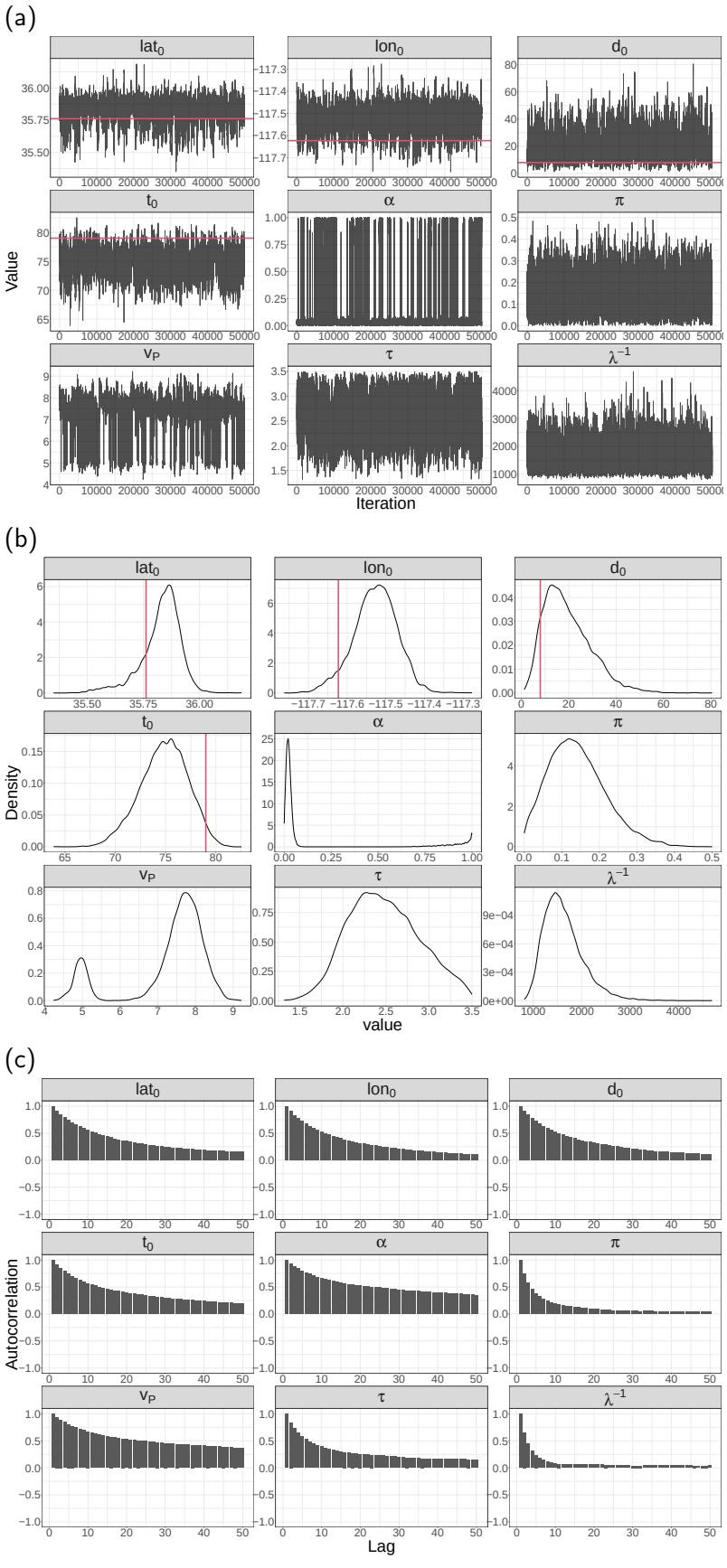


Figure C.4: Diagnostics plot for the Californian case. Traceplots (a), densities (b) and autocorrelations (c). Red lines represent the EMSC earthquake parameters.

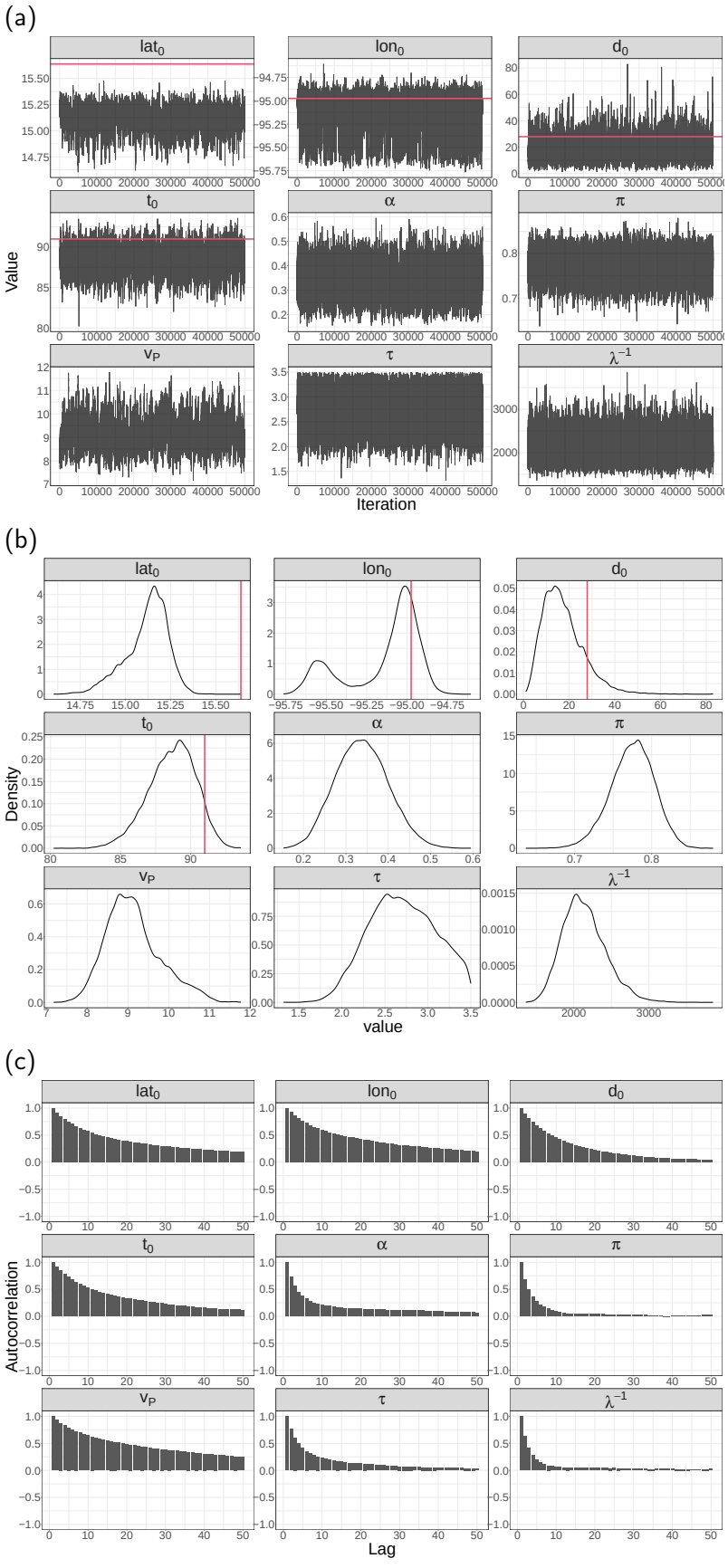


Figure C.5: Diagnostics plot for the Mexican case. Traceplots (a), densities (b) and autocorrelations (c). Red lines represent the EMSC earthquake parameters.

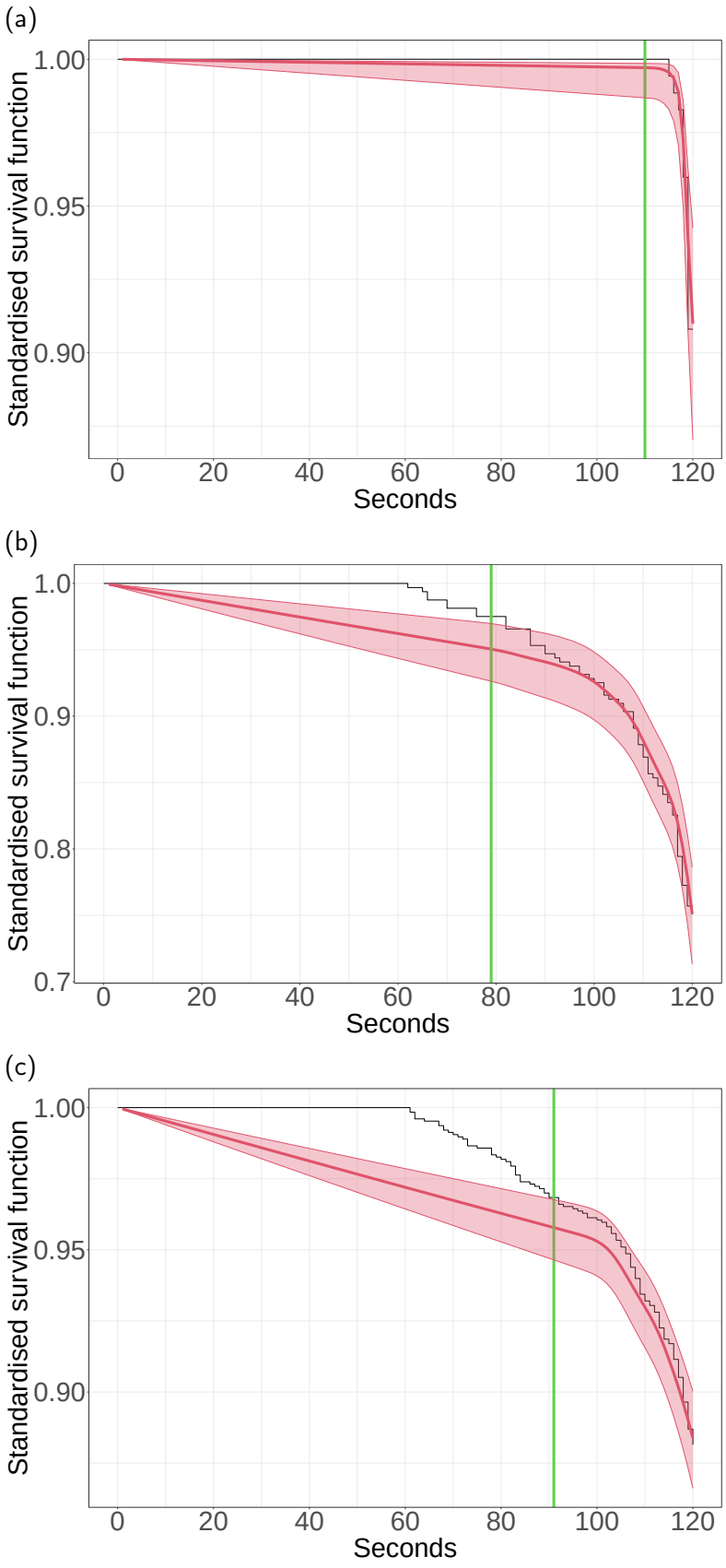


Figure C.6: Comparison between the Kaplan-Meier estimator (black line), the standardized survival function estimated by SEQM (red interval) and the true earthquake origin time (green line) for the three cases: Pazarcik (a), Californian (b) and Mexican (c).