Areal data:

disease mapping and small area estimation

Victoire Michal

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University
Montréal, Québec
April 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Dedication

À Mamirène. To Ervin Keith James.

Acknowledgements

I am deeply grateful to my supervisor, Alexandra M. Schmidt, for these past five years. Alex, thank you for your guidance, your constant availability, your contagious passion, your encouragements, your wisdom, your laughs, and your (many) calming messages. I cannot imagine completing a PhD with anyone else, you are an incredible human.

I am also grateful to my co-supervisor, Jon Wakefield, for the numerous enlightening discussions, the availability and valuable guidance he provided even from afar, the (hilarious) jokes, the infectuous curiosity, and the confidence he repeatedly shared.

From the department of Epidemiology, Biostatistics and Occupational Health (EBOH), I must acknowledge Katherine Hayden, Dolores Coleto, and André Yves Gagnon. I am sure we can count in the millions the number of questions I asked you throughout the years. Thank you for facilitating all my administrative tasks. From EBOH, I am also thankful to my committee member, Jill Baumgartner, for kindly sharing part of the data analysed in this thesis.

I am appreciative of the financial support I received from the *Fonds de recherche du Québec – Nature et technologies*, the Natural Sciences and Engineering Research Council of Canada, the *Institut de Valorisation des Données*, and from my supervisor.

Je remercie ensuite la présidente et le vice-président de mon fan club: ma maman (Mong) et Nanou. Ce troupeau de deux personnes me pousse continuellement à me dépasser et m'apporte depuis toujours un soutien financier et émotionnel sans faille, et ce, même à l'international! Merci d'avoir anticipé et de m'avoir envoyée toute seule en Angleterre à 9 ans, cette thèse aurait été beaucoup moins lisible autrement.

J'aimerais maintenant avoir un petit mot pour tous les copains qui ont rendu ma vie plus gaie ces dernières années. D'abord, je pense à ceux aux côtés de qui j'ai vécu cette expérience, soient les McGillois: Janie, Daniel, Armando, Julien, Renaud, Vanessa, Marc P., et Niki.

Ensuite, je pense aux UdeM-iens qui me sont restés chers: Gab, Isabelle, Alexis, et Véro. Je suis aussi infiniment reconnaissante envers tous ceux qui m'ont sorti la tête du guidon (la thèse): Justine, Andréanne, Cam, Julien R, Stan, et Davi, au Québec, puis Mallau, Choup's, Kik, Michou, Ben, et Elo, en France.

J'ai gardé le meilleur pour la fin: Marc-Antoine Nicolas Savard, mon humain préféré. Merci de ton soutien sans relâche ces cinq (9?) dernières années, merci de rendre ma vie tous les jours un peu plus rigolote et musicale, merci de toujours faire attention à ce que j'ai à boire et à manger, merci de me supporter, merci de m'aimer.

Preface

This thesis focuses on the analysis of areal data in three settings, namely a purely spatial and a spatio-temporal disease mapping contexts, and small area estimation. This manuscript-based thesis contains six chapters: Chapter 1 briefly introduces the concepts studied, and Chapter 2 provides a more detailed literature review, Chapters 3–5 contain three original manuscripts that are linked by the topic of areal data analysis, and Chapter 6 concludes this thesis with a discussion of future work avenues. The references and appendices of each manuscript are all available after the concluding Chapter 6.

Chapters 1, 2, and 6 were conceived and written by Victoire Michal (VM) and revised by Alexandra M. Schmidt (AMS).

Chapter 3 was suggested by AMS and further conceptualised by VM. VM developed the methodology, designed and conducted the simulation studies and the data analysis, and wrote the manuscript. Substantial feedback came from AMS at every step of the process, in particular, troubleshooting the simulation studies, and revising the manuscript. Regarding the data analysis, Laís Picinini Freitas helped to obtain the data, to introduce the application in the manuscript, and to provide insights on the results. Oswaldo Gonçalves Cruz provided further valuable insights regarding the data application.

Chapter 4 was conceptualised via discussions between VM and AMS. The methodology development, design of the simulation studies, data analyses, and manuscript writing were conducted by VM. AMS provided valuable feedback, guidance, and revision at every step of the process.

Chapter 5 was conceptualised via discussions between VM, Jon Wakefield (JW), and AMS. VM developed the methodology, designed and conducted the simulation studies and the data analysis, and wrote the manuscript. JW and AMS provided valuable guidance at every step of the process. The Ghanaian data application arose from discussions with Alicia Cavanaugh, Brian E. Robinson, and Jill Baumgartner.

Abstract

This thesis focuses on the analysis of areal data, where an outcome is observed across different areas of a region. Areal data commonly arise in disease mapping or small area estimation (SAE). We aim to provide flexible spatial and spatio-temporal models in the analysis of disease mapping data. Furthermore, we investigate different methods for SAE, including machine learning (ML) approaches.

When the number of cases of a disease is recorded across different areas within a region, disease mapping is useful to estimate the areal relative risk. The number of cases in an area is often assumed to follow a Poisson distribution whose log risk may be written as the sum of fixed and random effects. The BYM2 model decomposes each latent effect into a weighted sum of independent and spatial effects. In the first manuscript, we extend the BYM2 model to allow for heavy-tailed latent effects and accommodate potentially outlying risks, after accounting for the fixed effects. We assume a scale mixture wherein the variance of the latent process changes across areas and allows for outlier identification. We explore two prior specifications of the scaling parameters and compare the proposed model to another proposal in the literature, in simulation studies and in the analysis of Zika cases from the 2015-2016 epidemic in Rio de Janeiro.

Further, disease counts are increasingly recorded over time and across areas, and spatio-temporal disease mapping models help understand the spread of the disease over time. Commonly, the areal number of cases is assumed to follow a Poisson distribution, where the log risk varies with space and time. Models have been proposed to account for a spatio-temporal trend in the latent effects. In the second manuscript, we extend a spatio-temporal model to allow for heavy-tailed effects to accommodate and identify outliers. At each time point, we assume the latent effects to be spatially structured and include scaling parameters in the precision matrix to allow for heavy tails. We investigate the performance of the proposed model through simulation studies and analyse the weekly evolution of COVID-19

cases across Montreal and France during the second wave.

When an outcome is measured across a fraction of the areas of a region through a survey that samples few units per area, SAE methods are useful to obtain reliable estimates at the areal level. In the third manuscript, we propose a comparison of different approaches for model-based small area prediction when there are abundant auxiliary data for the sampled and non-sampled areas. Random forest (RF) and LASSO approaches are compared with a frequentist forward selection procedure and a Bayesian shrinkage method. To provide uncertainty quantification of estimates obtained from RF and LASSO methods, we propose a modification of the split conformal (SC) procedure that relaxes the assumption of exchangeable data. Through simulation studies, we assess the performance of the proposed SC procedure and compare the four modelling approaches. Further, we estimate the areal mean household log consumption in the Greater Accra Metropolitan Area using data available from the sixth Ghanaian Living Standard Survey (GLSS) and the 2010 Population and Housing Census. The dependent variable is measured only in the GLSS for 3% of all the areas, and 174 covariates are available from both datasets. For this analysis, a cross-validation study showed that the Bayesian shrinkage method yielded smaller bias and MSE.

The methods proposed in the three manuscripts of this thesis contribute to the literature on disease mapping, SAE, and ML. The first two add to the disease mapping literature by accommodating and identifying outlying areas in spatial and spatio-temporal models. The third manuscript contributes to the SAE literature by studying model-based approaches in a high-dimensional setting, and to the ML literature by proposing a procedure to provide uncertainty quantification of ML estimates.

Abrégé

Cette thèse traite de l'analyse de données régionales, où une variable est observée dans différentes régions d'un territoire. Les données régionales surviennent en cartographie des maladies (CM) ou lors d'estimation pour petits domaines (EPD). Notre but est de développer des modèles spatiaux et spatio-temporels flexibles en CM. De plus, nous étudions diverses méthodes en EPD, dont des méthodes d'apprentissage automatique (AA).

Quand le nombre de cas d'une maladie est observé dans diverses régions, la CM sert à estimer le risque relatif régional. On suppose souvent que le nombre de cas régional suit une loi Poisson dont le risque log est la somme d'effets fixes et latents. Le modèle BYM2 définit chaque effet latent en la somme pondérée d'effets indépendants et spatiaux. Dans le premier manuscrit, nous modifions le modèle BYM2 pour inclure des effets à queue lourde et s'adapter aux risques aberrants, après prise en compte des effets fixes. Nous supposons un mélange d'échelles où la variance latente change selon les régions et permet d'identifier les valeurs aberrantes. Nous envisageons deux lois *a priori* pour les paramètres d'échelle et comparons le modèle proposé à un autre via des études par simulation et dans l'analyse de l'épidémie de Zika de 2015-2016 à Rio de Janeiro.

En outre, l'enregistrement récurrent du nombre régional de cas malades est de plus en plus courant et les modèles spatio-temporels de CM aident à comprendre la propagation de la maladie. On suppose que le nombre de cas régional suit une loi Poisson dont le risque log varie dans l'espace et le temps. Des modèles d'effets latents à tendance spatio-temporelle ont été proposés dans la littérature. Dans le deuxième manuscrit, nous élargissons un modèle spatio-temporel existant afin d'inclure des effets à queue lourde pour s'adapter aux régions aberrantes et les identifier. À chaque point dans le temps, nous supposons que les effets sont structurés spatialement et incluons des paramètres d'échelle dans la matrice de précision pour permettre des queues lourdes. Nous évaluons le modèle proposé via des études par simulation et analysons l'évolution hebdomadaire de la COVID-19 durant la deuxième vague, à Montréal

et en France.

Quand une variable est observée dans peu de régions via un sondage sélectionnant peu d'unités par région, l'EPD mène à des estimations régionales fiables. Dans le troisième manuscrit, nous comparons différentes méthodes basées sur le modèle pour la prédiction en EPD en présence de nombreuses covariables. On compare des approches par forêt aléatoire (FA) et le LASSO à une sélection ascendante fréquentiste et à une méthode de contraction bayésienne. Pour mesurer l'incertitude des estimations par FA et le LASSO, nous proposons de modifier la procédure de prédiction conforme scindée (PCS) afin d'assouplir l'hypothèse d'échangeabilité des données. Nous évaluons la procédure PCS proposée via des études par simulation et comparons les quatre approches. De plus, nous estimons la consommation log moyenne des ménages dans la région métropolitaine du Grand Accra avec les données du sixième *Ghanaian Living Standard Survey* (GLSS) et du *Population and Housing Census* de 2010. La variable dépendante est observée uniquement par le GLSS dans 3% des régions et 174 covariables sont disponibles. Une étude par validation croisée démontre que la méthode par contraction bayésienne génère de plus petits biais et EQM.

Les méthodes proposées dans ces trois manuscrits contribuent à la littérature sur la CM, l'EPD et l'AA. Les deux premiers participent à la littérature sur la CM en s'adaptant et identifiant les régions aberrantes, via des modèles spatiaux et spatio-temporels. Le troisième manuscrit contribue à la littérature d'EPD en étudiant des approches basées sur le modèle dans un contexte de haute dimension. Il contribue aussi à la littérature en AA, en proposant une procédure pour mesurer l'incertitude des estimations par AA.

Table of contents

List of Tables

List of Figures

Abbreviations

AIC Akaike information criterion

BIC Bayesian information criterion

CAR conditional autoregressive

CB-MB Corpas-Burgos and Martinez-Beneito

CHSLD centre d'hébergement et de soins de longue durée

COVID-19 coronavirus disease 2019

EA enumeration area

EBLUP empirical best linear predictor

EHPAD établissement d'hébergement pour personnes âgées dépendantes

GAMA Greater Accra Metropolitan Area

GLSS Ghananian Living Standard Survey

i.i.d. independent and identically distributed

IBGE Intituto Braseileiro de Geografia e Estatística

ICAR intrinsic conditional autoregressive

INSEE Institut National de la Statistique et des Études Économiques

INSPQ Institut National de la Santé Publique du Québec

LASSO least absolute shrinkage and selection operator

MCMC Markov Chain Monte Carlo

MSE mean squared error

OOB out-of-bag

PCAR proper conditional autoregressive

SAE small area estimation

SC split conformal

SDG sustainable development goals

SINAN Sistema de Informação de Agravos de Notificação

SMR standardised morbidity ratio

SUS Sistema Único de Saúde

UN United Nations

Chapter 1

Introduction

In this thesis, I focus on the modelling of areal data in two different contexts, namely disease mapping and small area estimation (SAE). In the case of disease mapping, I am interested in the identification of potentially outlying areas, which might be helpful to better understand the spread of a disease and prioritise interventions. In the case of SAE, I investigate different modelling approaches when few areas are sampled whereas numerous covariates are available. Further, I propose a procedure to provide uncertainty quantification of complex estimates (e.g., using the least absolute shrinkage and selection operator (LASSO), or random forests) when data are not exchangeable. This is an important endeavour because the exchangeability assumption is a strong one in the SAE context, and SAE heavily relies on auxiliary information, due to the small areal sample sizes.

In disease mapping, areal data correspond to the number of cases of a disease recorded across the areas of a region of interest. Further, in a spatio-temporal disease mapping context, the disease counts are recorded across areas and over time. Commonly, in both purely spatial and spatio-temporal frameworks, the counts are modelled following a Poisson distribution with a log risk that includes fixed and latent effects. In a purely spatial setting, the random effects are areal components that are assumed to be spatially structured in the

sense that areas that are close to each other adjust similarly, whereas areas that are further apart do not have strong autocorrelation. Various areal models have been proposed in the disease mapping literature to allow for this spatial smoothing. For instance, Besag (1974) proposed the intrinsic conditional autoregressive (ICAR) model, which assumes a purely spatial autocorrelation of the areal effects. However, it can be argued that when the data are not only spatially structured, the ICAR model does not perform well. Hence, proposals such as Besag et al. (1991); Leroux et al. (1999); Riebler et al. (2016) model the areal effects following a combination of spatial and independent structures. In a spatio-temporal setting, the random effects are both spatially and temporally structured. This may be achieved by decomposing the latent effects into the sum of temporal, spatial, and spatio-temporal interaction components (Knorr-Held, 2000; Ugarte et al., 2012), or by only including spatio-temporal interaction terms (Rushworth et al., 2014). In spatio-temporal models, the aim is to borrow strength from neighbouring areas and from the past to smooth the risk surface through time and across space.

However, these proposed purely spatial and spatio-temporal disease mapping models do not accommodate spatial discrepancies, or specifically, outlying areas. Richardson et al. (2004) argue that disease mapping models should perform two essential tasks: to smooth the areal random noises, and to detect and adapt to true heterogeneity. Although there are various proposals in the spatial and spatio-temporal literature that allow for spatial discrepancies (see, e.g., Lawson and Clark (2002); Anderson et al. (2014); Lee and Lawson (2016); Rushworth et al. (2017)), they do not aim to identify potentially outlying areas. On the other hand, in a purely spatial context, Congdon (2017) proposed a scale mixture prior distribution for the latent effects that allows for the identification of potentially outlying relative risks, after accounting for covariates. This thesis extends the work of Congdon (2017) in two different directions. First, I propose a purely spatial disease mapping extension of Congdon's prior that aims to ease interpretation and prior assignment of the model parameters, while identifying potentially outlying areas. Then, the second case is to propose a spatio-temporal

model that aims to identify potentially outlying areas, which, to the best of my knowledge, has not been considered yet.

On the other hand, in the SAE context, areal data arise from a survey whose areal sample sizes are small, and the aim is to produce estimates across all areas of a region of interest. It is worth mentioning that SAE does not only apply to geographical areas and may consist in producing estimates for any domains of a finite population, but in this thesis, I only focus on regions divided into non-overlapping areas. Further, although SAE is part of the survey sampling theory, the design-based estimators are not reliable, and a model-based approach is usually adopted (Tzavidis et al., 2018). Although model-based estimation may be performed at the unit-level or at the area-level (Rao and Molina, 2015), in this thesis, I only focus on areal level models. Finally, in this setting, because of the small areal sample sizes (even zero, in the case of out-of-sample areas), it is common to expand the auxiliary information through exterior sources and produce areal estimates using these covariates.

In SAE, commonly, estimates rely on associations between the outcome and available covariates, particularly when there are many out-of-sample areas (Tzavidis et al., 2018; Erciulescu and Opsomer, 2022). However, there is no consensus on variable selection approaches (Ghosh, 2020). In the frequentist framework, a common variable selection method is the least absolute shrinkage and selection operator (LASSO, Tibshirani (1996, 2011)). However, inference is not straightforward for LASSO estimates, because the distribution of the regression coefficients is not continuous (Dezeure et al., 2015). Further, SAE has more recently been extended to machine learning approaches (e.g., random forests (Krennmair and Schmid, 2022)). Similarly to the LASSO approach, one issue with machine learning methods, or, specifically, random forests, is how to provide uncertainty quantification of the resultant estimates. Procedures have been proposed in the machine learning literature to compute prediction intervals for random forest point estimates (Lei et al., 2018; Zhang et al., 2019). These methods, however, rely on the assumption of exchangeable data, which may not be

the case in a SAE context. This thesis extends the work of Lei et al. (2018) to propose a procedure to compute prediction intervals of complex area-level estimates when data are not exchangeable, in the context of SAE.

This thesis is organised as follows. Chapter 2 provides a more detailed literature review of disease mapping and SAE methods. Then, in Chapter 3, a purely spatial disease mapping model that identifies potentially outlying areas is proposed. This proposal is an alternative to that of Congdon (2017), and their similarities and differences are discussed. The performance of the two prior specifications of the proposed model is evaluated through extensive simulation studies. Finally, the cases of Zika, a vector-borne disease, recorded across the 160 neighbourhoods of Rio de Janeiro during the first epidemic (2015-2016) are analysed to identify potentially outlying areas with respect to the relative risk of Zika.

In Chapter 4, an extension of a spatio-temporal disease mapping model is proposed to accommodate and identify potentially outlying areas. Two prior specifications of the proposed model are considered and evaluated through simulation studies, to assess the performance of the proposed model in the presence of neighbouring and distant outliers. Further, to showcase the ability of the proposed approach to identify potential outlying areas, the model is fitted to weekly COVID-19 cases and hospitalisations across the 33 boroughs of Montreal and the 96 French departments, respectively, during the second wave.

In Chapter 5, four model-based SAE approaches are compared in the case where the number of out-of-sample areas and the number of auxiliary information are high. Further, we propose a procedure to provide uncertainty quantification of complex estimates (e.g., LASSO and random forests), when data are not exchangeable. We prove that the proposed procedure yields prediction intervals of the right coverage rate and confirm this theoretical result through simulation studies. Finally, in the Greater Accra Metropolitan Area (GAMA), the mean household log consumption is estimated at the enumeration area (EA) level using the sixth Ghanaian Living Standard Survey (GLSS), which comprises 3% of all EAs in the

GAMA. To augment the auxiliary information to the entire GAMA, the 2010 Population and Housing Census is used.

Chapters 3 to 5 are stand-alone manuscripts. Chapter 3 is under a second round of revision in the journal *Statistical Methods in Medical Research*. Chapter 4 is under revision for the journal *Spatial Statistics*. Chapter 5 has been accepted for publication in the *Journal of Survey Statistics and Methodology*.

Chapter 2

Literature review

In this chapter, I discuss the theory that the three manuscripts of this thesis build upon. Section 2.1 reviews common disease mapping models that are proposed to analyse areal data while accommodating spatial autocorrelation. In particular, Section 2.1.1 provides an overview of the literature on disease mapping models that allow for spatial discontinuities and relax the amount of smoothing between neighbouring areas. Then, Section 2.1.2 discusses extensions of disease mapping models to the spatio-temporal framework. Section 2.2 introduces small area estimation methods that are useful to analyse areal data that arise from a survey with small areal sample sizes. Finally, Section 2.2.1 reviews methods proposed in the literature when a high-dimensional vector of auxiliary information is available.

Throughout this chapter and this thesis, vectors and matrices are denoted in bold. Further, in this chapter, we assume that a region of interest is partitioned into n non-overlapping areas, which are indexed by $i = 1, \ldots, n$.

2.1 Disease mapping

In spatial statistics, the types of observations are commonly divided into three categories (Cressie, 2015): geostatistical data, spatial point pattern data, and areal data. In the case of geostatistical data, a variable of interest is observed at fixed points in space and the aim is to model the outcome by taking into account its location. In the case of spatial point pattern data, the locations where an event has occurred is the response of interest, and the goal is to estimate how the points are distributed across the region. In these two cases, the locations are assumed to be continuous over a region of interest. On the other hand, the case of areal data, or lattice data, refers to an outcome observed across a lattice, which may be irregular, within a region of interest (e.g., France is divided into 96 departments). In that case, the region (France) is divided into a finite set of disjoint areas (departments) and the aim is to model the variable of interest while accounting for its location as it is expected that neighbouring areas tend to have similar realisations of the process being observed. This thesis focuses on the analysis of data recorded across different areas of a region of interest.

In particular, when the number of cases of a disease is recorded across the different areas that form a region of interest, disease mapping methods are used to reliably estimate the areal relative risk of that disease. The number of cases in an area is commonly assumed to follow a Poisson distribution whose mean is the product of an offset and the relative risk of the disease. The offsets correspond to the expected number of cases, were the disease counts uniformly distributed across the region (Banerjee et al., 2014). The basic estimate of the relative risk is the standardised morbidity ratio (SMR), which is the ratio between the areal count and offset and corresponds to the maximum likelihood estimate in the frequentist framework. However, in areas with small offsets, this estimate is unreliable as the variance is inversely proportional to the offset. Hence, models where the relative risk is decomposed in the log scale as the sum of an overall rate, and fixed and areal random effects have been proposed to borrow strength from neighbouring areas and obtain reliable disease risk

estimates (Wakefield, 2007; Banerjee et al., 2014). The inclusion of random effects also accommodates overdispersion in the Poisson model that would otherwise assume equal mean and variance for each area. Further, with the development of Markov Chain Monte Carlo (MCMC) methods and founding work of Besag et al. (1991), disease mapping methods are commonly incorporated in Bayesian hierarchical models (Banerjee et al., 2014; Lawson, 2018; MacNab, 2022). The disease mapping models developed in Chapters 3 and 4 of this thesis fall under this framework.

Regarding the areal disease risks, one might naturally expect that areas that are close to each other are more correlated than distant areas. Besag (1974) introduced the intrinsic conditional-autoregressive (ICAR) prior, where the spatial autocorrelation between areal effects is accounted for through spatial weights. Let a region be comprised of n disjoint areas and let $\boldsymbol{b} = [b_1, \ldots, b_n]^\top$ be a set of random effects included in a Bayesian hierarchical disease mapping model. Besag (1974) assumes the following set of full conditional distributions:

$$b_i \mid \boldsymbol{b}_{(-i)}, \sigma \sim \mathcal{N}\left(\frac{1}{d_i}\sum_{j=1}^{n} w_{ij} b_j, \frac{\sigma^2}{d_i}\right), \; i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{b}_{(-i)} = [b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n]^\top$. The neighbourhood structure $\boldsymbol{W} = [w_{ij}]$ is defined through the spatial weights w_{ij} and let $d_i = \sum_{j=1}^{n} w_{ij}$ be the sum of spatial weights for each area. The most common spatial weights are 0–1 weights such that $w_{ij} = 1$ if areas i and j share a border, and 0 otherwise. In that case, $d_i = \sum_{j \sim i} w_{ij}$, where $j \sim i$ means that area j is a neighbour of i, corresponds to the number of neighbours that the ith area has. Other neighbourhood structures could be considered (Banerjee et al., 2014). For example, w_{ij} may be defined as a function of the distance between the centroids of areas i and j, or one could define $w_{ij} = 1$ if area j is part of the set of K-nearest neighbours of area i. The most common 0–1 neighbourhood structure is considered throughout this thesis. Regardless of the choice for $\boldsymbol{W}$, the use of Brook's lemma (Brook, 1964) on the set of ICAR full conditionals (2.1)

leads to a joint distribution for $\boldsymbol{b}$ that is proportional to

$$\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{b}^\top \boldsymbol{Q}\boldsymbol{b}\right) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i\neq j} w_{ij}(b_i - b_j)^2\right), \tag{2.2}$$

where $\boldsymbol{Q} = \boldsymbol{D} - \boldsymbol{W}$, for $\boldsymbol{D} = diag(d_i)$ (Banerjee et al., 2014). Throughout this thesis, for a vector $\boldsymbol{a} = [a_1, \ldots, a_n]^\top$, the notation $diag(\boldsymbol{a})$ or $diag(a_i)$ is used indiscriminately. Because $\boldsymbol{Q}$ is not a positive definite "precision" matrix, this joint distribution, denoted ICAR$(\sigma^2, \boldsymbol{Q})$ hereafter, is not a proper multivariate normal distribution. This impropriety implies that data could not be modelled through an ICAR$(\sigma^2, \boldsymbol{Q})$ distribution. Nevertheless, a work around to deal with the impropriety of the ICAR distribution and guarantee the propriety of the resulting posterior distribution is to impose a sum-to-zero constraint such that $\sum_{i=i}^{n} b_i = 0$ (Banerjee et al., 2014). The necessity for this constraint stands out from the right-hand side of equation (2.2), where it can be seen that were a constant added to all latent effects, it would not be identifiable without the sum-to-zero constraint. Alternatively, it is possible to make the ICAR distribution proper through the inclusion of another parameter. More specifically, $\boldsymbol{Q}$ is altered into becoming a positive definite precision matrix. The proper conditional-autoregressive (PCAR) prior introduces $\boldsymbol{Q}_\alpha = \boldsymbol{D} - \alpha\boldsymbol{W}$, which is a valid precision matrix for $|\alpha| < 1$ (Banerjee et al., 2014). The parameter α can be either fixed or estimated; in this case, from a Bayesian point of view, a prior distribution must be assigned.

In the literature, different models built on the ICAR have been proposed to allow for spatially structured latent effects (see, e.g., section 3.1 of Riebler et al. (2016) for a review). Besag et al. (1991) introduced the BYM model where each random effect b_i is decomposed into the sum of two components, θ_i and u_i. The two sets of random effects $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]^\top$ and $\boldsymbol{u} = [u_1, \ldots, u_n]^\top$ are assumed independent and one follows an ICAR prior structure, whereas the other follows an independent structure across space. More specifically, Besag et al. (1991) assume

$$b_i = \theta_i + u_i, \;\; i = 1, \ldots, n, \tag{2.3}$$

with $\boldsymbol{\theta} \sim \mathcal{N}_n(\mathbf{0}, \sigma_\theta^2 \boldsymbol{I}_n)$ and $\boldsymbol{u} \sim \text{ICAR}(\sigma_u^2, \boldsymbol{Q})$, where $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix. The introduction of the independent and identically distributed (i.i.d.) effects $\boldsymbol{\theta}$ in this convolution model relaxes the assumption of a purely spatially structured variation. However, it is important to mention that while σ_θ^2 is the variance of the *marginal* distribution of the unstructured effects, σ_u^2 is the variance of a *conditional* distribution of the spatial effects and hence depends on the neighbourhood structure under study (Sørbye and Rue, 2014). Therefore, interpretation and prior assignment for these parameters should be done with care. Moreover, while the sum $\theta_i + u_i$ is identifiable, the variance parameters σ_θ^2 and σ_u^2 suffer from an identifiability issue (MacNab, 2011; Lawson, 2018). The model proposed by Leroux et al. (1999) overcomes this concern by introducing a single areal latent effect b_i whose covariance structure includes a spatial dependence parameter λ. Through the introduction of this mixing parameter, the precision matrix of $\boldsymbol{b}$ is the weighted sum of a spatially structured matrix and an unstructured one. Leroux et al. (1999) assume $\boldsymbol{b} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 \boldsymbol{Q}_L^{-1}\right)$, where $\boldsymbol{Q}_L = (1-\lambda)\boldsymbol{I}_n + \lambda \boldsymbol{Q}$ is a valid precision matrix for $\lambda \in [0, 1)$. The joint Leroux prior corresponds to the following set of full conditional distributions:

$$b_i \mid \boldsymbol{b}_{(-i)}, \lambda, \sigma \sim \mathcal{N}\left(\frac{\lambda}{1-\lambda+\lambda d_i} \sum_{j=1}^{n} w_{ij} b_j, \frac{\sigma^2}{1-\lambda+\lambda d_i}\right), \; i = 1, \ldots, n. \tag{2.4}$$

From the decomposition of $\boldsymbol{Q}_L$ and (2.4), one can see that when $\lambda = 0$, the latent effects are purely random with no spatial structure, whereas $\lambda = 1$ results in the $\text{ICAR}(\sigma^2, \boldsymbol{Q})$ distribution.

To allow for a spatial structure as well as an independent one using two sets of random effects, Dean et al. (2001) proposed a reparametrisation of the BYM model (2.3) where, similar to Leroux et al. (1999), they include a spatial dependence parameter, λ. They decompose $b_i = \sigma\left(\sqrt{1-\lambda}\theta_i + \sqrt{\lambda} u_i\right)$, $i = 1, \ldots, n$, where the unstructured effects $\boldsymbol{\theta} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I}_n)$ and the spatially structured $\boldsymbol{u} \sim \text{ICAR}(1, \boldsymbol{Q})$ are independent. This weighted sum leads to the joint distribution $\boldsymbol{b} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2\left[(1-\lambda)\boldsymbol{I}_n + \lambda \boldsymbol{Q}^-\right]\right)$, where $\boldsymbol{Q}^-$ is the Moore-Penrose

generalised inverse of $\boldsymbol{Q}$. This generalised inverse is computed because $\boldsymbol{Q}$ is not of full rank. To see that the Dean model is a reparametrisation of the BYM model (2.3), note that $b_i = \theta_i^{\text{BYM}} + u_i^{\text{BYM}}$, where $\boldsymbol{\theta}^{\text{BYM}} = \left[\theta_1^{\text{BYM}}, \ldots, \theta_n^{\text{BYM}}\right]^\top \sim \mathcal{N}_n(\mathbf{0}, \sigma_\theta^2 \boldsymbol{I}_n)$ and $\boldsymbol{u}^{\text{BYM}} = \left[u_1^{\text{BYM}}, \ldots, u_n^{\text{BYM}}\right]^\top \sim \text{ICAR}(\sigma_u^2, \boldsymbol{Q})$ are independent, with $\sigma_\theta^2 = \sigma^2(1-\lambda)$ and $\sigma_u^2 = \sigma^2\lambda$. One difference between the Leroux and Dean priors, other than the use of two sets of random effects, is that the spatial and independent structures appear in the *precision* matrix of the Leroux prior, whereas in Dean's proposal, they are introduced in the *covariance* matrix.

Sørbye and Rue (2014) point out that in the cases of ICAR and Leroux distributed spatial effects (namely, all the spatial effects listed above), the variance parameters σ_u^2 and σ^2 lie in the *conditional* distributions of the random effects. Hence, their impact depends on the neighbourhood structure of the region of interest. These conditional variance parameters play a role in the amount of smoothing of the spatial effects across the areas. This implies that a prior imposed on the conditional variance parameter may not lead to the same level of smoothing when studying two different regions (e.g., the 96 French departments and the 160 districts of Rio de Janeiro). In particular, Best et al. (1999) discuss the sensitivity to prior assignments in the BYM model. Therefore, in the case of the ICAR prior, Sørbye and Rue (2014) suggest scaling the spatial effects by a factor h in order to guarantee that σ^2, or σ_u^2, approximately corresponds to the *marginal* variance of the spatial components, and is independent of the neighbourhood structure of the region. In particular, they compute h as the generalised variance of the spatial effects: $h = \exp\left[(1/n)\sum_{i=1}^n \ln\left(Q_{ii}^-\right)\right]$, where A_{ij} denotes the element on the ith row and the jth column of a matrix $\boldsymbol{A}$. Riebler et al. (2016) argue that although the Leroux prior cannot be scaled, one may build on Sørbye and Rue (2014) and modify the Dean model to include scaled spatial effects. They propose the BYM2 model, which decomposes each random effect as the following weighted sum:

$$b_i = \sigma\left(\sqrt{1-\lambda}\theta_i + \sqrt{\lambda}u_i^\star\right), \quad i = 1, \ldots, n, \tag{2.5}$$

with unstructured components $\boldsymbol{\theta} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I}_n)$ independent of the scaled spatial effects $\boldsymbol{u}^\star = [u_1^\star, \ldots, u_n^\star]^\top = \boldsymbol{u}/\sqrt{h}$, where $\boldsymbol{u} \sim \text{ICAR}(1, \boldsymbol{Q})$. Alternatively, one may write $\boldsymbol{u}^\star \sim \text{ICAR}(1, \boldsymbol{Q}_\star)$, with scaled "precision" matrix $\boldsymbol{Q}_\star = h\boldsymbol{Q}$, such that $\mathbb{V}(u_i^\star) \simeq 1,\ i = 1, \ldots, n$. Hence, the scaling process of the spatial components implies that σ^2 is approximately the *marginal* variance of each latent effect: $\mathbb{V}(b_i \mid \sigma, \lambda) = \sigma^2 [(1-\lambda)\mathbb{V}(\theta_i) + \lambda \mathbb{V}(u_i^\star)] \simeq \sigma^2 [(1-\lambda) \times 1 + \lambda \times 1] = \sigma^2$. It is worth mentioning that similar to Dean et al. (2001), the covariance matrix for $\boldsymbol{b}$ is a combination of the unstructured and spatially structured matrices, $\sigma^2 [(1-\lambda)\boldsymbol{I}_n + \lambda \boldsymbol{Q}_\star^-]$. Finally, Riebler et al. (2016) point out that due to the scaling process of the spatial effects, the variance and mixing parameters, σ^2 and λ, now lie in the *marginal* distribution of the random effects. Therefore, their influence on the smoothing of the random effects is independent of the neighbourhood structure of the region of interest, which eases their prior assignment and interpretation. The work summarised in the first manuscript (Chapter 3) of this thesis builds on the BYM2 model to take advantage of the interpretability of the model parameters.

2.1.1 Spatial discontinuity

All the models described in the previous section assume constant variability of the latent effects across space. However, it is reasonable to imagine that some areas may have abnormally high or low disease risks which might not be well accommodated by covariates and smooth spatial effects. Richardson et al. (2004) explicitly state how important it is for disease mapping models to be able to adapt between smoothing and adjusting to abrupt changes in the risk surface. While the models from Section 2.1 aim to smooth the risk surface, more recently, other models have been proposed to allow for spatial autocorrelation while simultaneously accommodating spatial discontinuities.

In order to adjust to abrupt changes in the risk surface, Lawson and Clark (2002) build on the BYM model (2.3) and further decompose the spatial effect into a mixture of two spatially structured components. They assume $b_i = \theta_i + \lambda_i u_i^{(1)} + (1-\lambda_i) u_i^{(2)},\ i = 1, \ldots, n,$

where the unstructured components $\boldsymbol{\theta} \sim \mathcal{N}_n(\mathbf{0}, \sigma_\theta^2 \boldsymbol{I}_n)$ are independent of the two independent sets of spatial effects, $\boldsymbol{u}^{(1)} = \left[u_1^{(1)}, \ldots, u_n^{(1)}\right]^\top$ and $\boldsymbol{u}^{(2)} = \left[u_1^{(2)}, \ldots, u_n^{(2)}\right]^\top$. The first vector of spatial components is assigned an ICAR prior, $\boldsymbol{u}^{(1)} \sim \text{ICAR}(\sigma_{u1}^2, \boldsymbol{Q})$, while the second vector is assumed to follow a "jump" model whose joint distribution is proportional to $\exp\left(-1/(2\sigma_{u2}^2) \sum_{i \neq j} w_{ij} \left|u_i^{(2)} - u_j^{(2)}\right|\right)$. Note how this jump model resembles the ICAR prior (2.2), using the L_1 distance instead of the L_2. This model aims to allow for jumps, or local discrepancies, in the spatial surface. Yan (2007) proposes an alternative way to introduce a new set of spatially structured components in the BYM model to allow for spatial heteroscedasticity, and hence spatial discontinuity. Yan writes the BYM model (2.3) as $b_i \mid u_i, \sigma_\theta^2 \sim \mathcal{N}\left(u_i, \sigma_\theta^2\right),\ i = 1, \ldots, n$, and points out that one can account for heteroscedasticity by allowing the variance σ_θ^2 to vary across space. Specifically, Yan decomposes each latent effect as the sum $b_i = \theta_i + u_i^{(1)},\ i = 1, \ldots, n$, with ICAR spatial components $\boldsymbol{u}^{(1)} \sim \text{ICAR}(\sigma_{u1}^2, \boldsymbol{Q})$ independent of the heteroscedastic random effects $\theta_i \sim \mathcal{N}\left(0, \sigma_{\theta_i}^2\right)$, where $\ln\left(\sigma_{\theta_i}^2\right) = \varsigma + u_i^{(2)},\ i = 1, \ldots, n$, and $\boldsymbol{u}^{(2)} \sim \text{ICAR}(\sigma_{u2}^2, \boldsymbol{Q})$. However, Congdon (2017) argues that the inclusion of three different sets of random effects in the models proposed by Lawson and Clark (2002) and Yan (2007) leads to identifiability issues.

Another approach to adjust for discrepancies in the risk surface lies in clustering methods, which are multiple-step procedures. First, clusters are elicited to separate the areas of the region of interest based on the observed data; then the clustering information is incorporated within a disease mapping model through added cluster effects, which may be fixed or random. Anderson et al. (2014) propose a $(n+1)$-step approach where n different sets of clusters are first defined, and then n models are fitted to the data, one for each set of clusters. They provide an algorithm to elicit the n different sets of clusters based on the neighbourhood structure and the differences between the areal log SMRs, where various difference measures are discussed. Then, for each set of k clusters of areal indices, $\mathcal{C}_j,\ j = 1, \ldots, k$, each areal latent effect is decomposed into the sum of a spatial random effect and cluster fixed effects as follows: $b_i = u_i + \sum_{j=1}^{k} \mathbb{1}_{[i \in \mathcal{C}_j]} \beta_j,\ i = 1, \ldots, n$, where $\mathbb{1}_{[\cdot]}$ denotes the indicator function, $\boldsymbol{u} \sim$

ICAR($\sigma_u^2, \boldsymbol{Q}$), and $\beta_j \overset{i.i.d.}{\sim} \mathcal{N}(0, 10)$. Finally, the best clustering partition is defined as the one that leads to the smallest deviance information criterion (DIC, Spiegelhalter et al. (2002)). Santafé et al. (2021) note that when the number of areas is large, the clustering method proposed by Anderson et al. (2014) is not computationally feasible. They propose a different clustering approach that is a two-step procedure. First, they introduce a new algorithm to elicit clusters, namely the density-based spatial clustering (DBSC) algorithm that leads to a single cluster partition. Then, based on the resulting set of clusters, they consider three different models. If the DBSC algorithm results in no clusters, the latent effects are assumed to follow the Leroux prior (2.4), $\boldsymbol{b} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{Q}_L^{-1})$. If k clusters $\mathcal{C}_j,\ j = 1, \ldots, k$, are elicited, for k small, then each latent effect is decomposed as $b_i = u_i + \sum_{j=1}^{k} \mathbb{1}_{[i \in \mathcal{C}_j]} \beta_j,\ i = 1, \ldots, n$, with $\boldsymbol{u}$ modelled through the Leroux prior, and $\beta_j \overset{i.i.d.}{\sim} \mathcal{N}(0, 10)$. Finally, if k is large, then the areal effects are $b_i = u_i + \sum_{j=1}^{k} \mathbb{1}_{[i \in \mathcal{C}_j]} \delta_j,\ i = 1, \ldots, n$, where the cluster random effects are further assumed to follow a Leroux prior, $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_k]^\top \sim \mathcal{N}_k\left(\boldsymbol{0}, \sigma_\delta^2 \boldsymbol{Q}_{L,\delta}^{-1}\right)$. The $k \times k$ precision matrix is defined as $\boldsymbol{Q}_{L,\delta} = (1 - \lambda_\delta)\boldsymbol{I}_k + \lambda_\delta\left(\boldsymbol{D}_\delta - \boldsymbol{W}_\delta\right)$, where the neighbourhood matrix relative to the clusters, $\boldsymbol{W}_\delta = \left[w_{\ell j}^{(\delta)}\right]$, is defined based on the adjacency between the areas within clusters ℓ and j, and $\boldsymbol{D}_\delta = diag\left(\sum_{j=1}^{k} w_{\ell j}^{(\delta)}\right)$. Adin et al. (2022) further develop the clustering approach proposed by Santafé et al. (2021) to allow for the inclusion of covariates.

A third approach to adjust to changes in the risk surface, after accounting for the fixed effects, is that of Congdon (2017), which is a single-step method that aims to accommodate and identify potentially outlying areas. Congdon proposes a modification of the Leroux prior (2.4) by including scaling mixture components, $\boldsymbol{\kappa} = [\kappa_1, \ldots, \kappa_n]^\top$, where $\kappa_i > 0,\ i = 1, \ldots, n$, whose role is to allow for discrepancies. Congdon assumes

$$b_i \mid \boldsymbol{b}_{(-i)}, \boldsymbol{\kappa}, \lambda, \sigma^2 \sim \mathcal{N}\left(\frac{\lambda}{1 - \lambda + \lambda d_i} \sum_{j=1}^{n} w_{ij} \kappa_j b_j, \frac{\sigma^2}{\kappa_i (1 - \lambda + \lambda d_i)}\right),\ i = 1, \ldots, n, \tag{2.6}$$

where $\kappa_i \overset{i.i.d.}{\sim} \text{Gamma}(\nu/2, \nu/2)$. These scaling mixture parameters are termed "outlier

indicators" in Congdon (2017), as $\kappa_i < 1$ indicates that area i is an outlier, after accounting for the fixed effects. Specifically, let area i be an outlier, then $\kappa_i < 1$ inflates the conditional variance in (2.6), which allows the latent effect to differ from the overall surface. Moreover, let neighbouring areas $i \sim j$ be non-outlying and outlying, respectively. Then $\kappa_j < 1$, and the contribution of b_j in the conditional mean of b_i is decreased. Therefore, the ith area borrows less strength from its outlying neighbour j, than from its other neighbours whose $\kappa \not< 1$. Note that $\boldsymbol{\kappa} = \mathbf{1}_n = [1, \ldots, 1]^\top$ yields the Leroux prior (2.4). Additionally, the full conditionals (2.6) result in the joint distribution $\boldsymbol{b} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 \boldsymbol{Q}_C^-\right)$, where $\boldsymbol{Q}_C = \left[Q_{C_{ij}}\right]$ has diagonal elements $Q_{C_{ii}} = \kappa_i(1 - \lambda + \lambda d_i)$, $i = 1, \ldots, n$, and off-diagonal elements $Q_{C_{ij}} = -\lambda w_{ij} \kappa_i \kappa_j$, $i \neq j$. Although this symmetric matrix is not always a valid precision matrix for $\lambda \in [0, 1)$, the diagonal dominance condition (Banerjee et al., 2014) states that for $\boldsymbol{Q}_C$ to be symmetric positive definite, it is sufficient that $Q_{C_{ii}} > \sum_{j \neq i} |Q_{C_{ij}}|$, $\forall i \Leftrightarrow \lambda < \min_i \left\{1 / \left(1 - d_i + \sum_{j \neq i} w_{ij} \kappa_j\right)\right\}$. One appeal of Congdon's approach, compared to a clustering procedure (e.g., Santafé et al. (2021)), is that in a hierarchical Bayesian model, all the model parameters, including the scaling mixture components, are estimated in a single step. The first two manuscripts of this thesis (Chapters 3 and 4) build on the model proposed by Congdon (2017) to accommodate outlying areas in purely spatial and spatio-temporal settings. The main aim of the work presented in these first two chapters is to specifically identify potential outliers, which may help decision makers prioritise interventions.

It is worth mentioning that alternative approaches that rely on the estimation of the neighbourhood matrix $\boldsymbol{W}$ have been proposed to accommodate discontinuities in the estimated surface. The goal of these proposals is to adjust for spatial discrepancies without necessarily identifying outliers, as is the case for the model proposed by Congdon (2017). For a binomial outcome, Dean et al. (2019) propose a two-step procedure where they first test for statistical differences between the observed proportions in neighbouring areas. When two neighbouring proportions are found to be statistically different, the spatial structure is updated such that the two areas are not considered as neighbours. Specifically, the up-

dated neighbourhood matrix $\boldsymbol{W}_{\text{updated}} = \left[w_{ij}^{(updated)}\right]$ has elements $w_{ij}^{(\text{updated})} = 0$, if $w_{ij} = 0$, $w_{ij}^{(\text{updated})} = 0$, if neighbouring areas $i \sim j$ ($w_{ij} = 1$) are found to have statistically different proportions, and $w_{ij}^{(\text{updated})} = 1$, otherwise. Then, using the updated neighbourhood structure, a Leroux prior is assigned to the latent effects. More recently, Corpas-Burgos and Martinez-Beneito (2020) propose to adapt to spatial discontinuities by relaxing the 0–1 structure assumption, and estimate a vector of parameters $\boldsymbol{c} = [c_1, \ldots, c_n]^\top$ in order to alter the matrix $\boldsymbol{W}$ as follows: $\boldsymbol{W}_{\text{CBMB}} = diag\left(\boldsymbol{c}^{1/2}\right)\boldsymbol{W}diag\left(\boldsymbol{c}^{1/2}\right)$. Hence, $\boldsymbol{W}_{\text{CBMB}} = \left[w_{ij}^{(\text{CBMB})}\right]$ has elements $w_{ij}^{(CBMB)} = 0$, if $w_{ij} = 0$, and $w_{ij}^{(\text{CBMB})} = \sqrt{c_i c_j}$, otherwise. Further, they assume $c_i \overset{i.i.d.}{\sim} \text{Gamma}(\nu, \nu)$, $i = 1, \ldots, n$. The authors extend both the ICAR prior (2.2) and the Leroux model (2.4) in the univariate and multivariate disease mapping settings. In particular, one parametrisation of their so-called adaptive Leroux prior is such that $\boldsymbol{b} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2\left[(1-\lambda)diag\left(\boldsymbol{c}^{1/2}\right) + \lambda\left(\boldsymbol{D}^{\text{CBMB}} - \boldsymbol{W}^{\text{CBMB}}\right)\right]^{-1}\right)$, where $\boldsymbol{D}^{\text{CBMB}} = diag\left(\sum_{j=1}^{n} w_{ij}^{(CBMB)}\right)$. This distribution corresponds to the n full conditionals $b_i \mid \boldsymbol{b}_{(-i)}, \boldsymbol{c}, \lambda, \sigma \sim \mathcal{N}\left(\lambda\sum_{j\neq i} w_{ij}\sqrt{c_j}b_j / \left(1-\lambda+\lambda d_i^{(c)}\right), \sigma^2 / \left(\sqrt{c_i}\left[1-\lambda+\lambda d_i^{(c)}\right]\right)\right)$, $i = 1, \ldots, n$, where $d_i^{(c)} = \sum_{j\neq i} w_{ij}\sqrt{c_j}$. One may notice that this distribution resembles the full conditional (2.6) proposed by Congdon (see, e.g., Table A.6 in Appendix A.8 of the first manuscript). However, Corpas-Burgos and Martinez-Beneito (2020) remark that in the case of a univariate outcome, their parametrisation may suffer from an identifiability issue. This is not the case with the proposal by Congdon (2017).

2.1.2 Spatio-temporal framework

We now focus on the case of disease counts recorded over time and across the areas of a region of interest, which is the main focus of the second manuscript (Chapter 4). Spatio-temporal disease mapping models aim to estimate the evolution of the disease relative risk across space. Commonly, spatio-temporal extensions of the hierarchical Bayesian disease mapping models described in Section 2.1 assume that the number of cases in an area at a point in time follows a Poisson distribution whose mean is the product of an offset and the relative

risk of the disease, which varies over time and across space. The offsets correspond again to the expected number of cases for a uniform spread of the disease. These may vary across space and time (see, e.g., Bernardinelli et al. (1995); Ugarte et al. (2012)), or depend on space only (see, e.g., Freitas et al. (2021)). The first spatio-temporal model for areal counts was proposed by Bernardinelli et al. (1995). In the log scale, they decompose the relative risk as the sum of an overall rate, a random spatial component, a temporal fixed effect, and a space-time interaction term. Specifically, let μ_{it} be the relative risk at time $t = 1, \ldots, T$, in area $i = 1, \ldots, n$, they assume: $\ln(\mu_{it}) = \beta_0 + b_{it}$, where the spatio-temporal latent effect is $b_{it} = u_i^{(1)} + \beta \text{time}_t + u_i^{(2)} \text{time}_t$, with covariate time_t. Different priors are discussed for the two vectors of spatial effects, $\boldsymbol{u}^{(1)}$ and $\boldsymbol{u}^{(2)}$, including an independent normal distribution, the ICAR prior (2.1), and the BYM model (2.3). This specification of the log risk, however, does not allow for random temporal effects.

In the case of a binomial outcome, Knorr-Held (2000) extends this spatio-temporal model such that the log odds are decomposed into $\beta_0 + b_{it}$, where $b_{it} = v_t^{(1)} + v_t^{(2)} + \theta_i + u_i + \varepsilon_{it}$. Knorr-Held includes unstructured temporal and spatial effects, $\boldsymbol{v}^{(1)} = \left[v_1^{(1)}, \ldots, v_T^{(1)}\right]^\top \sim \mathcal{N}_T(\boldsymbol{0}, \sigma_{v1}^2 \boldsymbol{I}_T)$ and $\boldsymbol{\theta} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_\theta^2 \boldsymbol{I}_n)$, respectively, as well as temporally and spatially structured components, $\boldsymbol{v}^{(2)} = \left[v_1^{(2)}, \ldots, v_T^{(2)}\right]^\top$ and $\boldsymbol{u} \sim \text{ICAR}(\sigma_u^2, \boldsymbol{Q})$, respectively. A first-order random walk is assigned to $\boldsymbol{v}^{(2)}$, with joint distribution proportional to

$$\exp\left(-1/(2\sigma_{v2}^2) \sum_{t=2}^{T} \left(v_t^{(2)} - v_{t-1}^{(2)}\right)^2\right) \propto \exp\left(-1/(2\sigma_{v2}^2) \boldsymbol{v}^{(2)\top} \boldsymbol{R} \boldsymbol{v}^{(2)}\right), \tag{2.7}$$

or $\boldsymbol{v}^{(2)} \sim \mathcal{N}_T(\boldsymbol{0}, \sigma_{v2}^2 \boldsymbol{R}^-)$, where the $T \times T$ temporal structure matrix $\boldsymbol{R} = [R_{ij}]$ is a tridiagonal matrix with upper and lower diagonal elements $R_{i-1\ i} = R_{i\ i+1} = -1$, and with main diagonal elements $R_{11} = R_{nn} = 1$ and $R_{ii} = 2$, $i = 2, \ldots, n-1$. The matrix $\boldsymbol{R}$ may be seen as a temporal counterpart of $\boldsymbol{Q}$, where each time point has two neighbours: the previous and the following time points. Knorr-Held discusses four types of models for the space-time interaction component ε_{it}, namely, an unstructured parametrisation,

a purely temporal evolution, a purely spatial autocorrelation, and a fully spatio-temporal structure. Knorr-Held argues that the fourth parametrisation is the most interesting one, that is, $\boldsymbol{\varepsilon} = [\varepsilon_{11}, \ldots, \varepsilon_{n1}, \ldots, \varepsilon_{1T}, \ldots, \varepsilon_{nT}]^\top \sim \mathcal{N}_{n\times T}\left(\mathbf{0}, \sigma_\varepsilon^2 \left(\boldsymbol{R} \otimes \boldsymbol{Q}\right)^-\right)$, where $\otimes$ denotes the Kronecker product. Finally, when this fourth interaction parametrisation is considered, Knorr-Held further proposes to remove the unstructured temporal effects, such that $b_{it} = v_t + \theta_i + u_i + \varepsilon_{it}$, with $\boldsymbol{v} = [v_1, \ldots, v_T]^\top \sim \mathcal{N}_T\left(\mathbf{0}, \sigma_v^2 \boldsymbol{R}^-\right)$ and the sum of spatial effects $\theta_i + u_i,\ i = 1, \ldots, n$, corresponds to the BYM model (2.3).

More recently, Ugarte et al. (2012) build on Knorr-Held (2000) and decompose the spatio-temporal latent effects such that $b_{it} = u_i + v_t + \varepsilon_{it}$, with temporal effects assigned a random walk prior (2.7), and spatio-temporal effects assumed to follow the fourth interaction parametrisation proposed by Knorr-Held, $\boldsymbol{\varepsilon} \sim \mathcal{N}_{n\times T}\left(\mathbf{0}, \sigma_\varepsilon^2 \left(\boldsymbol{R} \otimes \boldsymbol{Q}\right)^-\right)$. The proposal by Ugarte et al. (2012) is particularly appealing because they reduce the parameter space by including a single set of spatial effects $\boldsymbol{u}$, which are modelled according to the Leroux prior (2.4). Rushworth et al. (2014) propose to further reduce the parameter space and assume that the spatio-temporal latent effects only include space-time interaction terms. Additionally, they propose a temporal extension of the Leroux prior (2.4) to model the interaction components. Specifically, they assume the vector of spatio-temporal effects $\boldsymbol{b} = [b_{11}, \ldots, b_{n1}, \ldots, b_{1T}, \ldots, b_{nT}]^\top$ to be modelled as follows:

$$\boldsymbol{b}_{\cdot 1} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 \boldsymbol{Q}_L^{-1}\right), \quad \text{and} \quad \boldsymbol{b}_{\cdot t} |\, \boldsymbol{b}_{\cdot t-1} \sim \mathcal{N}_n\left(\alpha \boldsymbol{b}_{\cdot t-1}, \sigma^2 \boldsymbol{Q}_L^{-1}\right),\ t = 2, \ldots, T, \tag{2.8}$$

where $\boldsymbol{b}_{\cdot t} = [b_{1t}, \ldots, b_{nt}]^\top,\ t = 1, \ldots, T$, and $\alpha \in [0, 1]$ is a temporal smoothing parameter, which may be seen as the temporal counterpart of λ. If $\alpha = 0$, the spatio-temporal effects are purely spatially structured, with $\boldsymbol{b}_{\cdot t}$ independent of $\boldsymbol{b}_{\cdot t-1}$, and if $\alpha = 1$, $\boldsymbol{b}$ is fully structured across space and over time. The model proposed in the second manuscript (Chapter 4) of this thesis builds on the Rushworth model (2.8).

In this section, the spatio-temporal distributions listed thus far are temporal extensions of

the spatial priors for areal data described in Section 2.1. However, Rushworth et al. (2014) mention that their proposal does not accommodate spatial discrepancies in the risk surface. Models have been proposed in the literature on spatio-temporal disease mapping to try to allow for spatial disparities over time. For example, as an extension of the clustering approaches discussed in Section 2.1.1, Lee and Lawson (2016) allow for jumps in the risk surface over time by including cluster effects in the spatio-temporal components. They assume $b_{it} = \beta_{Z_{it}} + \varepsilon_{it}$, with $\boldsymbol{\varepsilon}$ distributed according to the Rushworth model (2.8), and where the k cluster-specific intercepts are indicated by $Z_{it} \in \{1, \ldots, k\}$ and are assigned a uniform prior. Alternatively, Rushworth et al. (2017) extend the Rushworth model (2.8) such that the spatial neighbourhood structure $\boldsymbol{W}$ is estimated from the data. Specifically, in the logit scale, the non-zero spatial weights, $\ln\left(w_{ij}/(1-w_{ij})\right),\ i \sim j$, are assumed to follow a Leroux prior (2.4) such that the back transformation yields spatial weights estimated between 0 and 1.

It is worth mentioning that additional ways to model spatio-temporal areal counts have been proposed in the literature. For example, Nobre et al. (2005) consider a modification of the ICAR prior (2.1) that is similar to a dynamic linear model, where the conditional variance parameter is allowed to vary with time in the log scale. Specifically, they assume $\boldsymbol{b}_{\cdot t} \sim \text{ICAR}(\sigma_t^2, \boldsymbol{Q}),\ t = 1, \ldots, T$, with $\ln\left(\sigma_t^2\right) \sim \mathcal{N}\left(\ln\left(\sigma_{t-1}^2\right), \sigma_\sigma^2\right),\ t = 1, \ldots, T$, and $\ln\left(\sigma_0^2\right) = 0$. Similarly, Napier et al. (2016) allow the variance parameter to evolve through time within the Leroux model (2.4). They also include a purely temporal effect v_t and assume $b_{it} = v_t + \varepsilon_{it}$, where $\boldsymbol{\varepsilon}_{\cdot t} = [\varepsilon_{1t}, \ldots, \varepsilon_{nt}]^\top \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma_t^2 \boldsymbol{Q}_L^{-1}\right),\ t = 1, \ldots, T$. Independent inverse gamma priors are assigned to $\sigma_t^2,\ t = 1, \ldots, T$. Therefore, in these alternative specifications of spatio-temporal models, the interaction between space and time happens through the inclusion of time-dependent parameters within spatially structured random effects.

2.2 Small area estimation

In disease mapping (Section 2.1), we commonly assume that the data are observed across all areas of the region of interest. However, areal data may also arise from a survey where an outcome of interest is observed within some areas of a region (Lawson, 2018). In particular, when the areal sample sizes are small, small area estimation (SAE) methods are useful to obtain reliable estimates at the areal level (Rao and Molina, 2015). Interestingly, another name for disease mapping is "small area health studies". SAE is part of the survey sampling theory and has become increasingly popular in the last 50 years (see, e.g., Pfeffermann (2013); Ghosh (2020)), in particular among statistical official organisations (e.g., World Bank (2015); Census Bureau (2018)).

As part of the survey sampling theory, SAE methods may be divided into design-based and model-based approaches (Rao and Molina, 2015). The design-based framework assumes that within a finite population of interest, all variables are fixed and the randomness of an observed sample comes from the sampling process. In that case, an estimator of a quantity of interest relies on the sampling weights that result from the sampling design. On the other hand, the model-based framework assumes that the sample is fixed, and the outcome is treated as a random variable, as is the case in classical statistics. In that setting, an estimator of a quantity of interest relies on model assumptions and not on the sampling design.

In SAE, the quantity of interest is an areal summary of the response variable. Through design-based methods, this may be computed by direct estimators, which only use the response variable and sampling weights within a particular area to produce an estimate (e.g., weighted estimators introduced by Horvitz and Thompson (1952) or Hájek (1971)). Hence, one cannot produce estimates for areas that are missing from the sample. Indirect estimators, such as model-assisted estimators (e.g., GREG estimator, Särndal et al. (2003)) are proposed to borrow strength from other areas and covariates to allow for estimates in non-sampled

areas. The aim of indirect methods is to increase the effective areal sample sizes. However, when areal sample sizes are small, design-based estimates are commonly not reliable (e.g., low precision) and model-based methods are favoured (Tzavidis et al., 2018).

Let a region of interest be divided into n areas of sizes $N_i,\ i = 1, \ldots, n$, and let $k = 1, \ldots, N_i$ be the unit index. Assume that the interest is to estimate $\overline{y}_i = (1/N_i) \sum_{k=1}^{N_i} y_{ik},\ i = 1, \ldots, n$, the areal means of outcome y. This is the objective of the work summarised in the third manuscript of this thesis (Chapter 5), however, it is worth mentioning that other targets of inference may be of interest (e.g., non-linear quantity to measure poverty at the areal level (Molina et al., 2014)). Model-based SAE relies on a model assumption for the response variable y and vector of auxiliary information $\boldsymbol{x}$. In particular, model-based SAE often uses exterior sources of information (e.g., census) to augment the survey auxiliary variables. In the model-based framework, SAE can further be divided into unit-level and area-level approaches. In the first case, unit-level responses are linked to unit-level auxiliary variables. For example, for a continuous outcome, Battese et al. (1988) propose to model the response as follows: $y_{ik} = \boldsymbol{x}_{ik}^{\top}\boldsymbol{\beta} + b_i + e_{ik}$, where each areal random effect, b_i, is assumed independent of the unit-level error, e_{ik}, with zero means and variances σ_b^2 and σ_e^2, respectively. However, it may be difficult to obtain unit-level auxiliary information for the entire finite population, while areal summaries may be more accessible. In area-level models, the outcome and covariates are aggregated at the areal level before modelling. The first area-level SAE model is proposed by Fay and Herriot (1979), they assume $\widehat{\overline{y}}_i = \boldsymbol{x}_i^{\top}\boldsymbol{\beta} + b_i + e_i$, where $\widehat{\overline{y}}_i$ is the direct estimator of the areal mean response computed from the sample, $\boldsymbol{x}_i$ is the vector of area-level auxiliary variables known for all areas, b_i is an areal random effect with mean 0 and variance σ_b^2, and e_i is the sampling error with mean 0 and known design variance of the areal direct estimator, $\sigma_{e,i}^2$, computed using the sampling design information. Historically, model-based SAE is conducted under the frequentist framework, and empirical best linear unbiased predictors (EBLUPs) are computed. For example, under the Fay and Herriot model, the areal EBLUP is $\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}} + \widehat{b}_i = \widehat{\gamma}_i\widehat{\overline{y}}_i + (1 - \widehat{\gamma}_i)\,\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}$, with $\widehat{\gamma}_i = \widehat{\sigma}_b^2 / \left(\widehat{\sigma}_b^2 + \sigma_{e,i}^2\right)$, where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}_b^2$ denote the

estimators of $\boldsymbol{\beta}$ and σ_b^2, respectively. However, it is important to mention that model-based SAE approaches have been extended to the Bayesian framework (see, e.g., Datta and Ghosh (1991); Gómez-Rubio et al. (2010); Molina et al. (2014)) and, more recently, to machine learning procedures (see, e.g., Krennmair and Schmid (2022)). The third manuscript of this thesis investigates area-level SAE methods under all three settings (Chapter 5).

2.2.1 Variable selection and machine learning approaches

In model-based methods, when abundant auxiliary variables are available for all areas from the survey and exterior sources, it may be necessary to select a subset of covariates to model the outcome. Under any framework, variable selection is a common research topic in statistics (see, e.g., Porwal and Raftery (2022) for a comparison of 21 different selection methods under the frequentist and Bayesian paradigms). In the frequentist framework, the most common variable selection procedures are multiple-step approaches, where different models are fitted to the data by iteratively adding or removing covariates, based on a chosen comparison criterion. Specifically, a *forward* selection procedure starts with a model that includes only an intercept and adds one covariate at a time, based on the resulting criterion. Conversely, a *backward* elimination procedure starts with a full model and each step consists in removing one covariate based on the computed criterion. In both cases, once the set of relevant auxiliary variables is defined, a final model that includes these covariates is fitted to the data. Multiple criteria have been proposed in the literature, Wakefield (2013) lists the most widely used ones, including the Akaike information criterion (AIC, Akaike (1998)), the Bayesian information criterion (BIC), Mallow's C_p (Mallows, 1973), and the adjusted R^2. Although it is beyond the scope of this thesis, it is worth mentioning that design-based model comparison criteria have been proposed in the survey sampling literature (see, e.g., Lumley and Scott (2015)).

On the other hand, regularisation methods, or shrinkage methods, have also been proposed to select a subset of covariates and model the data in a single step. These regularisation

methods impose a constraint on the regression parameters in order to shrink the irrelevant coefficients towards 0. For example, the ridge regression (Hoerl and Kennard, 1970) assumes the constraint $\sum_{j=1}^{p} \beta_j^2 \leq c$, for some $c \geq 0$, where p is the total number of covariates. Alternatively, the ridge regression may be written in the following Lagrangian form:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ (1/2n) \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{2.9}$$

for some $\lambda > 0$. A different shrinkage approach is the least absolute shrinkage and selection operator (LASSO, Tibshirani (1996, 2011)), which imposes the constraint $\sum_{j=1}^{p} |\beta_j| \leq c$, or, in the Lagrangian form:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ (1/2n) \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2.10}$$

for some $\lambda > 0$. This is a popular variable selection method, as it leads to regression coefficients that are exactly zero and excludes the corresponding covariates from the model (Hastie et al., 2015). However, the LASSO yields non-linear estimates and inference should be conducted with care. For instance, Dezeure et al. (2015) argue that bootstrap approaches may not be adequate to assess the uncertainty of estimates obtained through the LASSO, due to the non-continuity of the distribution of the regression parameters.

In the Bayesian framework, variable selection consists in imposing an informative prior on the regression parameters. A covariate is said to be selected if the posterior credible interval of its corresponding coefficient does not include 0. Multiple shrinkage methods have been proposed in the Bayesian framework, where coefficients are assigned prior distributions that peak at 0. In particular, for a model of the form $y_i \overset{i.i.d.}{\sim} \mathcal{N}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2 \right),\ i = 1, \ldots, n$, then assuming the prior $\beta_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2/\lambda),\ j = 1, \ldots, p$, leads to a maximum a posteriori estimator of the form (2.9) (Reich and Ghosh, 2019). Therefore, Bayesian ridge regression consists in assigning each β parameter a normal prior with mean 0 and variance σ^2/λ. Similarly, the Bayesian LASSO

(Hans, 2010) assumes $\beta_j \overset{i.i.d.}{\sim} \mathcal{DE}(0, \lambda/\sigma^2),\ j = 1, \ldots, p$, where $\mathcal{DE}(a, b)$ denotes the double exponential distribution with mean a and scale b. This double exponential prior leads to a maximum a posteriori estimator of the form (2.10) (Reich and Ghosh, 2019). An alternative is the horseshoe prior proposed by Carvalho et al. (2010) that assumes $\beta_j \sim \mathcal{N}(0, \lambda_j^2\tau^2)$, and $\tau, \lambda_j \sim \mathcal{HC}(0, 1),\ j = 1, \ldots, p$, where $\mathcal{HC}(a, b)$ denotes the half-Cauchy distribution with location a and scale b. This popular shrinkage prior (Datta and Ghosh, 2013; Porwal and Raftery, 2022) is studied in a SAE context in the third manuscript of this thesis (Chapter 5).

Alternatively, machine learning approaches, and in particular random forests, may be used to naturally select relevant covariates. It is worth mentioning that random forests are becoming popular in the survey sampling literature (Breidt and Opsomer, 2017; Dagdoug et al., 2023), and are a novelty in the SAE context (Krennmair and Schmid, 2022; Newhouse, 2023). Breiman (2001) proposed a random forest algorithm, with a collection of B regression trees, where each tree consists in repetitively partitioning the data points into subgroups, which are called nodes, based on covariate splits. Each tree is grown on a bootstrap sample of the original dataset. The point estimates that result from Breiman's random forest procedure correspond to the average over the B estimates from the B trees. In each tree, a sequence of covariate splits leads to a number of final nodes, and an estimate is computed as the mean of the responses within the adequate final node. Therefore, in addition to naturally select auxiliary variables through covariate splits, random forests present the advantage that non-linear relationships between an outcome and covariates are inherently accommodated. However, the uncertainty assessment of random forest point estimates is not always straightforward (Wager and Athey, 2018), in particular, when the interest lies in predictions for new data points. Similar to the case of the LASSO, Wager et al. (2014) argue that bootstrap approaches may not be adequate for random forest uncertainty assessment, since numerous trees would need to be grown, which may be computationally inefficient. Although they are outside the scope of this thesis, it is worth mentioning that various Jackknife procedures

have been introduced to compute uncertainty intervals (see, e.g., Wager et al. (2014); Wager and Athey (2018); Lei et al. (2018)). These methods should be considered with care as they are proposed for different random forest algorithms. Recently, Zhang et al. (2019) proposed the so-called out-of-bag (OOB) prediction intervals for random forest point estimates. They argue that since the first step of a random forest algorithm is to select a bootstrap sample of the original dataset, there exist for each data point $(y_i, \boldsymbol{x}_i)$ a random forest that does not include $(y_i, \boldsymbol{x}_i)$. This smaller random forest comprises all the B_i trees grown on bootstrap samples that do not include $(y_i, \boldsymbol{x}_i)$. Therefore, from a single random forest procedure, each y_i has an OOB prediction, $\widehat{y}_i^{OOB}$, which is the prediction based on the random forest made of B_i trees. Let $d_i = y_i - \widehat{y}_i,\ i = 1, \ldots, n$, be the OOB errors, and let $d_{(\alpha)}$ be the $1-\alpha$ empirical quantile of the d's. The OOB prediction interval of a new data point, $\boldsymbol{x}$, is defined as $\left[\widehat{y} + d_{(\alpha/2)}, \widehat{y} + d_{(1-\alpha/2)}\right]$, where $\widehat{y}$ is the point estimate that results from inputting $\boldsymbol{x}$ in the random forest made of B trees. The simulation studies summarised in Zhang et al. (2019) show that the OOB prediction intervals perform similarly to the prediction intervals computed from the split conformal (SC) procedure proposed by Lei et al. (2018). However, one advantage of the SC procedure over the OOB approach is that the OOB procedure is tied to the random forest algorithm, whereas SC inference may be applied to a variety of modelling methods. In particular, in addition to random forests, Lei et al. (2018) consider the SC procedure to compute prediction intervals for estimates obtained through a LASSO regression (2.10). The first step of the SC procedure is to divide the dataset $\{(y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\}$ into two equal sized samples, S_1 and S_2. Then, the modelling method of interest (e.g., random forest, LASSO) is trained on S_1. Predictions are obtained for all data points in the remaining dataset, $\widehat{y}_i,\ i \in S_2$, and absolute residuals are computed, such that $R_i = |y_i - \widehat{y}_i|,\ i \in S_2$. Finally, the SC prediction interval of a new data point, $\boldsymbol{x}$, is $\left[\widehat{y} \pm R_{(\alpha)}\right]$, where $R_{(\alpha)}$ is the $1-\alpha$ empirical quantile of the R's and $\widehat{y}$ is the point estimate computed for the new data point $\boldsymbol{x}$. The SC procedure relies on the assumption of exchangeable data and guarantees that the prediction intervals yield the correct coverage (Angelopoulos and Bates, 2021). However,

the assumption of exchangeable data is strong, in particular in a SAE context, and some extensions of the SC procedure have been proposed to accommodate non-exchangeable data (see, e.g., Tibshirani et al. (2019); Barber et al. (2023)).

2.3 Summary

This chapter has provided an overview of the literature on disease mapping and SAE, which are methods used to analyse and provide estimates at the areal level. Regarding disease mapping methods, I discussed the issue of spatial discontinuity and reviewed models previously proposed to adapt to changes in the risk surface. I also discussed spatio-temporal models available in the literature to analyse data recorded across different areas of a region and over time. Finally, regarding SAE, I discussed model-based methods and different approaches to deal with numerous auxiliary variables.

Chapter 3

A Bayesian hierarchical model for disease mapping that accounts for scaling and heavy-tailed latent effects

Preamble to Manuscript 1. In disease mapping, Congdon (2017) proposed a modification of the Leroux prior (Leroux et al., 1999), where the latent effects are spatially structured and include independent scaling mixture components that aim to identify areas with potentially outlying disease risks, after accounting for the effect of covariates. Riebler et al. (2016) introduced the so-called BYM2 model, which decomposes each latent effect into the sum of an unstructured and a scaled spatially structured component, where the scaling process aims to ease interpretation and prior assignment of the model parameters. However, the BYM2 model assumes the variance of the latent effects is constant across areas. The goal of this manuscript is to relax this assumption and investigate if we are able to estimate the parameters of the proposed model.

This manuscript proposes an alternative disease mapping model to that of Congdon (2017). The proposed model is an extension of the BYM2 to allow for heavy-tailed latent effects

through the introduction of scaling mixture components to accommodate and identify potential outliers. Two prior specifications of the proposed model are investigated: one with independent scaling mixture parameters, and one where they are spatially structured.

The contributions of this manuscript include (i) a new disease mapping model that aims to identify potential outliers after accounting for covariates, (ii) a spatially structured distribution for the scaling mixture components, (iii) a comparison of the interpretation of the parameters included in the proposed and Congdon's models, (iv) a thorough simulation study to investigate the ability of the proposed model in identifying potential outliers compared to Congdon's prior, (v) a study on how the proposed model may help in the analysis of the first Zika epidemic (2015-2016) in Rio de Janeiro.

This manuscript is under a second round of review for the journal *Statistical Methods in Medical Research*.

A Bayesian hierarchical model for disease mapping that accounts for scaling and heavy-tailed latent effects

Victoire Michal[1], Alexandra M. Schmidt[1], Laís Picinini Freitas[2], Oswaldo Gonçalves Cruz[3].

[1]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada*

[2]*School of Public Health, University of Montreal, Montreal, Canada & Centre de Recherche en Santé Publique, Montreal, Canada*

[3]*Programa de Computação Científica (PROCC), Oswaldo Cruz Foundation, Rio de Janeiro, Brazil*

Abstract

In disease mapping, the relative risk of a disease is commonly estimated across different areas within a region of interest. The number of cases in an area is often assumed to follow a Poisson distribution whose mean is decomposed as the product between an offset and the logarithm of the disease's relative risk. The log risk may be written as the sum of fixed effects and latent random effects. The BYM2 model decomposes each latent effect into a weighted sum of independent and spatial effects. We build on the BYM2 model to allow for heavy-tailed latent effects and accommodate potentially outlying risks, after accounting for the fixed effects. We assume a scale mixture structure wherein the variance of the latent process changes across areas and allows for outlier identification. We propose two prior specifications for this scale mixture parameter. These are compared through simulation studies and in the analysis of Zika cases from the first (2015-2016) epidemic in Rio de Janeiro city, Brazil. The simulation studies show that, in terms of the model assessment criterion WAIC and outlier detection, the two proposed parametrisations perform better than the model proposed by Congdon (2017) to capture outliers. In particular, the proposed parametrisations are more efficient, in terms of outlier detection, than Congdon's when outliers are neighbours. Our analysis of Zika cases finds 23 out of 160 districts of Rio as potential outliers, after accounting for the socio-development index. Our proposed model may help prioritise interventions and identify potential issues in the recording of cases.

3.1 Motivation

The first Zika cases in the Americas were identified in 2015, when it was considered a benign disease. However, in October 2015 an unprecedented increase in the number of microcephaly cases in neonates was reported in the Northeast of Brazil and was later associated with the Zika virus infection during pregnancy (Lowe et al., 2018). The Zika virus is transmitted to humans by the bite of infected *Aedes* mosquitoes, the same vectors that transmit dengue, chikungunya and yellow fever. Dengue is the most prevalent *Aedes*-borne disease in the world and around 3.9 billion people in 129 countries are at risk of acquiring the disease (World Health Organization, 2020). Because of climate change, the global distribution of *Aedes* mosquitoes is expanding, increasing the number of people exposed to *Aedes*-borne diseases.

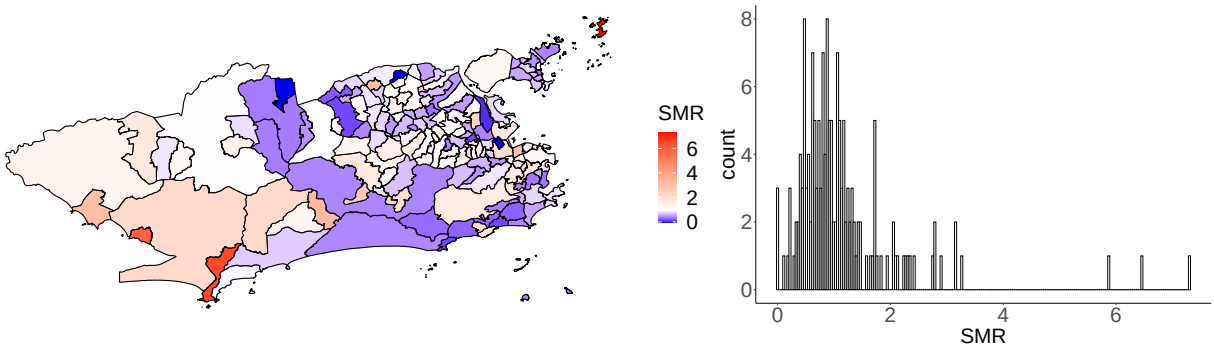
In the city of Rio de Janeiro, Brazil, the first Zika epidemic occurred between 2015 and 2016, with more than 35 thousand confirmed cases (Freitas et al., 2019). The city is the second-largest in Brazil, with approximately 6.3 million inhabitants, and its main tourist destination. Rio de Janeiro has a tropical climate and a favourable environment for the *Ae. aegypti* mosquitoes, which are highly adapted to urban settings. Despite efforts to control the vector population, the city has suffered from dengue epidemics every three to four years, in general (Nogueira et al., 1999; Honório et al., 2009; dos Santos et al., 2019). The widespread presence of the mosquito also allowed the entry and rapid dispersion of Zika and chikungunya viruses (Freitas et al., 2019). This epidemiological scenario highlights the need for novel strategies to help design interventions that are more effective in decreasing the burden of established *Aedes*-borne diseases and preventing emerging and re-emerging arbovirus diseases from causing new outbreaks. In this sense, we propose a model that has the potential to help prioritise interventions by identifying areas with outlying risks with respect to the entire region and with respect to their neighbours, while accounting for covariates.

Motivating the proposed model, we have available the Zika counts aggregated by neighbourhood for the period of the first Zika epidemic in the city of Rio de Janeiro. The data come from the Brazilian Notifiable Diseases Information System (SINAN – *Sistema de Informação de Agravos de Notificação*). In Brazil, cases attending healthcare facilities with a suspected diagnosis of Zika are reported to this system, usually by the physician. The standardised morbidity ratios (SMR) for the Zika counts by neighbourhood during the study period are presented in Figure 3.1. Although the epidemic affected most of the city, some neighbourhoods seem to have been hit harder than others and some, not at all. The diversity of the territory of Rio de Janeiro is possibly an important factor influencing this. Regarding the city's geography, for instance, there are mountains that separate different areas which may act as a natural barrier for the spread of the disease. Additionally, Rio's territory is heterogeneous in terms of demographic, socio-economic, and environmental characteristics that are involved in the distribution of *Aedes*-borne diseases (Freitas et al., 2021).

For this analysis, we have available the socio-development index, an index that includes indicators related to sanitation, education and income, and for which higher values represent better socio-economic conditions. In places with inadequate sanitary conditions, the female *Ae. aegypti* can more easily find any type of container filled with water to deposit her eggs. In Rio de Janeiro, a city with great social disparities, the socio-development index ranges from 0.282 (in Grumari, a neighbourhood in the West region) to 0.819 (in Lagoa, South region) (Prefeitura do Rio de Janeiro, 2018).

3.1.1 Literature review

In the last 30 years, the field of disease mapping has experienced an enormous growth. This is because it is an important tool for decision makers to obtain reliable areal estimates of disease rates over a region of interest. Disease mapping methods, or ecological regression, further help understanding the underlying associations between covariates and the disease risk. Commonly, Bayesian hierarchical models are used to model the disease cases observed



(a) Map of the SMR

(b) Histogram of the SMR

Figure 3.1: Map and histogram of the SMR distribution for the Zika counts across the 160 neighbourhoods of Rio de Janeiro, between 2015 and 2016.

across the different areas that form a region of interest. The number of cases in an area is assumed to follow a Poisson distribution whose mean is decomposed as the product of an offset and the relative risk of the disease. Further, in the log scale, the relative risk is decomposed as the sum of covariates and latent (unobserved) areal effects. The latent components accommodate overdispersion as this decomposition of the log-relative risk can be seen as a Poisson-lognormal mixture model, if the latent effects follow a normal prior distribution.

Usually, these latent effects follow a spatial structure, *a priori*, such that neighbouring locations will adjust similarly after accounting for the available covariates. Indeed, it seems natural to expect that areas that are close to each other are more correlated than areas that are further apart. Let $\boldsymbol{b} = [b_1, \ldots, b_n]^\top$ be the vector of latent effects for the n areas of the region of interest. Different models have been proposed in the literature for the b's. First, a commonly used spatial model for the latent effects that does not accommodate outliers is the intrinsic conditional auto-regressive (ICAR) prior (Besag, 1974). Under the ICAR prior distribution, it is assumed that $b_i \mid \boldsymbol{b}_{(-i)}, \sigma_b^2 \sim \mathcal{N}\left((1/d_i)\sum_{j=1}^n w_{ij} b_j, \sigma_b^2/d_i\right)$, $i = 1, \ldots, n$, where $\boldsymbol{b}_{(-i)} = [b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n]^\top$, $\boldsymbol{W} = [w_{ij}]$ is a $n \times n$ matrix of weights, w_{ij}, that defines the neighbourhood structure and where $d_i = \sum_{j=1}^n w_{ij}$. Note that σ_b^2 is the variance parameter of the *conditional distribution* of b_i given its neighbours. It can be shown (Baner-

jee et al., 2014) that the joint distribution of $\boldsymbol{b}$ is proportional to $\exp\left[-(1/2\sigma_b^2)\boldsymbol{b}^\top \boldsymbol{Q}\boldsymbol{b}\right]$, with $\boldsymbol{Q} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D} = diag(d_i)$. The spatial weights are often set as $w_{ij} = 1$ if areas i and j share a border and $w_{ij} = 0$, otherwise. To ease the notation, let $\boldsymbol{b} \sim \text{ICAR}(\sigma_b, \boldsymbol{Q})$ denote the multivariate ICAR distribution. Using this common adjacency matrix, the joint ICAR distribution is not a proper multivariate normal distribution as the "precision" matrix, $\boldsymbol{Q}$, is not positive definite. One issue with the ICAR model is that it does not perform well when there is no underlying spatial structure in the data (Riebler et al., 2016).

To accommodate the presence of independent latent effects, Besag et al. (1991) proposed the so-called BYM model, where each areal latent effect is decomposed as the sum of an unstructured component and a spatially structured component. As pointed out by MacNab (2011), this model presents an identifiability issue as the two variance components cannot be distinguished. To avoid the introduction of two random effects for each area, like in the BYM model, Leroux et al. (1999) proposed an alternative distribution for the latent spatial effects that includes a spatial dependence parameter, λ. The latter is a mixing parameter in the unit interval that allows the variance of the latent effects to be decomposed into a weighted sum between an unstructured and a spatially structured variance components. On the other hand, regarding the BYM model, Sørbye and Rue (2014) argued that scaling the spatially structured effects is essential to ease interpretation and prior assignment of the variance parameter of the latent effects, independently of the neighbourhood structure. Hence, Riebler et al. (2016) proposed the BYM2 model, that decomposes the latent effects into a weighted sum of unstructured random noises with unit variance and scaled structured components. The vector of latent spatial effects is scaled according to the neighbourhood structure. This BYM2 model is a modification of the Dean model (Dean et al., 2001), which is itself a modification of the BYM model. In the BYM2 model, the decomposition of the ith latent effect is as follows:

$$b_i = \sigma_B \left(\sqrt{1-\lambda}\theta_i + \sqrt{\lambda} u_i^\star \right), \; i = 1, \ldots, n, \tag{3.1}$$

where $\lambda \in [0,1]$ and $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is independent of the scaled spatially structured components, $\boldsymbol{u}^\star = [u_1^\star, \ldots, u_n^\star]^\top \sim \text{ICAR}(1, \boldsymbol{Q}_\star)$. Let the matrix $\boldsymbol{Q}_\star^-$ be the generalised inverse of $\boldsymbol{Q}_\star$, which is a scaled version of the ICAR "precision" matrix, $\boldsymbol{Q}$: $\boldsymbol{Q}_\star = h\boldsymbol{Q}$. The scaling factor, h, is proportional to the generalised variance that arises from an ICAR model, $h = \exp\left[(1/n)\sum_{i=1}^n \ln\left(\boldsymbol{Q}_{ii}^-\right)\right]$. Note that the scaling factor only depends on the graph of the region under study. This scaled ICAR prior corresponds to $\boldsymbol{u}^\star = \left[u_1/\sqrt{h}, \ldots, u_n/\sqrt{h}\right]^\top$, for $\boldsymbol{u} \sim \text{ICAR}(1, \boldsymbol{Q})$. As stated in Sørbye and Rue (2014), this scaling process allows each structured component to have a variance of approximately 1. For further discussion on the scaling process, refer to section 3.2 of Riebler et al. (2016). It results that $\mathbb{V}(b_i \mid \sigma_B) = \sigma_B^2\left[(1-\lambda)\mathbb{V}(\theta_i) + \lambda\mathbb{V}(u_i^\star)\right] \simeq \sigma_B^2\left[(1-\lambda)\times 1 + \lambda \times 1\right] = \sigma_B^2$. Hence, a *marginal* variance, σ_B^2, is defined for the latent effects and all the parameters can be interpreted for all neighbourhood structures.

Spatial heteroscedasticity is not explicitly considered in the previous models. However, it is reasonable to imagine that some areas may have abnormally high or low disease risks. Richardson et al. (2004) emphasised the importance for disease mapping models to be able to differentiate and adapt between smoothing the risk surface and capture abrupt changes in relative risks. This issue of spatial heteroscedasticity has been increasingly considered over the recent years. For instance, regarding geostatistical data, Palacios and Steel (2006) proposed a log-normal scale mixture of a Gaussian process to accommodate heavy tails.

To allow for disparities, Congdon (2017) proposed a modification of the Leroux prior by including scale mixture parameters. More specifically, Congdon (2017) assumes

$$b_i \mid \boldsymbol{b}_{(-i)}, \boldsymbol{\kappa}, \lambda, \sigma_C^2 \sim \mathcal{N}\left(\frac{\lambda}{1-\lambda+\lambda d_i}\sum_{j=1}^n w_{ij}\kappa_j b_j, \frac{\sigma_C^2}{\kappa_i(1-\lambda+\lambda d_i)}\right), \; i = 1, \ldots, n, \qquad (3.2)$$

with $\kappa_i \overset{i.i.d.}{\sim} \text{Gamma}(\nu/2, \nu/2)$, $i = 1, \ldots, n$ and $\nu \sim \text{Exp}(1/\mu_\nu)$, for some value of μ_ν fixed by the analyst. These positive parameters, $\boldsymbol{\kappa} = [\kappa_1, \ldots, \kappa_n]^\top$, allow for discrepancies in the neighbouring estimated risks, while the usual CAR-type priors aim to locally smooth

the risk surface. The scale mixture parameters are termed outlier indicators as $\kappa < 1$ captures outliers. Again, for $\lambda \in (0,1)$, σ_C^2 is the variance parameter of the *conditional distribution* of b_i given its neighbours. This implies that the interpretation of σ_C^2 differs with every spatial structure, which renders its prior assignment not straightforward and makes interpretation difficult. It can be shown (Congdon, 2017) that the joint distribution of the latent effects is $\boldsymbol{b} \mid \sigma_C^2, \lambda, \boldsymbol{\kappa} \sim \mathcal{N}\left(\mathbf{0}, \sigma_C^2 \boldsymbol{Q}_C^-\right)$, where the "precision" matrix has diagonal elements $\boldsymbol{Q}_{C_{ii}} = \kappa_i(1 - \lambda + \lambda d_i)$ and off-diagonal elements $\boldsymbol{Q}_{C_{ij}} = -\lambda w_{ij}\kappa_i\kappa_j$. The diagonal dominance condition (Rue and Held, 2005) states that a sufficient condition for a symmetric matrix $\boldsymbol{Q}_C$ to be symmetric positive definite is $Q_{C_{ii}} > \sum_{j \neq i} |Q_{C_{ij}}|, \ \forall i$. Hence, it is sufficient that $\lambda \in [0,1)$ and $\lambda < \min_i \left\{1/\left(1 - d_i + \sum_{j \neq i} w_{ij}\kappa_j\right)\right\}$, for $\boldsymbol{Q}_C$ to be a valid precision matrix. Note that if $\boldsymbol{\kappa} = \mathbf{1}_n$, then Congdon's prior is the Leroux prior, which is proper for $\lambda \in [0,1)$. This mixture differs from the commonly used normal-gamma model, as the scale mixture components appear both in the mean and in the variance of the conditional distribution. Because the scale mixture components appear in the conditional mean, areas that share a border with an outlying area give this outlier a lower weight. Let neighbouring areas i and j be outliers, and let area k be a neighbour of i and not an outlier. Then, b_j contributes by a weight of $\kappa_j < 1$ to the conditional mean of b_i, whereas b_k contributes by a factor of $\kappa_k > \kappa_j$. This is a drawback when there are multiple outlying areas that are neighbours, as they will not borrow strength from each other.

Different from Congdon (2017), Dean et al. (2019) addressed local discrepancies by changing the neighbouring structure according to the observed data. This approach differs from Congdon's as it is a two-step procedure that implies changing the neighbourhood structure. Other models have been proposed to allow the strength of the spatial autocorrelation to vary over a region of interest. Corpas-Burgos and Martinez-Beneito (2020) proposed the so-called adaptive ICAR and adaptive Leroux models, which are modifications of the ICAR and Leroux models, by estimating the weights in the matrix $\boldsymbol{W}$. The adaptive Leroux model they proposed (CB-MB model) can be tied to Congdon's model (3.2). For $\lambda = 0$, Congdon's model

yields independent latent effects with variance divided by the scaling mixture component. Similarly, when $\lambda = 0$, the CB-MB model yields independent latent effects with variance divided by the spatial weight (see, e.g., Table A.6 in Appendix A.8). However, Corpas-Burgos and Martinez-Beneito point out that a single dataset is not enough to learn about those weights; so they suggest that their method is more suitable when modelling a multivariate outcome, where the neighbourhood structure is the same for the different outcomes. On the other hand, MacNab (2023) recently proposed a model that allows the spatial mixing parameter, λ, to change across space. This approach allows the underlying structure of the latent effects of the areas to differ from their neighbours, when necessary. The model proposed by MacNab differs from our proposal because it points out which structure, between the independent and spatially structured included in the BYM2 model, is more important for each region. The method proposed by MacNab does not allow for different variances across the region of interest, nor the identification of outlying areas.

The main aim of this paper is to propose a method to accommodate and identify outlying areas, following a single step inference procedure. We propose a modification of the BYM2 prior (3.1) that is able to identify outlying areas, after accounting for the effect of covariates. A scale mixture is introduced in the BYM2 model. The proposed model keeps the appealing property of parameter interpretation while capturing potentially outlying areas and allowing the neighbouring outlying areas to borrow strength from each other. Areas may be outliers with respect to the whole region of interest, namely areas with extreme disease risks; or with respect to their neighbours, termed spatial outliers. Throughout, the term "outlier" refers to both types of outliers: extremes and spatial outliers. This paper is organised as follows: Section 3.2 describes the proposed model, then a simulation study showcases the performance of the proposed model in section 3.3. Additionally, the application of the proposed model to the data presented in section 3.1 from the 2015-2016 Zika epidemic in the 160 neighbourhoods of Rio de Janeiro is shown in section 3.3. Section 3.4 concludes with a discussion.

3.2 Proposed model

Let a region of interest be partitioned into n non-intersecting areas. Let Y_i be the number of cases in area i, $i = 1, \ldots, n$, and E_i, the expected number at risk in that area. The counts are modelled through the following Poisson model:

$$Y_i \mid E_i, \mu_i \sim \mathcal{P}(E_i \mu_i),$$

where μ_i denotes the relative risk in area i and E_i is an offset. Commonly, the risk is decomposed in the log scale as follows:

$$\ln(\mu_i) = \beta_0 + \boldsymbol{x}_i \boldsymbol{\beta} + b_i,$$

where β_0 is the overall log risk, $\boldsymbol{x}_i$ is a p-dimensional vector with the explanatory variables in area i, associated with the p coefficients $\boldsymbol{\beta}$, and b_i is a random effect for area i. This latent effect is included in order to allow for overdispersion in the Poisson model that would otherwise assume equal mean and variance for area i. The latent areal effects can also accommodate an assumed underlying spatial structure in the data. To that end, a spatial structure is defined through the matrix $\boldsymbol{W} = [w_{ij}]$. Throughout this paper, we assume that two areas are said to be neighbours if they share a border. This implies that $w_{ij} = 1$ if areas i and j are neighbours and $w_{ij} = 0$, otherwise. In this setting, $d_i = \sum_{j=1}^{n} w_{ij}$ corresponds to the number of neighbours of area i. To model the latent areal effects accounting for such 0-1 spatial structure, we propose a modification of the BYM2 prior (3.1), that is, we assume

$$b_i = \frac{\sigma}{\sqrt{\kappa_i}} \left(\sqrt{1-\lambda}\theta_i + \sqrt{\lambda} u_i^\star \right), \quad i = 1, \ldots, n, \tag{3.3}$$

where $\sigma > 0$ is divided by the scaling mixture component $\kappa_i > 0$, and where $\lambda \in [0, 1]$. The component θ_i is assumed independent of $u_i^\star$. In particular, $\boldsymbol{\theta} \equiv [\theta_1, \ldots, \theta_n]^\top \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{u}^\star \equiv [u_1^\star, \ldots, u_n^\star]^\top \sim \text{ICAR}(1, \boldsymbol{Q}_\star)$. Components θ_i and $u_i^\star$ are termed the unstruc-

tured and the scaled structured component, respectively. Like in the BYM2 model (Riebler et al., 2016) (3.1), the "precision" matrix is such that $\boldsymbol{Q}_\star = h\boldsymbol{Q}$, where the scaling factor, h, is computed from the neighbourhood structure (see section 3.1.1). It results that, $\mathbb{V}(b_i \mid \sigma, \kappa_i) = (\sigma^2/\kappa_i)\left[(1-\lambda)\mathbb{V}(\theta_i) + \lambda\mathbb{V}(u_i^\star)\right] \simeq (\sigma^2/\kappa_i)\left[(1-\lambda)\times 1 + \lambda \times 1\right] = \sigma^2/\kappa_i$. Hence, σ^2/κ_i represents the approximate *marginal* variance of the ith area's latent effect. Moreover, the variance-covariance matrix, $\boldsymbol{V}$, of the proposed latent effects, $\boldsymbol{b}$, is given by $\boldsymbol{V} = \sigma^2\boldsymbol{K}^{-1}\times\ [(1-\lambda)\boldsymbol{I} + \lambda\boldsymbol{Q}_\star^-]$, where $\boldsymbol{K} = diag(\kappa_i)$. Thus, the parameter λ represents the weight of the spatial effect in the variance of the latent process. Note that this distribution is a proper multivariate normal for small values of λ, depending on the neighbourhood structure. Indeed, the diagonal dominance condition (Rue and Held, 2005) implies that it is sufficient that $\lambda \in [0,1)$ and $\lambda < \min_i \left\{1/\left(1 - \boldsymbol{Q}_{\star_{ii}}^- + \sum_{j\neq i}|\boldsymbol{Q}_{\star_{ij}}^-|\right)\right\}$ for the covariance matrix, $\boldsymbol{V}$, to be valid.

In a nutshell, the proposed model uses interpretable parameters to accommodate outlying areas while identifying them. The proposed model points at neighbourhoods that need heavy-tailed latent effects, through the introduction of the scale mixture components, $\boldsymbol{\kappa} = [\kappa_1, \ldots, \kappa_n]^\top$. Area i is identified as an outlier when $\kappa_i < 1$. Different from Congdon (3.2), the proposed model makes use of parameters that intervene on the *marginal* distribution of the latent effects. Therefore, their prior assignment is simplified as their interpretation remains the same regardless of the neighbourhood structure. This concerns the weight of the spatial structure λ, the marginal variance σ^2, as well as the scaling mixture parameters $\kappa_1, \ldots, \kappa_n$ when the κ's are assumed independent across the region.

We now compare the interpretation and roles of the scale mixture components $\boldsymbol{\kappa}$ in the proposed model and in Congdon's model. To interpret the scale mixture components $\boldsymbol{\kappa}$, the importance of the spatial structure in the data, measured by λ, must be taken into account. When $\lambda = 0$, both models reduce to independent latent effects without spatial structure. In

that case, $\kappa_i < 1$ only impacts the *marginal* variance of the ith latent effect and identifies an outlying area that showcases an extreme disease risk, after accounting for covariates. When $\lambda = 1$, the proposed latent effects become $b_i = (\sigma/\sqrt{\kappa_i})(u_i/\sqrt{h})$, $i = 1, \ldots, n$. The κ's intervene on the *marginal* variances and $\kappa_i < 1$ acts as an outlier indicator by inflating the ith *marginal* variance and hence allowing the ith effect to differ from the overall mean structure. Additionally, when $\lambda = 1$, the *conditional* distribution of the latent effects may be written as follows:

$$b_i \mid \boldsymbol{b}_{(-i)}, \sigma^2, \boldsymbol{\kappa} \sim \mathcal{N}\left(\frac{1}{d_i}\sum_{j=1}^{n} w_{ij}\sqrt{\frac{\kappa_j}{\kappa_i}}b_j, \frac{\sigma^2/h}{\kappa_i d_i}\right), \quad i = 1, \ldots, n. \tag{3.4}$$

We compare the *conditional* distributions (3.2) and (3.4) considering the case where neighbouring areas i and j are both outliers with $\kappa_i, \kappa_j < 1$ and $i \sim j$. In both distributions (3.2) and (3.4), the ith and jth *conditional* variances are inflated by κ_i and κ_j, respectively. Regarding the *conditional* means, in the proposed model, $\kappa_j/\kappa_i \simeq \kappa_i/\kappa_j \simeq 1$ and outlying effects are allowed to borrow strength from neighbouring outliers. However, in Congdon's model, the mutual weights of b_i and b_j are deflated and areas i and j contribute less to their mutual latent effects. This feature of borrowing strength in the proposed model is attractive in the case where neighbouring areas have extreme disease risks.

In the next subsection, different prior distributions are discussed for the scale mixture components.

3.2.1 Prior specification of the scale mixture component

A natural choice, and used by Congdon (2017), is to assume:

$$\kappa_i \overset{i.i.d.}{\sim} \text{Gamma}\left(\nu/2, \nu/2\right), \; i = 1, \ldots, n, \quad \text{and} \quad \nu \sim \text{Exp}(1/\mu_\nu), \tag{3.5}$$

where the hyperparameter's mean μ_ν controls the magnitude of ν. When $\lambda = 0$, marginalising the proposed distribution (3.3) of the latent effect, b_i, with respect to κ_i yields a Student-t distribution with μ_ν degrees of freedom, that is t_{μ_ν}. The introduction of κ_i hence allows for heavier tails than a Gaussian distribution for the latent effects. In this case, μ_ν corresponds to choosing the degrees of freedom of the resulting t distribution, which impact the moments of the distribution as well as its tails. A large μ_ν results in a distribution close to being normal, which is inadequate to capture outliers. On the other hand, $\mu_\nu < 3$ implies a t distribution whose variance is not defined. Some simulation studies showed that setting $\mu_\nu = 4$ performed well, which is the value suggested by Gelman et al. (2004).

Another possible prior specification for the κ's is to borrow ideas from Palacios and Steel (2006) who proposed the inclusion of a scale mixture component in the variance of a Gaussian process. The authors suggest the usual gamma mixing is not always appropriate, as not all positive moments exist. Additionally, they point out that the t distribution that results from marginalising over the gamma scaling mixture parameters may still overestimate the overall variance and struggle to detect specific outlying areas. In particular, they assume that the scale mixture component follows a log-Gaussian process with the same spatial structure as the one defined for the main Gaussian process. Here, we propose a scaled log proper CAR prior distribution for the κ's. This form of discretisation of the method proposed by Palacios and Steel (2006) is applied to the latent effects, b_i, $i = 1, \ldots, n$, which include both the structured and unstructured components, in order to keep the interpretative property of the parameters. This contrasts with the method proposed by Palacios and Steel (2006) as they introduced a scale mixture only for the spatially dependent components, leaving the unstructured components untouched. Let the scale mixture components be modelled as follows:

$$\begin{gathered} \ln(\kappa_i) \equiv -\frac{\nu_\kappa}{2} + z_i, \;\; i = 1, \ldots, n, \\ \text{where} \quad \boldsymbol{z} \equiv [z_1, \ldots, z_n]^\top \mid \nu_\kappa \sim \mathcal{N}\left(\mathbf{0}, \nu_\kappa \boldsymbol{Q}_{\alpha,\star}^{-1}\right) \quad \text{and} \quad \nu_\kappa \sim \text{Exp}(1/\mu_{\nu_\kappa}), \end{gathered} \tag{3.6}$$

where $\boldsymbol{Q}_{\alpha,\star} = h\boldsymbol{Q}_\alpha = h_\alpha[\boldsymbol{D} - \alpha\boldsymbol{W}]$ is again a precision matrix that is scaled by h_α, which is computed based on $\boldsymbol{D} - \alpha\boldsymbol{W}$. The parameter α guarantees $\boldsymbol{Q}_\alpha$ to be a valid precision matrix for $\alpha \in [0, 1)$ (Banerjee et al., 2014). For this proper distribution to be close to an ICAR prior, we impose $\alpha = 0.99$. The proper CAR distribution is scaled in order to approximately have that $\mathbb{V}[\ln(\kappa_i) \mid \nu_\kappa] \simeq \nu_\kappa \times 1$. Similarly to Palacios and Steel (2006), this prior implies $\mathbb{E}(\kappa_i \mid \nu_\kappa) \simeq 1$, which corresponds to a constant marginal variance across the areal latent effects, and $\mathbb{V}(\kappa_i \mid \nu_\kappa) = [\exp(\mathbb{V}(\ln(\kappa_i \mid \nu_\kappa)) - 1)] \exp(2\mathbb{E}(\ln(\kappa_i \mid \nu_\kappa)) + \mathbb{V}(\ln(\kappa_i \mid \nu_\kappa))) \simeq [\exp(\nu_\kappa) - 1] \exp(-\nu_\kappa + \nu_\kappa) = \exp(\nu_\kappa) - 1, \ \forall i$. For ν_κ close to 0, κ is close to 1 with a small variance. A bigger ν_κ allows the κ's to differ greatly from 1 and to be closer to 0, when necessary. Palacios and Steel (2006) suggest that a reasonable prior mean for ν_κ is $\mu_{\nu_\kappa} = 0.2$. The simulation studies we conducted suggest that a sensible choice for μ_{ν_κ} is $\mu_{\nu_\kappa} = 0.3$, which yields $[0.2, 2.4]$ as the 95% prior credible interval for the κ's. This includes, $\kappa_i = 1$ while allowing for departure from $\kappa_i = 1$, to accommodate the potentially outlying random effect of area i. This prior specification for the κ's allows the mixture components to borrow strength from neighbouring κ's. This may be of particular interest when outlying areas are neighbours.

3.2.2 Inference procedure

Following the specifications discussed in the previous section, the resultant posterior distributions, regardless of the prior specification for κ_i, do not have a closed analytical form. Therefore, the posterior distributions are approximated through computational methods. In particular, Markov Chain Monte Carlo (MCMC) methods are considered. The Hamiltonian Monte Carlo method implemented in the R package `rstan` (Stan Development Team, 2020) is used for the simulation studies and real data application that follow. Morris et al. (2019) note that the No U-Turn Sampler implemented in `rstan` is more efficient than other MCMC samplers to obtain reliable estimates of the posterior distributions induced by the complex autoregressive type of models that are of interest in this paper.

One way to approximate a proper posterior distribution when assigning an ICAR prior, is to add a sum-to-zero constraint on the parameters in order to distinguish them from any added constant. This is necessary due to the invariance of the ICAR distribution to the addition of a constant (Rue and Held, 2005). The sum-to-zero constraint is applied to the spatial components of the proposed model, $\boldsymbol{u}$, that need to be distinguished from the global intercept, β_0. More precisely, we add a soft sum-to-zero constraint, that is $\sum_{i=1}^{n} u_i \sim \mathcal{N}(0, (n/1000)^2)$. The `rstan` implementation of the BYM2 model is discussed by Morris et al. (2019) and the code for the proposed model, which is a modification of the BYM2, is available in Appendix A.1.

The scaling factor, h, needed in the BYM2 and in the proposed model, is computed through the R package `R-INLA` (Integrated Laplace Approximation, Rue et al. (2009), `www.r-inla.org`) as explained by Riebler et al. (2016).

3.3 Data analyses

In this section, we present the results of a simulation study that was conducted to assess the performance of the proposed model. The results from fitting the proposed model to data obtained from the first Zika epidemic that took place between 2015 and 2016 in Rio de Janeiro are also shown. In both cases, we consider the two parametrisations of the proposed model, which correspond to the two prior specifications of the scaling mixture components described in section 3.2.1. In the simulation study and in the data application, the proposed model is compared to Congdon's model (Congdon, 2017). Out of completeness, we also consider the two prior specifications for the κ's for Congdon's model. Namely, Congdon's model is fitted with the κ's following the original independent prior Gamma distributions (3.5), as well as with spatially structured κ's (3.6).

In the simulation study, we generate data for the 96 French departments and contaminate some areas. The goal is to check whether our proposed model is able to identify the generated

outliers. Then, in the Zika data analysis in Rio de Janeiro, we compare the results of our proposed model to Congdon's as well as the BYM2 (Riebler et al., 2016) and Leroux (Leroux et al., 1999) models. We identify some potentially outlying districts which might be of interest to decision makers.

3.3.1 Simulation study: neighbouring outliers in France

In this section, we present the results from a simulation study wherein some arbitrary neighbouring areas in France are contaminated into outlying areas, to assess the performance of the proposed model in comparison to the one proposed by Congdon. The design of the simulation study is inspired by Richardson et al. (2004), where the goal is to assess the ability of the proposed model to both smooth over non-contaminated areas while capturing and identifying the contaminated ones. Richardson et al. (2004) emphasised the importance for disease mapping models to adapt to these abrupt changes in the risk surface.

In this simulation study, 20 departments are contaminated such that 2 groups of 10 neighbouring outliers are created. Out of simplicity, there are no covariates included in the generating process nor when fitting the models. First, all $n = 96$ latent effects, which correspond to log relative risks in this covariate-free simulation study, are set to 0: $b_i = 0,\ i = 1, \ldots, n$. Then, the offsets $[E_1, \ldots, E_n]^\top$ are computed based on the 2019 department size estimates available on the Institut National de la Statistique et des Études Économiques (INSEE) website (`https://statistiques-locales.insee.fr/#c=indicator`). We define five offset categories based on the empirical offset quantiles. The first category corresponds to the smallest offsets and the fifth category, to the largest ones. The categories are termed "Small" for $E \leq 568$, "Medium low" for $E \in (568, 906]$, "Medium" for $E \in (906, 1428]$, "Medium high" for $E \in (1428, 2399]$ and "High" for $E > 2399$. Based on these categories, we select 20 departments to be outliers, such that each group of 10 neighbouring outliers contains 2 areas of each offset category. Within each such pair of departments, the relative risks are contaminated into outliers by setting $b_i = \ln(0.5)$ and $b_{i'} = \ln(1.5)$. The resulting outliers

are mapped in the left panel of Figure 3.2, highlighting the offset sizes and imposed relative risks. Finally, $R = 100$ populations of size $n = 96$ are created according to a hierarchical Poisson model, that is, $Y_i \sim \mathcal{P}(E_i \exp[b_i])$. The only source of randomness across the 100 replicates comes from the repeated sampling from a Poisson distribution.

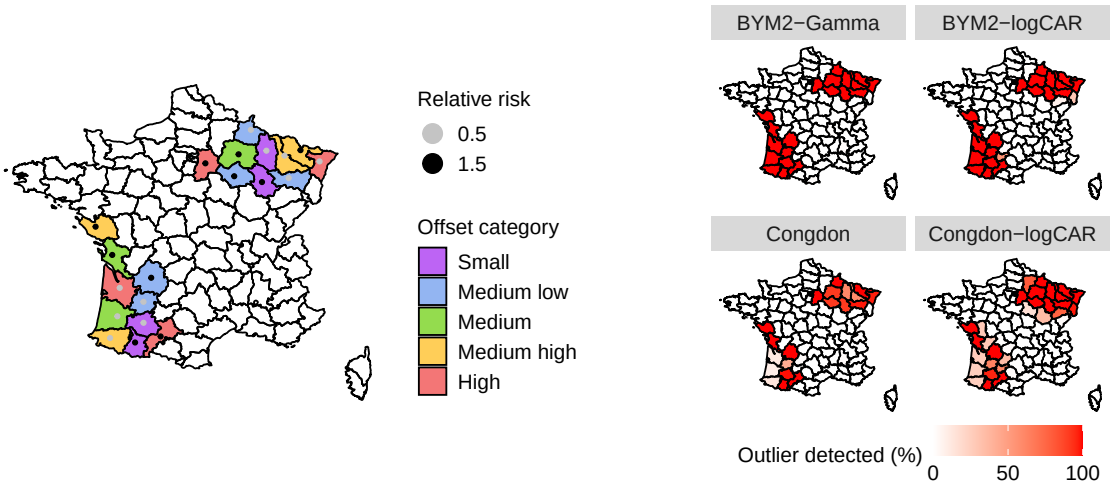


Figure 3.2: Left panel: French departments arbitrarily chosen to be outliers in the simulation study. Colours depict the offset category based on the empirical offset quantiles. The points represent the relative risk set to each outlying district. Right panel: Percentage of times among 100 replicates that the outliers were identified by each model, in the simulation study. The outliers are pointed out when $\kappa_u < 1$, where κ_u is the upper bound of the posterior 95% credible interval of κ.

Using the two scale mixtures described in section 3.2.1, the Congdon model is compared to the proposed model. The first version of the proposed model is denoted BYM2-Gamma and the second, BYM2-logCAR. The original Congdon model is termed Congdon, whereas the one with spatially structured scale mixture components is denoted Congdon-logCAR. For the four models, the intercept is given a quite vague prior: $\beta_0 \sim \mathcal{N}(0, 10^2)$ and the mixing parameter, λ, is assigned a uniform, $\mathcal{U}(0, 1)$, prior distribution. The same $\mathcal{N}_+(0, 1)$ prior is considered for σ, which is a *marginal* standard deviation in the proposed model, while it is a *conditional* standard deviation in Congdon's. Finally, in the BYM2-Gamma and Congdon models, the prior distribution for the κ's is described in (3.5) with $\nu \sim \text{Exp}(1/4)$. For the BYM2-logCAR and Congdon-logCAR parametrisations, the κ's follow *a priori* the

distribution in (3.6) and we set $\nu \sim \text{Exp}(1/0.3)$.

The models are fitted through the R package `rstan` (Stan Development Team, 2020). For each dataset, the MCMC procedure consists of 2 chains of 20,000 iterations with a 10,000 burn-in period and a thinning factor of 10. Convergence of the chains is assessed through trace plots, effective sample sizes and the $\widehat{R}$ statistic (Gelman and Rubin, 1992; Vehtari et al., 2021).

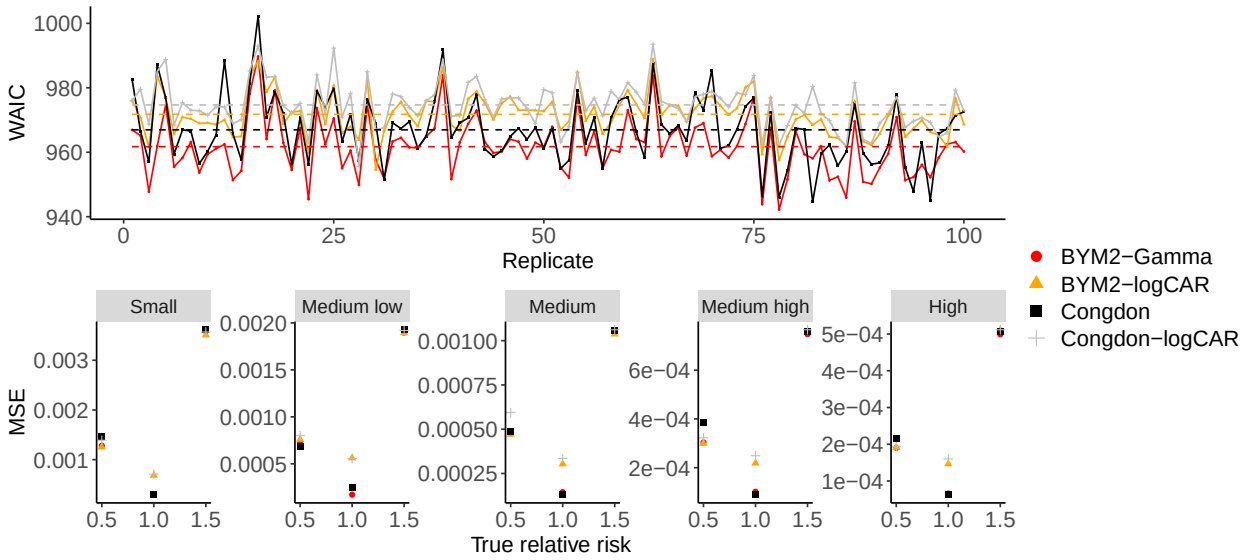


Figure 3.3: Top panel: WAIC across the 100 replicates for the proposed models and Congdon's, in the simulation study. Dashed lines: mean WAIC for each model. Bottom panel: MSE over the 100 replicates for the proposed models and Congdon's according to the true relative risk and the offset size, in the second simulation study.

In terms of WAIC (Watanabe and Opper, 2010), for which smaller values are preferred, the proposed BYM2-Gamma model yields the smallest value among the four models, as shown in Figure 3.3, with an average WAIC of 962 versus 967, 972 and 975 for Congdon, BYM2-logCAR and Congdon-logCAR, respectively. In terms of MSE, Figure 3.3 shows that all models perform similarly: on average over the 100 replicates and all areas, the BYM2-Gamma's MSE is 0.0003, versus 0.0004 for Congdon and 0.0005 for both models with the logCAR parametrisation.

Regarding the detection of outliers, Table 3.1 and the right panel of Figure 3.2 show how often each model accurately detects departments as outliers (sensitivity) and non-outliers (specificity), depending on the offset category. That is, the sensitivity is equal to the percentage of outliers detected among the contaminated departments over the 100 replicates. The specificity is the percentage of departments not identified as outliers among the ones whose true relative risk is equal to 1, over the 100 replicates. The definition for sensitivity and specificity are taken from Richardson et al. (2004). Area i is detected as an outlier when $\kappa_{u,i} < 1$, where $\kappa_{u,i}$ is the upper bound of the 95% posterior credible interval of κ_i. Congdon's model with spatially structured κ's tends to identify more outliers than truly present in the data (overall specificity of 93%, versus 99.9% for both BYM2-Gamma and Congdon, and 98.7 for BYM2-logCAR). More importantly, while both parametrisations of the proposed model always identify all the contaminated areas, overall, the two versions of Congdon's model miss 22% and 13% of the outliers. That is, the proposed spatially structured prior for the κ's allows Congdon's model to identify 10% more outliers than the model with independent mixture components.

	Offset category	BYM2-Gamma	BYM2-logCAR	Congdon	Congdon-logCAR
Sensitivity	Small	100.0	100.0	87.7	99.0
	Medium low	100.0	100.0	86.4	92.6
	Medium	100.0	100.0	66.7	75.0
	Medium high	100.0	100.0	68.0	81.2
	High	100.0	100.0	77.0	81.7
	Overall	100.0	100.0	78.1	86.8
Specificity	Small	100.0	99.2	99.9	89.2
	Medium low	99.9	96.1	99.9	90.1
	Medium	99.7	99.9	99.9	92.6
	Medium high	99.9	98.1	100.0	93.5
	High	100.0	100.0	100.0	100.0
	Overall	99.9	98.7	99.9	93.1

Table 3.1: Sensitivity and specificity of the outlier detection for each model, depending on the offset size, in the simulation study.

Further simulation studies

To further assess the performance of the proposed model, other simulation studies were

conducted. In Appendices A.3 and A.4, two simulation studies show the ability of the two versions of the proposed model to recover the true parameters when data are generated from the model itself. This suggests that the proposed model does not suffer from identifiability issues. In particular, the proposed model is able to identify and distinguish, for each district, the outlier indicators, the spatial components and the unstructured components, individually. Appendix A.5 presents a simulation study without contaminating any areas into outliers, which results in the proposed model performing well compared to the prior by Congdon (2017), in terms of WAIC and in terms of outlier detection, where Congdon's model wrongly identifies non-outlying areas as outliers. Appendix A.6 presents the results from a simulation study where arbitrary distant areas in France are contaminated into outliers. Again, the goal is to assess the ability of the proposed model to identify these outliers. As discussed in Section 3.2, in that scenario where outliers are far from each other, the proposed model performs similarly to Congdon's model. To show that the performance of the proposed model is independent of the neighbourhood structure under study, we present in Appendix A.7 the results from two simulation studies that use the map of Rio de Janeiro, where some districts are contaminated into outliers. A third simulation study shown in Appendix A.7.3 aims to resemble the data analysis presented in Section 3.3.2, wherein a covariate is included, and relative risks vary more over the region of interest. We found that the proposed model performed better in identifying the outliers, compared to Congdon's model.

3.3.2 Cases of Zika during the 2015-2016 epidemic in Rio de Janeiro

The total numbers of cases of Zika were recorded across the 160 neighbourhoods of Rio de Janeiro during the first epidemic, which took place between 2015 and 2016. Let Y_i be the disease count in district $i = 1, \ldots, 160$. A hierarchical Poisson model is fitted to these data with offsets, E, computed from, P, the areal population sizes, $E_i = P_i \left(\sum_j Y_j / \sum_j P_j \right)$. We consider a socio-development index, x, as an explanatory variable for the number of cases. Identifying districts with potentially outlying risks, after accounting for the covariate, may

be useful for decision makers to understand how to prevent Zika and where to start from. The distribution of Zika is described through a map and a histogram of the standardised morbidity ratio (SMR), Y/E, in Figure 3.1 in section 3.1. Some districts seem to present different SMR values than the mean surface, such as the island Paquetá, Barra de Guaratiba and Pedra de Guaratiba, with SMRs of 7.3, 6.5 and 5.9, respectively. In the lower tail of the SMR distribution, three districts did not record any cases and thus present null SMRs, namely Gericinó, Vasco da Gama and Parque Colúmbia. However, the SMR being an exploratory tool, one cannot conclude that high or low SMR values necessarily indicate outlying districts. Therefore, we are interested in comparing which districts are identified as potential outliers, after accounting for the socio-development index, by the two versions of the proposed model and Congdon's. The same priors are defined for the parameters as in the simulation study presented in section 3.3.1 and the two versions of the proposed model and Congdon's are again denoted BYM2-Gamma, BYM2-logCAR, Congdon and Congdon-logCAR. We further compare the performance of the four models to the BYM2 and Leroux models which do not accommodate potential outliers.

All models are fitted in `rstan` (Stan Development Team, 2020) with 2 chains of 20,000 iterations thinned by 10 and of which 10,000 are burnt. As assessed by the trace plots, the effective sample sizes and the $\widehat{R}$ statistics, the two chains have mixed well for all six models and convergence is attained. Appendix A.2 presents the trace plots, effective sample sizes and $\widehat{R}$ statistics for a selection of parameters from the two parametrisations of the proposed model. The proposed BYM2-Gamma model took 15 minutes to run while the proposed BYM2-logCAR needed 11 minutes. In comparison, Congdon's model converged in 22 minutes and the Congdon-logCAR, in 11 minutes.

The results from the fitted models are presented in Table 3.2 and Figure 3.4. In terms of WAIC, the proposed BYM2-Gamma model performs best among the six considered. There is an important performance gain when accommodating outliers (BYM2-Gamma, BYM2-

logCAR, Congdon and Congdon-logCAR: 1335, 1342, 1337 and 1339, respectively, vs BYM2 and Leroux: 1371 and 1374, respectively). Congdon's prior does not seem to perform significantly worse than the BYM2-Gamma model. Interestingly, even though the proposed model has 160 more parameters than Congdon's, its effective number of parameters is similar (80 vs 81). The models are further compared in terms of MSE, where MSE $= (1/N)\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2$, where $\widehat{Y}_i$ is the fitted value, that is, the estimated mean of the posterior predictive distribution. All models yield similar values, between 243.5, for the Congdon-logCAR model, and 245.7 for the Leroux model.

Regarding the intercept, β_0, the proposed models and Congdon's give similar results, whereas the Leroux and BYM2 models yield smaller posterior means and lower credible interval bounds. This is probably due to the difference in the spatial effects that are allowed to be more extreme in the Congdon, Congdon-logCAR, BYM2-Gamma and BYM2-logCAR models. All six models indicate a negative relationship between the development index and the risk of Zika, with negative posterior 95% credible intervals for β that do not include 0. We cannot directly compare the parameters λ and σ between the BYM2-type models and Leroux-type priors, as these lie in the *marginal* and *conditional* distributions of the latent effects, respectively. Marginally, the BYM2-type models yield similar weights of the spatially structured components on the latent effects (posterior means for λ of 0.6 and 0.7). For the Leroux-type models, the point estimates for λ show slightly more difference (e.g. 0.6 for Leroux and 0.8 for Congdon). This difference may be due to the presence of outliers in the data, which results in the Leroux model finding more random noise in the latent effects. The same observation can be made for the marginal and conditional standard deviation, σ, regarding the BYM2-type models and the Leroux-type models, respectively. The posterior credible interval for σ is significantly higher in the BYM2 model compared to the two parametrisations of the proposed model, and in the Leroux model compared to the two versions of Congdon's model. Indeed, the proposed models are able to estimate a smaller overall variance for the latent effects, which is then adjusted through the κ's when needed.

Finally, it can be noted that there seems to be enough information in the data to learn about the hyperparameter ν. This parameter was assigned a prior mean of 4 and prior 95% credible interval of $[0.1, 14.7]$ for the BYM2-Gamma and Congdon models and resulted in posterior means of about 2 and posterior 95% credible intervals of about $[1, 3]$. The BYM2-logCAR and Congdon-logCAR models assigned an exponential distribution with mean 0.3 for ν, inducing a prior 95% credible interval of $[0.0, 1.1]$, and yielded posterior credible intervals of $[0.7, 2.3]$ and $[0.9, 2.9]$, showing the need for some κ's to be different from 1, *a posteriori*.

	BYM2	BYM2-logCAR	BYM2-Gamma	Congdon	Congdon-logCAR	Leroux
Model fit						
WAIC	1371.2	1342.3	1335.6	1337.5	1339.2	1373.9
p_W	88.6	82.3	80.0	81.0	81.1	89.2
MSE	244.8	243.7	244.1	244.3	243.5	245.7
Parameters' posterior summaries						
	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)
β_0	1.6 (0.4,2.8)	2.5 (1.3,3.5)	2.5 (1.7,3.4)	2.4 (1.4,3.2)	2.0 (1.0,3.0)	1.2 (-0.1,2.4)
β	-2.8 (-4.8,-0.8)	-4.2 (-5.8,-2.3)	-4.3 (-5.6,-2.9)	-4.0 (-5.4,-2.6)	-3.7 (-5.1,-1.9)	-1.9 (-4.1,-0.1)
λ	0.7 (0.4,0.9)	0.6 (0.2,0.9)	0.7 (0.3,0.9)	0.8 (0.5,0.9)	0.6 (0.2,0.9)	0.6 (0.2,0.9)
σ	0.8 (0.7,0.9)	0.4 (0.3,0.5)	0.4 (0.3,0.5)	0.6 (0.4,0.8)	0.6 (0.4,0.8)	1.2 (0.9,1.5)
ν	-	1.4 (0.7,2.3)	2.2 (1.4,3.3)	1.9 (1.3,2.8)	1.7 (0.9,2.9)	-

Table 3.2: Results from the analysis of Zika reported cases in Rio de Janeiro in 2015-2016. Model assessment (WAIC) and parameter posterior summaries: posterior mean and 95% credible interval (CI) for BYM2, BYM2-logCAR, BYM2-Gamma, Congdon and Leroux.

We now focus on the outliers detected by the proposed models and Congdon's, as shown in Figure 3.4. District i is again found to be a potential outlier, after accounting for the socio-development index, if $\kappa_{u,i}$, the upper bound of the posterior 95% credible interval of κ_i, is below 1. In Figure 3.4, the blue and red coloured districts help distinguish the detected outliers on the lower tail of the SMR distribution from the ones on the upper tail. After accounting for the socio-development index, some districts are pointed out by the four models, such as Gericinó, Parque Colúmbia, Vasco da Gama and Maré, on the lower tail of the SMR distribution, Barra de Guaratiba and Bonsucesso, on the upper tail. However,

Congdon's model and both versions of the proposed approach do not point out Paquetá in the upper tail, whereas the Congdon-logCAR model detects it. This may be explained by the offset size of Paquetá, which is among the smallest in the entire region of Rio. Note, however, that the BYM2-Gamma model is close to identifying Paquetá as an outlier as it results in $\kappa_u = 1.04$ for this district. Neither of the four models identify Pedra de Guaratiba, which has a high SMR, as shown in Figure 3.1. Interestingly, the district of São Cristóvão is detected as an outlier by all models except the BYM2-Gamma model, with $\kappa_u = 1.2$. The Congdon models detect few more potential outliers than both versions of the proposed model. Our simulations have shown that the Congdon models tend to detect non-outliers more often than the BYM2-Gamma model. We believe that this explains the differences in the outliers identified after accounting for the socio-development index.

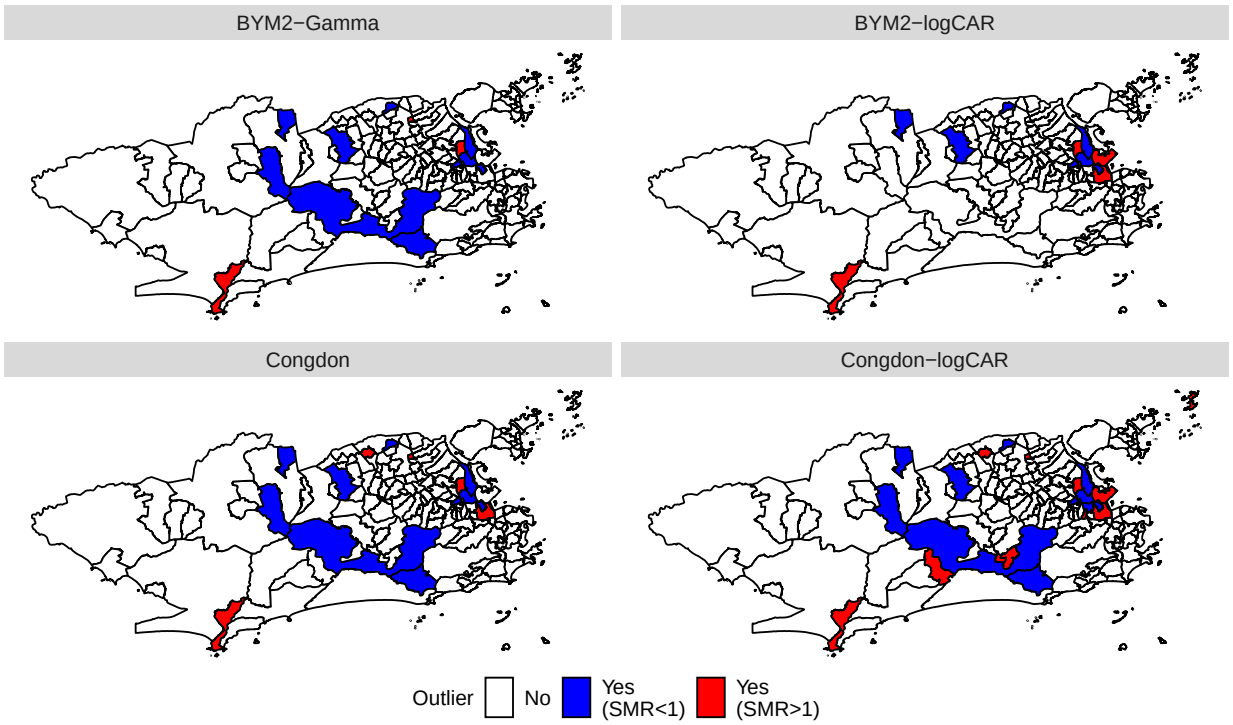


Figure 3.4: Maps of the outliers indicated by each model when analysing the Zika counts. The outliers are pointed out when $\kappa_u < 1$, where κ_u is the upper bound of the posterior 95% credible interval of κ. The outliers on the lower tail are distinguished from the ones on the upper tail of the SMR distribution.

3.4 Discussion

In this paper, we propose a disease mapping model that is able to identify areas with potentially outlying disease risks, after accounting for the effects of covariates. Outliers refer to areas with extreme risks - on the tail of the risk distribution - as well as spatial outliers, after accounting for covariates. Spatial outliers correspond to areas whose risk differs from their neighbours, after accounting for covariates. The proposed model is a scale mixture of the BYM2 model (Riebler et al., 2016). Two different prior specifications are proposed for the scale mixture components in order to compare independent components and spatially structured components. Our model allows for a straightforward interpretation of the parameters, that is common to every data application, while accommodating outliers. The parameters' interpretation is eased by the scaling process of the latent spatially structured components (Sørbye and Rue, 2014).

A simulation study presents the performance of the two versions of the proposed model compared to the one by Congdon (2017), as well as a version of Congdon's model that uses our proposed spatially structured mixture components. The neighbourhood structure of France is used and the latent effects of some neighbouring departments are contaminated to control the presence of outliers. The BYM2-Gamma version of the proposed model always performs best in terms of WAIC and in terms of MSE. Regarding the detection of outliers, the two versions of the proposed model always identify the contaminated departments, compared to the two parametrisations of Congdon's model that miss up to 33% of the outliers. Additionally, the BYM2-Gamma version of the proposed model does not detect non-contaminated districts. Finally, in all of our simulation studies, the proposed model always performs at least as well as Congdon's, and often better, both in terms of WAIC, MSE and of outlier identification (see, e.g., Appendices A.6, A.7).

The cases of Zika that were recorded in Rio de Janeiro during the first 2015-2016 epidemic are analysed using the two parametrisations of the proposed model as well as the model by

Congdon (2017) and its version with spatially structured mixture components, the BYM2 (Riebler et al., 2016) and the Leroux prior (Leroux et al., 1999). All six models find that there is a fairly strong negative association between the socio-development index and the number of cases, meaning that richer districts have lower disease risks. This finding is consistent with previous studies conducted in Rio de Janeiro, one investigating the first chikungunya epidemic in the city (Freitas et al., 2021) and another also investigating Zika, but using a different methodological approach (Raymundo and de Andrade Medronho, 2021). These studies, including ours, indicate that improving sanitary conditions and reducing socio-economic disparities are of paramount importance to fight *Aedes*-borne diseases.

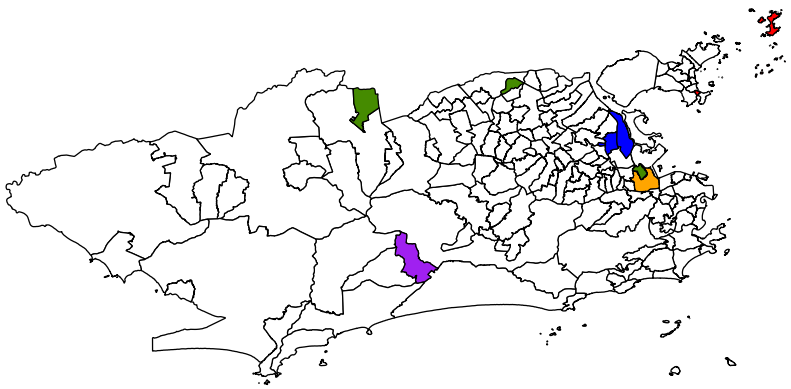


Figure 3.5: Map highlighting some districts identified as outliers by at least one model when analysing the Zika counts. Orange: São Cristóvão; Red: districts with small offsets; Blue: districts whose population sizes increased significantly after the 2010 census; Purple: districts combining both characteristics; Green: districts with zero cases recorded.

After accounting for the effect of the socio-development index, some neighbourhoods are detected as potential outliers by the proposed models and Congdon's, both in the lower and upper tails of the number of cases' distribution across the districts. Out of the 23 neighbourhoods identified as outliers, irrespective of the model, the proposed models BYM2-logCAR and BYM2-Gamma identified 11 (47.8%) and 14 (60.9%), respectively. The four models do not always point out the same districts as potential outliers. One possible explanation for that is the small offset sizes of some districts. The simulation study with neighbouring

outliers showed that, when the offset is small, the models that impose a spatially structured prior on the scaling mixture components tend to accurately identify outlying areas more often than the models with *a priori* independent mixture components. Regarding the analysis of Zika cases, Figure 3.5 shows in red and purple the districts identified as outliers by at least one of the four models and whose offsets are among the smaller 5%. For example, based on the results from the second simulation study, it is possible that, when analysing the Zika counts, Camorim (purple) and the island Paquetá (red) are missed by the BYM2-Gamma and Congdon models while they are pointed out by the Congdon-logCAR model (Figure 3.4) because of their smaller offset sizes (Figure 3.5).

Figure 3.5 highlights in green the districts with zero Zika cases recorded between 2015-2016: Parque Colúmbia, Gericinó and Vasco da Gama. These 3 districts are pointed out as outliers by the four models, as shown in Figure 3.4. One potential explanation for these zero recorded cases is that when the disease appeared for the first time in 2015, it was not immediately identified as Zika. Further, there is evidence that epidemics in Rio de Janeiro tend to spread starting from the north-east of the city (Freitas et al., 2019). It is then possible that when the authorities began registering the Zika cases, there were no cases to record in the two northern districts highlighted in blue, Parque Colúmbia and Gericinó. Another potential reason is that it is not uncommon in Rio de Janeiro for a person to report as their neighbourhood of residence a neighbourhood that actually shares a border with the one where they actually live. For instance, Parque Colúmbia and Gericinó are relatively new districts and the population might not yet be used to naming them as their districts of residence. Similarly, a person living in Vasco da Gama (southern green district) may report São Cristóvão (orange) as their district. This would artificially cause Vasco da Gama to record zero cases and be detected as a potential outlier. Further, if a given district is accounting for a proportion of the cases that are in fact from the neighbouring areas (e.g., São Cristóvão), this would artificially increase the risk of this district. In fact, Figure 3.4 shows that São Cristóvão is pointed out as a potential outlier by all models but the BYM2-

Gamma. Therefore, the inaccurate information on the district of residency may artificially create outliers.

Finally, artificial outliers may be caused by inaccurate information on the areal population sizes used to compute the offsets. While the disease counts were recorded during 2015-2016, the population sizes were extracted from the previous census, dating from 2010. Between 2010 and 2015-2016, the population sizes may have increased in some districts, without being reflected in the offsets in this analysis, causing the artificial detection of increased disease risks. Figure 3.5 highlights in blue and purple the districts identified as potential outliers and whose sizes have largely increased since 2010, according to more recent aerophotogrammetry flights by the Health Secretariat of the city. The eastern blue districts are pointed out as outliers by all four models in Figure 3.4. Further investigating these districts would help determine whether they do present outlying disease risks or if they are artificial outliers. An interesting side effect of the proposed model seems to be that by identifying outliers and further investigating the results, the authorities might better understand the population dynamics in the region of interest, in between censuses, and identifying potential issues in the accurate recording of cases.

Therefore, we suggest exploring both prior specifications for the scaling mixture components, using the proposed model and Congdon's, and further investigation on the detected districts should be conducted by decision makers and experts to fully comprehend the detected outlying behaviours. Also, it is important to emphasize that some socio-environmental factors that influence the burden and distribution of *Aedes*-borne diseases may be heterogeneous within the districts, our spatial unit of analysis. For example, the same district may have areas with *favelas* (slums) and areas with middle and upper class condominiums. The socio-development index will not capture this intra-district social inequality, and a recent study showed evidence about the presence of socio-economic inequalities in the distribution of dengue, Zika and chikungunya in two Latin American cities (Carabali et al., 2020). Another

possibility is the presence of large potential breeding sites, such as dumps and vacant lots. It is also worth mentioning that spatial confounding, which is beyond the scope of this work, is a potential issue that may affect the estimated latent effects (Dupont et al., 2022; Urdangarin et al., 2023) and identified potential outliers. Hence, interpretation of the results should be done with care.

To conclude, we believe our proposed model to be useful to decision makers. First, the parameters' interpretation eases the use of our model regardless of the data spatial structure. This may help decision makers to create a systematic procedure to analyse data with our proposed model, in which non-informative priors for the parameters could be defined for any spatial structure. Then, the introduction of scaling mixture components improves the recovering of the observed and potentially outlying disease risks, as assessed by the model performance criteria (WAIC and MSE). Finally, these mixture components together with high estimated risk ratios help identify all the potential outlying areas in which interventions may need to be prioritised.

Acknowledgements

Michal was partially supported from an award from the Fonds de Recherche Nature et Technologies (B2X - 314857). Schmidt and Cruz are grateful to IVADO (Fundamental Research Project, PRF-2019-6839748021).

Data Availability Statement

The Zika and the population data analysed in this study come from the Brazilian Notifiable Diseases Information System (SINAN - Sistema de Informação de Agravos de Notificação) and the Brazilian Institute of Geography and Statistics (IBGE - Instituto Brasileiro de Geografia e Estatística), respectively, and are publicly available at the Rio de Janeiro Secretariat

of Health website (http://www.rio.rj.gov.br/dlstatic/10112/7079759/4197436/ZIKASE2015.pdf and http://www.rio.rj.gov.br/dlstatic/10112/10617973/4260330/ZIKASE2016.pdf, for 2015 and 2016, respectively). Note that SINAN reflects data from the public health system (SUS - Sistema Único de Saúde) only, which does not include data from private hospitals and health plans. The sociodevelopment index data come from the Instituto Pereira Passos and can be found at www.data.rio.

Appendices

Supplementary material is available in Appendix A:

- A.1: Stan code for the proposed model
- A.2: Convergence diagnostics for the proposed model
- A.3: Simulation study: generating data from the proposed BYM2-Gamma model
- A.4: Simulation study: generating data from the proposed BYM2-logCAR model
- A.5: Simulation study: no outlying areas
- A.6: Simulation study: distant outliers in France
- A.7: Simulation studies on the map of Rio de Janeiro
- A.8: Comparison with the model proposed by Corpas-Burgos and Martinez-Beneito (2020)

Chapter 4

A spatio-temporal model to detect potential outliers in disease mapping

Preamble to Manuscript 2. In disease mapping, in a purely spatial setting, Congdon (2017) proposed a modification of the Leroux prior (Leroux et al., 1999) to include scaling parameters that identify potentially outlying areas, after accounting for fixed effects. Further, Rushworth et al. (2014) proposed an extension of the purely spatial Leroux prior to the temporal framework. That is, at each time point, the vector of spatially structured latent effects are assumed to be centred around the random effects at the previous time point. The Rushworth model, however, assumes the conditional variance of the latent effects to be constant across space and over time, which does not allow the capture of observations that might fall on the tail of the distribution of the latent effects.

This manuscript proposes a spatio-temporal disease mapping model that allows for spatial heteroscedasticity. The proposed model extends that of Congdon (2017) to the spatio-temporal setting, similarly to how Rushworth et al. (2014) extended the Leroux prior. That is, areal scaling mixture parameters are included in the Rushworth model to allow and identify potential outliers. Finally, two prior distributions are investigated for the scaling

parameters, namely, independent and spatially structured components.

The contributions of this manuscript include (i) a novel spatio-temporal disease mapping model that aims to identify areas whose risks are potentially outlying at some time points, after accounting for fixed effects, (ii) extensive simulation studies to investigate the ability of the proposed model to identify outliers, (iii) two data applications to showcase how the two prior specifications of the proposed model may help in the analysis of weekly COVID-19 cases and hospitalisations across Montreal and France, during the second wave.

This manuscript is under revision for the journal *Spatial Statistics*.

A spatio-temporal model to detect potential outliers in disease mapping

Victoire Michal, Alexandra M. Schmidt.

Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada

Abstract

Spatio-temporal disease mapping models are commonly used to estimate the relative risk of a disease over time and across areas. For each area and time point, the disease count is modelled with a Poisson distribution whose mean is the product of an offset and the disease relative risk. This relative risk is commonly decomposed in the log scale as the sum of fixed and latent effects. The Rushworth model allows for spatio-temporal autocorrelation of the random effects. We build on the Rushworth model to accommodate and identify potentially outlying areas with respect to their disease relative risk evolution, after taking into account the fixed effects. An area may display outlying behaviour at some points in time but not all. At each time point, we assume the latent effects to be spatially structured and include scaling parameters in the precision matrix, to allow for heavy tails. Two prior specifications are considered for the scaling parameters: one where they are independent across space and one with spatial autocorrelation. We investigate the performance of the different prior specifications of the proposed model through simulation studies and analyse the weekly evolution of the number of COVID-19 cases across the 33 boroughs of Montreal and the 96 French departments during the second wave. In Montreal, 6 boroughs are found to be potentially outlying. In France, the model with spatially structured scaling parameters identified 21 departments as potential outliers. We find that these departments tend to be close to each other and within common French regions.

4.1 Introduction

Since 1995, disease mapping models have been proposed to estimate the spatio-temporal trend of relative risks (Bernardinelli et al., 1995; Lawson, 2018). In the literature on spatio-temporal disease mapping, models commonly assume that the disease cases follow a Poisson distribution whose mean is the product between an offset and the relative risk, which varies through time and across space. The relative risk is usually written as the sum of fixed and latent effects. Commonly, the random effects are spatially structured and evolve through time (see, e.g., Lee et al. (2018) for an overview). Further, in their discussion, Rushworth et al. (2014) note that two neighbouring areas may behave differently over time, which is usually not accounted for in spatio-temporal disease mapping models. We propose a spatio-temporal model that identifies potential outliers with respect to the disease risk. In particular, the proposed model aims to identify any area whose behaviour over time differs from their neighbours or the rest of the region of interest. This provides decision makers with tools to help prioritise interventions and implement localised policies.

Knorr-Held (2000) proposed a spatio-temporal disease mapping model wherein the log relative risks are decomposed as the sum of fixed, temporal, spatial and space-time interaction effects. Four different interaction patterns are considered, the more complex case assuming interaction terms that are both spatially and temporally structured. The covariance matrix for this interaction term is constructed following Clayton (1996). It is given by the Kronecker product between a temporal random walk structure and a spatial intrinsic conditional autoregressive (ICAR) structure (Besag, 1974). Ugarte et al. (2012) built on Knorr-Held (2000) to model the spatial effects following a Leroux prior (Leroux et al., 1999), while keeping the covariance matrix for the space-time interaction term as the combination of a random walk and ICAR interaction structure. Rushworth et al. (2014) proposed to reduce the parameter space by including only the space-time interaction effects and further introduced the Leroux structure in the spatio-temporal structure. More specifically, let $\boldsymbol{b}_{\cdot t} = [b_{1t}, \ldots, b_{nt}]^\top$ be the

vector of areal latent effects at time t, they assume

$$\boldsymbol{b}_{\cdot 1} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{Q}_L^{-1}\right), \quad \text{and} \quad \boldsymbol{b}_{\cdot t} \mid \boldsymbol{b}_{\cdot t-1} \sim \mathcal{N}\left(\alpha \boldsymbol{b}_{\cdot t-1}, \sigma^2 \boldsymbol{Q}_L^{-1}\right), \; t = 2, \ldots, T, \tag{4.1}$$

with conditional variance parameter σ^2, temporal smoothing parameter $\alpha \in [0, 1]$ and spatial smoothing parameter $\lambda \in [0, 1)$. The precision matrix $\boldsymbol{Q}_L = (1 - \lambda)\boldsymbol{I} + \lambda(\boldsymbol{D} - \boldsymbol{W})$ is the one proposed by Leroux et al. (1999), with $\boldsymbol{W} = [w_{ij}]$ a $n \times n$ matrix of spatial weights and $\boldsymbol{D} = diag(d_i)$, for $d_i = \sum_{j \neq i} w_{ij}$. The spatial weights are commonly defined as $w_{ij} = 1$ if areas i and j share a border, and $w_{ij} = 0$ otherwise. Hence, Rushworth et al. (2014) assume a non-separable spatio-temporal structure, where the temporal trend appears in the conditional means and the spatial structure, in the precision.

On the other hand, the models mentioned above assume spatial homogeneity through time. In the purely spatial setting, this issue of spatial heterogeneity may be addressed by allowing the spatial structure $\boldsymbol{W}$ to be estimated from the data (see, e.g., Lee and Mitchell (2013); Dean et al. (2019); Corpas-Burgos and Martinez-Beneito (2020)). Another approach is to allow the spatial dependence parameter λ to vary across space, which accommodates local differences in the spatial structure (MacNab, 2023). Other authors proposed two-step procedures to elicit clusters of areas that behave similarly and include that information in the spatial model (Anderson et al., 2014; Santafé et al., 2021). Congdon (2017) allowed for disparities by modifying the Leroux prior to include scaling mixture components $\kappa_i > 0, i = 1, \ldots, n$, such that $\boldsymbol{b} = [b_1, \ldots, b_n]^\top \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{Q}_C^-)$, where the precision matrix has diagonal elements $\boldsymbol{Q}_{C,ii} = \kappa_i(1 - \lambda + \lambda d_i)$ and off-diagonal elements $\boldsymbol{Q}_{C,ij} = -\lambda w_{ij} \kappa_i \kappa_j$. This joint distribution proposed by Congdon (2017) corresponds to the following conditional distributions:

$$b_i \mid \boldsymbol{b}_{(-i)} \sim \mathcal{N}\left(\frac{\lambda}{1 - \lambda + \lambda d_i} \sum_{j=1}^{n} w_{ij} \kappa_j b_j, \frac{\sigma^2}{\kappa_i \left(1 - \lambda + \lambda d_i\right)}\right), \; i = 1, \ldots, n. \tag{4.2}$$

In this proposal, the scaling mixture parameters help identify potential outliers, wherein $\kappa_i < 1$ implies that the ith area is a potential outlier. Let area i be an outlier and a neighbour of area j. Then $\kappa_i < 1$ inflates the conditional variance for b_i while allowing b_j to allocate less weight to the outlying b_i in its conditional mean, and borrow more strength from its non-outlying neighbours.

In the spatio-temporal setting, proposals have been made to extend some methods discussed in the previous paragraph. For instance, Lee and Lawson (2016) proposed to include a piecewise constant intercept term to identify clusters of areas that behave similarly over space and time. Rushworth et al. (2017) proposed a spatio-temporal model where the spatial structure is estimated based on the data. Different from these methods, the main aim of this paper is to propose a spatio-temporal model that accommodates and specifically identifies potential outlying areas, after accounting for fixed effects. Specifically, we propose to extend Congdon's prior in equation (4.2) to the spatio-temporal setting, similarly to how Rushworth et al. (2014) (4.1) extended the Leroux prior (Leroux et al., 1999). Throughout this paper, the term outlier designates both areas that may behave differently from their neighbours (spatial outliers), and areas that present extreme risks.

4.1.1 Illustration

To investigate the benefits of the proposed model, we consider two examples related to the coronavirus disease 2019 (COVID-19) pandemic. Due to the spatial dimension of the disease, disease mapping and spatio-temporal methods have been widely used to analyse COVID-19 counts (see, e.g., Franch-Pardo et al. (2020) for a review) in order to help decision makers understand the disease and implement policies. Further, COVID-19 counts tend to show different behaviours over time and across areas (see, e.g., Figures 4.1 and 4.2 for the behaviour of COVID-19 standardised morbidity ratios (SMRs) during the second wave in Montreal and in France).

First, we have data available across the 33 boroughs of Montreal, where the number of COVID-19 cases have been recorded weekly during the second wave, between August 23rd 2020 and March 20th 2021 (see, e.g., Institut national de santé publique du Québec (2024) for a COVID-19 timeline in the province of Quebec). The data come from the *Institut national de la santé publique du Québec* (INSPQ). Figure 4.1 showcases the SMR distribution for the COVID-19 cases across space at three different time points. Even with limiting policies in place (Institut national de santé publique du Québec, 2024), some boroughs appear to have elevated SMRs at some time points but not all, while other boroughs seem to never show extreme values. Similar to Michal et al. (2022), three auxiliary variables are considered to analyse these weekly COVID-19 cases, namely the number of beds in long-term care centres (*Centres d'hébergement et de soins de longue durée*, CHSLDs), the median age by borough and the population aged 25-64 with a university degree (Ville de Montréal, 2016), as a proxy for the socio-economic status.

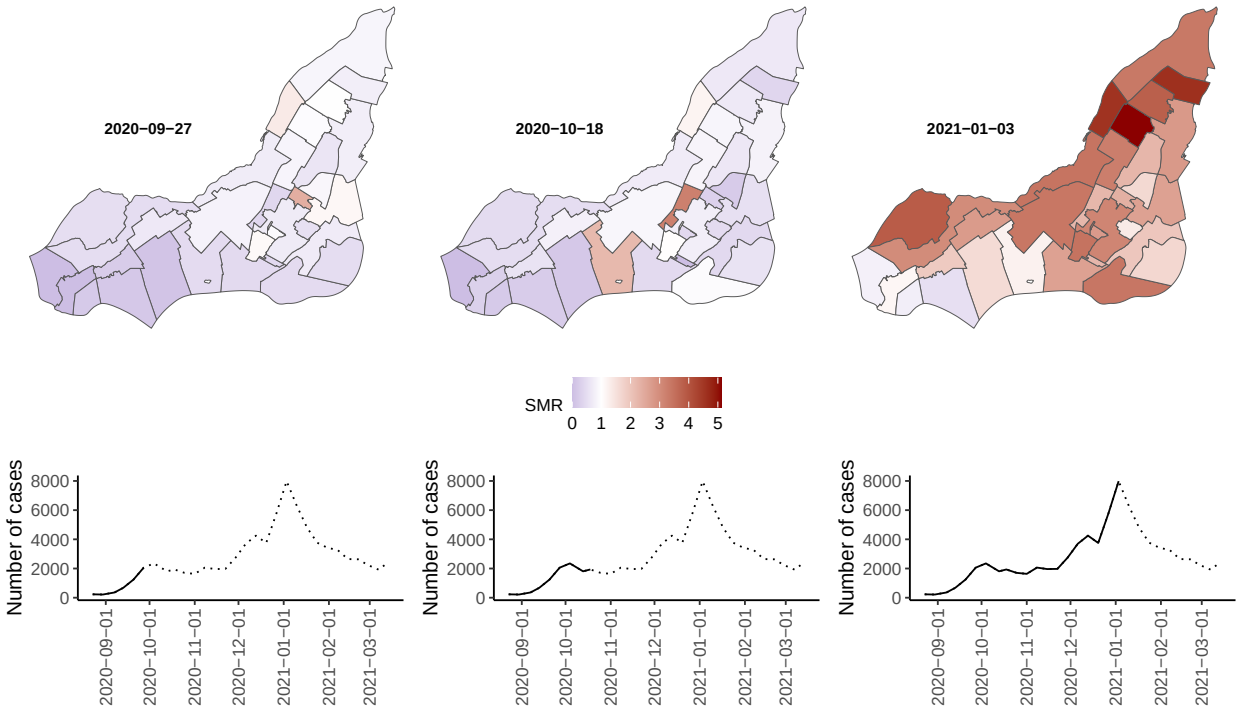


Figure 4.1: Maps of the SMR distribution across the boroughs of Montreal at three different time points (top) and distribution of the total number of COVID-19 cases over time (bottom).

As another example of the need to investigate outlying observations in spatio-temporal disease counts, we study the COVID-19 second wave in France. We have available the weekly counts of hospitalisation due to COVID-19 during the second wave, across the 96 French departments. In France, the second wave lasted 26 weeks between early July 2020, and the end of the year 2020 (Costemalle et al., 2021). The data are publicly available from the French national health agency (Santé publique France, 2023). Figure 4.2 shows the evolution of the COVID-19 SMRs across the French departments. It appears that some departments might behave differently than the others, in particular at the beginning of the second wave.

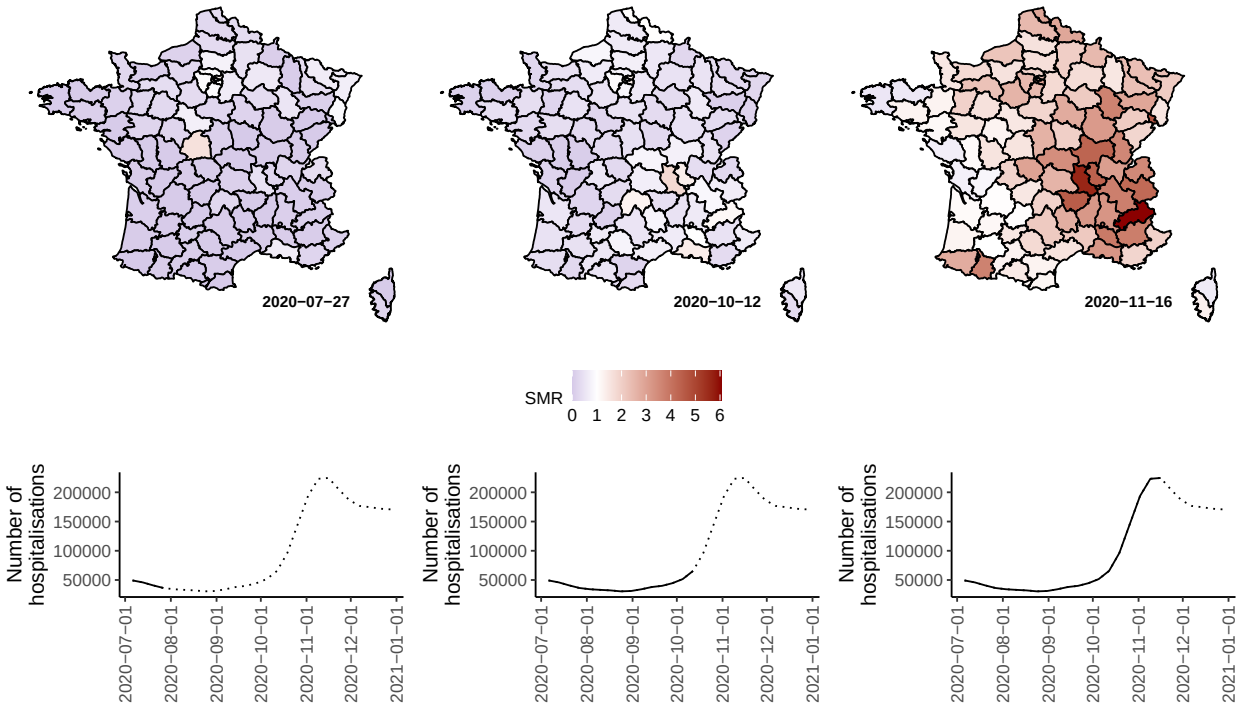


Figure 4.2: Maps of the SMR distribution across the French departments at three different time points (top) and distribution of the total number of COVID-19 hospitalisations over time (bottom).

This paper is organised as follows. Section 4.2 proposes the spatio-temporal model that accounts for and identifies outlying areas over time. Therein, the inference procedure, which is performed following the Bayesian paradigm, is also discussed. Section 4.3 shows the performance of the proposed model under different simulation scenarios (Section 4.3.1). Then, the proposed model is fitted to the COVID-19 data in Montreal (Section 4.3.2) and in France

(Section 4.3.3). Section 4.4 provides concluding remarks and points to potential future avenues of research.

4.2 Proposed model

Consider a region divided into n non-overlapping areas studied over T time points. Let Y_{it} be the number of cases of a disease recorded in area i at time t. Let E_i be the expected number at risk in area i, which we assume to be constant over time. The number of cases is modelled as follows:

$$Y_{it} \mid E_i, \mu_{it} \sim \text{Pois}\left(E_i \mu_{it}\right),$$

where μ_{it} is the relative disease risk for the ith area at time t, which is decomposed as

$$\log\left(\mu_{it}\right) = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} + b_{it},$$

where the intercept β_0 corresponds to the overall log risk across time and space, the p regression coefficients $\boldsymbol{\beta}$ multiply the vector of areal-level covariates $\boldsymbol{x}_i$, and b_{it} is a latent effect for the ith area at time t, which captures whatever is left after accounting for the covariates. A square $n \times n$ matrix of weights $\boldsymbol{W} = [w_{ij}]$ is defined to account for a spatial structure in the latent effects. Two areas i and j are said to be neighbours with a weight $w_{ij} = 1$ if they share a border and $w_{ij} = 0$, otherwise. From this 0-1 neighbourhood structure, the diagonal matrix $\boldsymbol{D} = diag(d_i)$, where $d_i = \sum_j w_{ij}$, corresponds to the matrix whose diagonal elements are the areal numbers of neighbours. We propose a modification of the Rushworth prior (4.1) to model the vector of latent effects $\boldsymbol{b} = [b_{11}, \ldots, b_{n1}, \ldots, b_{1T}, \ldots, b_{nT}]^\top$. Let $\boldsymbol{b}_{\cdot t} = [b_{1t}, \ldots, b_{nt}]^\top$, we assume

$$\boldsymbol{b}_{\cdot 1} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{Q}_C^-\right), \quad \text{and} \quad \boldsymbol{b}_{\cdot t} \mid \boldsymbol{b}_{\cdot t-1} \sim \mathcal{N}\left(\alpha \boldsymbol{b}_{\cdot t-1}, \sigma^2 \boldsymbol{Q}_C^-\right), \; t = 2, \ldots, T, \tag{4.3}$$

with temporal dependence parameter $|\alpha| < 1$, variance parameter $\sigma > 0$, and $\boldsymbol{Q}_C$ as proposed by Congdon (2017). The matrix $\boldsymbol{Q}_C$ has diagonal elements $\boldsymbol{Q}_{C_{ii}} = \kappa_i(1 - \lambda + \lambda d_i)$ and off-diagonal elements $\boldsymbol{Q}_{C_{ij}} = -w_{ij}\lambda\kappa_i\kappa_j$, where $\lambda \in [0, 1]$ and $\kappa_i > 0,\ i = 1,\ \ldots n$. The matrix $\boldsymbol{Q}_C$ may be written as

$$\boldsymbol{Q}_C = diag_1(\boldsymbol{\kappa}) \odot \left[(1 - \lambda)\boldsymbol{I} + \lambda\left(\boldsymbol{D} - \boldsymbol{W} \odot \boldsymbol{\kappa}\boldsymbol{\kappa}^\top\right)\right], \tag{4.4}$$

where $diag_1(\boldsymbol{\kappa})$ denotes the square matrix with diagonal elements $\boldsymbol{\kappa} = [\kappa_1, \ldots, \kappa_n]^\top$ and off-diagonal elements equal to 1. From expression (4.4), it is clear that the mixing parameter λ is a spatial dependence parameter and $\lambda = 0$ yields temporal latent effects that are independent across space, while $\lambda = 1$ implies fully structured spatio-temporal effects. Similarly, α appears as a temporal dependence parameter in the prior distribution (4.3), where $\alpha = 0$ leads to vectors of latent effects $\boldsymbol{b}_{\cdot t}$ that are spatially structured and independent over time, and $\alpha = 1$ implies fully structured spatio-temporal effects.

Our main contribution lies in the inclusion of the scaling parameters $\boldsymbol{\kappa}$. In expression (4.4), they appear in the diagonal elements of the precision matrix, and $\kappa_i < 1$ inflates the conditional variances of the latent effects for the ith area. Additionally, the κ's impact the spatial weights as follows: $\boldsymbol{W} \odot \boldsymbol{\kappa}\boldsymbol{\kappa}^\top = [w_{ij}\kappa_i\kappa_j]$. Hence, at any time t, $\kappa_i < 1$ implies an inflated conditional variance for the ith latent effect, and a decreased correlation between areas j and i when they are neighbours. Following Congdon (2017), these parameters act as outlier indicators and an area i is defined as a potential outlier if $\kappa_i < 1$. An area may be outlying at all time points, or at some points in time.

Further, the role of the scaling parameters can be studied from the conditional distributions of the latent effects that result from (4.3). For $t \geq 2$, the joint distribution of the latent effects, (4.3), corresponds to the set of n Gaussian conditional distributions with expectation $\mathrm{E}\left(b_{it} \mid \boldsymbol{b}_{\cdot t-1}, \boldsymbol{b}_{(-i)t}\right) = \alpha b_{it-1} + \lambda/(1 - \lambda + \lambda d_i)\sum_{j\sim i}\kappa_j(b_{jt} - \alpha b_{jt-1})$ and variance $\mathrm{V}\left(b_{it} \mid \boldsymbol{b}_{\cdot t-1}, \boldsymbol{b}_{(-i)t}\right) = \sigma^2/\left(\kappa_i(1 - \lambda + \lambda d_i)\right)$, where $\boldsymbol{b}_{(-i)t} = [b_{1t}, \ldots, b_{i-1t}, b_{i+1t}, \ldots, b_{nt}]^\top$ and

$j \sim i$ means that areas i and j are neighbours. Hence, for outlying area j a neighbour of area i, κ_j is smaller than 1 and the difference $b_{jt} - \alpha b_{jt-1}$ contributes less to the conditional mean of b_{it} than another neighbour ℓ whose κ_ℓ is greater or equal to 1.

Two priors are proposed for the scaling mixture components. First, following Congdon (2017), independent gamma priors are assigned to the scaling mixture components, $\kappa_i \overset{i.i.d.}{\sim} \text{Gamma}(\nu/2, \nu/2)$. This implies that $\mathbb{E}(\kappa_i \mid \nu) = 1$ and $\mathbb{V}(\kappa_i \mid \nu) = 2/\nu$, *a priori*; that is, the prior assumption is that area i is not an outlier, with small variance for large hyperparameter ν. Following Gelman et al. (2004) and Michal et al. (2024), we assume $\nu \sim \text{Exp}(1/4)$. Second, we investigate a discretisation of the continuous spatially structured scaling process proposed by Palacios and Steel (2006). For this discrete parametrisation of Palacios and Steel (2006), we follow Michal et al. (2024), who assign a proper conditional autoregressive (PCAR) prior to the scaling components. They assume $\ln(\kappa_i) \equiv -\nu/2 + z_i$, with scaled spatially structured $\boldsymbol{z} \equiv [z_1, \ldots, z_n]^\top \sim \mathcal{N}\left(\boldsymbol{0}, \nu \boldsymbol{Q}_{\rho,\star}^{-1}\right)$ and $\nu \sim \text{Exp}(1/0.3)$. The precision matrix $\boldsymbol{Q}_\rho = \boldsymbol{D} - \rho \boldsymbol{W}$, which is positive definite for $\rho \in [0, 1)$ (Banerjee et al., 2014), is scaled by $h_\rho = \exp\left[(1/n) \sum_{i=1}^n \ln\left(\boldsymbol{Q}_{\rho,ii}^{-1}\right)\right]$ such that $\boldsymbol{Q}_{\rho,\star} = h_\rho \boldsymbol{Q}_\rho$. This scaling of the spatially structured precision matrix yields approximate marginal variances $\mathbb{V}(\ln(\kappa_i) \mid \nu) \simeq \nu$, for any spatial structure under study (Riebler et al., 2016). Similar to the case of independent gamma priors on the scaling components, this spatially structured prior implies that $\mathbb{E}(\kappa_i \mid \nu) = 1$, which means that the ith area is not an outlier *a priori*. For further discussion regarding these two priors for the κ parameters, see Michal et al. (2024).

4.2.1 Inference procedure

The proposal (4.3) and the Rushworth model (4.1) do not yield posterior distributions with a closed form. Therefore, to approximate the posterior distribution of the resultant parameter vector we resort to Markov Chain Monte Carlo (MCMC) methods. Specifically, we use the `R` package `rstan` (Stan Development Team, 2020), which efficiently estimates posterior distributions from complex hierarchical models where spatial and temporal structures are

studied, using a sampler based on a Hamiltonian Monte Carlo algorithm (Morris et al., 2019).

To avoid a potential identifiability issue between the intercept and the latent effects in the proposed model (Rue and Held, 2005), we impose a sum-to-zero constraint on $\boldsymbol{b}_{\cdot 1}$, the latent effects for the first time point. In the MCMC procedure, a soft sum-to-zero constraint corresponds to assuming $\sum_{i=1}^{n} b_{i1} \sim \mathcal{N}(0, 0.001n)$ (Morris et al., 2019). The `rstan` code implemented to fit the proposed model in the simulation studies and data applications summarised in Section 4.3 is displayed in Appendix B.1. For more details on the data and code, see https://github.com/vicmic13/SpatioTemporal_DiseaseMapping_OutlyingAreas.

4.3 Data analyses

In Section 4.3.1, we present the results from a simulation study where the goal is to assess the performance of the proposed model (4.3) compared to the Rushworth model (4.1). Data are first generated from the Rushworth model and some areas are contaminated into outliers that we aim to identify. Note that in Appendix B.2, we present results from a simulation study where data are generated from the proposed model to check whether we can recover the parameters used to generate the data. The results suggest that we can estimate the parameters of the model and there does not seem to be any identifiability issue in the model. Finally, Sections 4.3.2 and 4.3.3 provide the analyses of the COVID-19 data across the 33 boroughs of Montreal and the 96 French departments.

4.3.1 Simulation study

The simulation study presented in this section is inspired by the analyses shown in Appendix C of Fonseca et al. (2023) and by the simulation studies summarised in Michal et al. (2024). We aim to investigate the performance of the proposed model when compared to the

Rushworth model, and assess its ability to identify outliers when the truth is known. The region of interest is Montreal, which is divided into $n = 33$ boroughs, and the time period is arbitrarily set to $T = 52$ time points, in order to mimic weekly data recorded over a year. A known set of boroughs is contaminated at some of the time points, but not all, into outlying areas. The goal is to identify these areas that sometimes present outlying risks.

Two simulation scenarios are considered to experiment with the prior specification of the scaling mixture components. In both scenarios, the overall log risk is $\beta_0 = -1$ and the latent effects $b_{it},\ i = 1, \ldots, n,\ t = 1, \ldots, T$ are generated according to the Rushworth model (4.1) with $\lambda = 0.7$, $\alpha = 0.85$, and $\sigma = 0.3$. Both simulation examples use offsets $E_1, \ldots, E_n$ taken from the analysis of COVID-19 cases shown in Section 4.3.2. The offsets are sorted into five categories based on their magnitude. The levels are termed "Small" ($E < 26$), "Medium low" ($E \in [26, 45)$), "Medium" ($E \in [45, 108)$), "Medium high" ($E \in [108, 147)$), and "High" ($E \geq 147$). In each simulation scenario, five boroughs (one per offset category) are then selected to be contaminated into outlying areas at given times. In one case, the selected boroughs are distant from each other and do not share a border, and in the second scenario, the five boroughs are neighbours. The maps on the left-hand side of Figure 4.3 highlight the selected outliers based on their offset size, for each simulation scenario. For j denoting one of the five selected areas in each scenario, we contaminate its latent effect as follows: $b_{jt}^{\text{contaminated}} = b_{jt} + r_{jt} \times c_{jt}$, with $c_{jt} \sim \mathcal{U}\left(\max(|b_{(1)t}|, |b_{(n)t}|), 1.5 \max(|b_{(1)t}|, |b_{(n)t}|)\right)$, where $b_{(1)t}$ and $b_{(n)t}$ denote the minimum and maximum generated latent effects at time t, respectively. The quantity $r_{jt} \in \{0, 1\}$ determines whether the jth area is outlying at time t as follows: for $t = 1$, $r_{jt} \sim \text{Ber}(0.4)$, and for $t \geq 2$, $r_{jt} = r_{jt-1}$ with probability 0.8, or $r_{jt} \sim \text{Ber}(0.4)$ otherwise. Figure B.4 in Appendix B.3 shows the latent effects generated from the Rushworth model before and after contamination. Finally, for each simulation scenario, $R = 100$ datasets of $n = 33$ boroughs and $T = 52$ time points are created according to the hierarchical Poisson model $Y_{it} \sim \text{Pois}(E_i \exp(\beta_0 + b_{it}))$.

Six models are fitted to the data generated for each simulation scenario. First, because Rushworth et al. (2014) discuss the necessity to estimate the parameter α, the Rushworth model (4.1) is fitted with $\alpha = 1$ and with a prior $\alpha \sim \mathcal{U}(-1, 1)$, denoted R(1) and R($\alpha$), respectively. Then, four versions of the proposed Heavy Rushworth model (4.3) are considered: two impose $\alpha = 1$ and the two others assume a uniform prior, $\alpha \sim \mathcal{U}(-1, 1)$. For each pair, one version, denoted HR($\cdot$), imposes independent gamma priors to the scaling parameters, $\kappa_i \sim$ Gamma$(\nu/2, \nu/2)$, with hyperparameter as discussed in Section 4.2, $\nu \sim$ Exp$(1/4)$, while the other, denoted HR-LPC($\cdot$), imposes the spatially structured prior for $\boldsymbol{\kappa}$ that is defined in Section 4.2 with hyperparameter $\nu \sim$ Exp$(1/0.3)$. The prior assignment and notation of the six models considered are summarised in Table 4.1.

Model	α	$\boldsymbol{\kappa}$	ν
R(1)	1	–	–
R(α)	$\mathcal{U}(-1, 1)$	–	–
HR(1)	1	Gamma$(\nu/2, \nu/2)$	Exp$(1/4)$
HR(α)	$\mathcal{U}(-1, 1)$	Gamma$(\nu/2, \nu/2)$	Exp$(1/4)$
HR-LPC(1)	1	log-PCAR	Exp$(1/0.3)$
HR-LPC(α)	$\mathcal{U}(-1, 1)$	log-PCAR	Exp$(1/0.3)$

Table 4.1: Notation and description of the six models fitted to the simulated data.

For each model, the MCMC procedure with two chains converged after 5,000 iterations with a burn-in period of 2,500 iterations and a thinning factor of 5, as assessed by the traceplots, effective sample sizes and $\widehat{R}$ statistic (Gelman and Rubin, 1992; Vehtari et al., 2021).

Figure 4.3 and Table 4.2 show how often the four versions of the proposed model identify the correct set of contaminated boroughs, depending on the simulation scenario. Area i is identified as an outlier when $\kappa_{u,i} < 1$, where $\kappa_{u,i}$ denotes the upper limit of the posterior 95% credible interval for κ_i. In Table 4.2, the sensitivity measures the frequency of correct outlier identification (%) and the specificity quantifies how often the models do not point out areas that are not contaminated (%). For both measures, higher values are preferred. When the contaminated boroughs are not neighbours, the proposals HR(α) and HR-LPC(α), which estimate α, perform better than the ones with fixed $\alpha = 1$. In particular, HR(1) and

HR-LPC(1) correctly identify Sainte-Anne-de-Bellevue (purple borough in Figure 4.3) only 26% and 5% of the time, respectively, while HR-LPC(α) and HR(α) reach 69% and 88% sensitivity values for this contaminated borough with a small offset. A similar result is obtained in the second scenario, where both HR(α) and HR-LPC(α) identify Montréal-Ouest (purple borough) at least 90% of the time, whereas the versions with fixed $\alpha = 1$ do not find this borough in more than 50% of the replicates. When the offsets are larger, in both simulation scenarios, the four versions of the proposed model accurately point out the correct outliers 100% of the time. In terms of smoothing, in both simulation scenarios, all models equally succeed in not identifying irrelevant areas (e.g., overall specificities above 99.6).

	Offset category	HR(1)	HR(α)	HR-LPC(1)	HR-LPC(α)
Distant outliers					
Sensitiviy	Small	26.0	88.0	5.0	69.0
	Medium low	100.0	100.0	100.0	100.0
	Medium	100.0	100.0	100.0	100.0
	Medium high	100.0	100.0	100.0	100.0
	High	100.0	100.0	100.0	100.0
	Overall	85.2	97.6	81.0	93.8
Specificity	Small	99.8	99.8	99.8	99.8
	Medium low	100.0	100.0	100.0	100.0
	Medium	100.0	100.0	99.8	100.0
	Medium high	99.2	100.0	100.0	100.0
	High	99.8	100.0	100.0	100.0
	Overall	99.8	99.9	99.9	99.9
Neighbouring outliers					
Sensitiviy	Small	50.0	90.0	40.0	91.0
	Medium low	100.0	100.0	99.0	100.0
	Medium	100.0	100.0	100.0	100.0
	Medium high	100.0	100.0	100.0	100.0
	High	100.0	100.0	100.0	100.0
	Overall	90.0	98.0	87.8	98.2
Specificity	Small	100.0	99.8	100.0	99.8
	Medium low	99.8	99.8	100.0	100.0
	Medium	99.7	99.8	99.7	99.8
	Medium high	99.0	99.6	100.0	100.0
	High	99.2	100.0	100.0	100.0
	Overall	99.6	99.8	99.9	99.9

Table 4.2: Sensitivity and specificity of the outlier detection for each version of the proposed model in both simulation scenarios, depending on the offset size and overall.

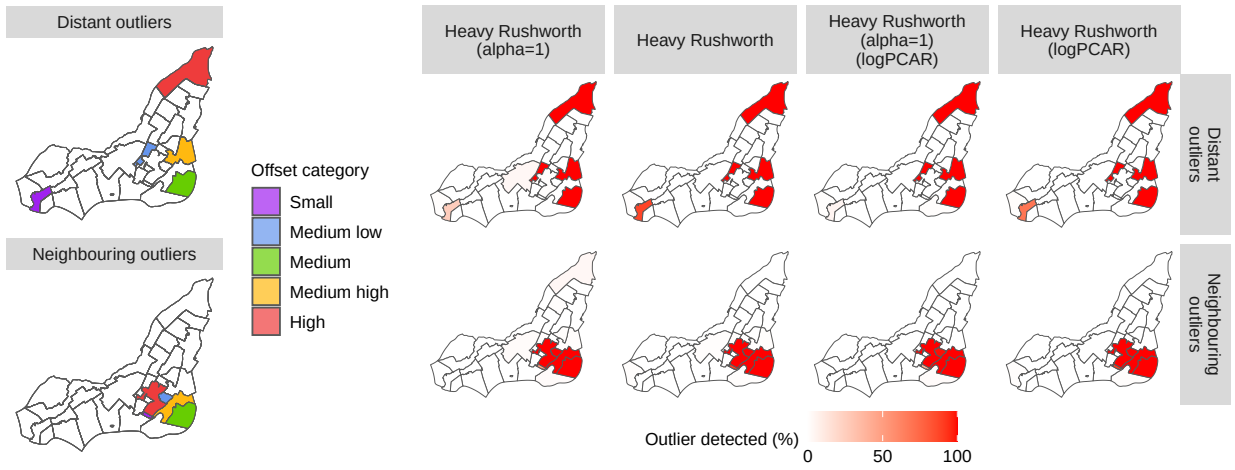


Figure 4.3: Left: maps of the selected boroughs of Montreal that were contaminated into outlying areas based on their offset sizes, for each simulation scenario. Right: Percentage of times each borough is detected as a potential outlier according to the four versions of the proposed model across the two simulation scenarios. A borough is defined as a potential outlier when $\kappa_u < 1$, where κ_u is the upper limit of the 95% posterior credible interval for κ.

Regarding the performances of the models, Table 4.3 summarises, for each simulation scenario, the average WAIC (Watanabe and Opper, 2010) computed across the 100 replicates for each model, as well as their average MSE within the contaminated and non-contaminated boroughs, and overall. For more details on these performance measures, Figure B.5 in Appendix B.3 shows the WAIC values across the 100 replicates for each model, as well as their average MSE across the different offset size categories. Smaller WAIC values are preferred and in both simulation scenarios, Table 4.3 shows that HR(1), HR(α), HR-LPC(1) and HR-LPC(α) perform better than R(1) and R(α) (e.g., when outliers are distant, R(α) yields 10,829 as the average WAIC, vs 10,757 for HR-LPC(α)). In terms of WAIC, HR(α) and HR-LPC(α) always perform better than the Rushworth models, or than the proposed models with fixed $\alpha = 1$. Further, in this case where the true temporal dependence parameter is $\alpha = 0.85$, when neighbouring boroughs are contaminated, the Rushworth model with unknown α is correctly pointed out by WAIC as the best model when compared to both versions of the proposed model where $\alpha = 1$ is fixed. It can also be noted that, as expected, when distant boroughs are contaminated, HR(α) performs slightly better than HR-LPC(α),

while the converse is observed with neighbouring outliers. Finally, in both scenarios, when α is estimated, both the Rushworth and Heavy Rushworth models (with independent and spatially structured $\boldsymbol{\kappa}$) perform better than their counterparts with fixed $\alpha = 1$. This result is sensible, as the data were generated with $\alpha = 0.85$.

With respect to the MSE based on fitted and observed values, regardless of the simulation scenario, the four versions of the proposed model tend to perform better than the Rushworth model among the contaminated boroughs (e.g., for neighbouring contaminated boroughs, the average MSEs are 8.0 and 40.0 for HR-LPC(α) and R(α), respectively). However, to a lesser extent, the converse is observed within the non-contaminated areas (e.g., overall MSEs of 14.4 and 7.7 for HR-LPC(α) and R(α), respectively, in the same scenario). Finally, in terms of MSE, regardless of the scenario, the proposed HR(α) and HR-LPC(α) tend to perform slightly better than HR(1) and HR-LPC(1) (e.g., for distant contaminated boroughs, overall average MSEs of 14.3 vs 17.6 for HR(α) and HR(1), respectively), which agrees with Rushworth et al. (2014).

		R(1)	R(α)	HR(1)	HR(α)	HR-LPC(1)	HR-LPC(α)
Distant outliers							
	WAIC	10,912.2	10,829.1	10,819.2	10,755.5	10,814.0	10,757.4
	p_W	647.5	631.3	568.7	569.4	573.0	572.2
MSE	Contaminated	44.4	33.9	7.4	6.6	8.9	7.1
	Not contaminated	10.6	8.4	19.5	15.7	18.2	15.3
	Overall	15.7	12.2	17.6	14.3	16.8	14.0
Neighbouring outliers							
	WAIC	10,861.0	10,797.2	10,818.8	10,758.2	10,811.7	10,756.9
	p_W	635.7	622.9	573.0	573.7	575.4	575.8
MSE	Contaminated	52.2	40.1	8.4	7.4	9.7	8.0
	Not contaminated	9.4	7.7	18.0	14.6	17.3	14.4
	Overall	15.9	12.6	16.5	13.5	16.2	13.4

Table 4.3: Average WAIC and MSE computed over the 100 replicates for each model and each simulation scenario under the different fitted models. The MSE results are distinguished between the contaminated boroughs, the non-contaminated ones, and overall.

4.3.2 Analysis of COVID-19 cases in Montreal during the second wave

The number of COVID-19 cases were recorded weekly across the $n = 33$ boroughs of Montreal during the second wave, which consisted of $T = 30$ weeks between August 23rd 2020 and March 20th 2021 (Institut national de santé publique du Québec, 2024). Let Y_{it} be the number of cases recorded in the ith borough during the tth week, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Following Section 4.2, the number of COVID-19 cases is modelled via a Poisson distribution whose offset E_i is computed using the total number of cases and the population size of each borough, denoted P. In particular, for each borough i, E_i is assumed constant over time and the offsets represent the overall expected number of cases were the disease spread uniformly across space, during the second wave. Hence, we compute $E_i = \left(\sum_{i,t} Y_{it} / \sum_i P_i\right) \times P_i/T$ (Freitas et al., 2021) and yield expected cases that range from 1.3 to 233.9. Further, three auxiliary variables that are measured at the borough level are standardised and included to model the COVID-19 cases: the median age, the percentage of the population aged 25-64 with a university diploma, and the number of beds in CHSLDs. Figure 4.1 in Section 4.1.1 shows the SMR distribution across the boroughs in Montreal at three different points in time, alongside the evolution of the total number of cases. It can be seen that some boroughs have elevated SMRs across some weeks, but not all. For example, during the week of October 18th 2020, Mont-Royal and Dorval have higher SMRs than the rest of Montreal, as well as higher SMRs than those observed in those two boroughs during previous weeks or following ones. The aim of this analysis is to identify potential outlying boroughs, after accounting for the covariates' effect. Similar to Section 4.3.1, four versions of the proposed model are fitted to these weekly counts. The models are again denoted HR(1), HR(α), HR-LPC(1) and HR-LPC(α), as summarised in Table 4.1. The performance of the proposed model is compared to that of the Rushworth model, both by fixing $\alpha = 1$ and imposing a prior $\alpha \sim \mathcal{U}(-1, 1)$. All prior specifications follow the same ones used in the simulation study (Section 4.3.1).

The six models are fitted in `R` using the `rstan` package (Stan Development Team, 2020). Convergence of two MCMC chains is attained after 10,000 iterations with a burn-in period of 5,000 and a thinning factor of 5. The diagnostics used to assess convergence are the trace plots, the effective sample sizes, and the $\widehat{R}$ statistic (Gelman and Rubin, 1992; Vehtari et al., 2021).

Table 4.4 shows the performance measures for each model, and the estimated posterior summaries for the parameters. In terms of WAIC, smaller values are preferred, and the proposed model always performs better than the two versions of the Rushworth model (e.g., 6779 vs 6870, for HR-LPC(α) and R(α), respectively). All models that allowed the temporal dependence parameter α to be estimated yielded smaller WAICs than the ones that fixed $\alpha = 1$. In fact, R(α), HR(α), and HR-LPC(α) resulted in 95% posterior credible intervals for this parameter that were smaller than 1 (approximately $(0.8, 0.9)$). Finally, regarding the WAIC values, HR(α) performs the best among the four versions of the proposed model, which seems to indicate that there is no need for spatially structured scaling mixture components when analysing COVID-19 cases in Montreal during the second wave. The overall MSE results agree with the WAIC ones: the models that estimate α perform better than the ones with fixed $\alpha = 1$ (19, 21 and 21, for R(α), HR(α) and HR-LPC(α), respectively, vs 23, 26 and 24 for their respective counterparts). The MSE results are further distinguished between the boroughs identified as potential outliers by the proposed model, and the rest of Montreal. In the potentially outlying areas, the four versions of the proposed model yield MSEs that are about 3 times smaller than the Rushworth ones. However, in the boroughs that are not found to be potential outliers by the proposed model, the MSEs that result from fitting the Rushworth models are 1.7 times smaller than the ones obtained from the proposed Heavy Rushworth models.

The three covariate effects are found to be weak by all models. The posterior 95% credible intervals for the coefficient corresponding to the percent of the population aged 25-64 with

a university diploma all include 0 (e.g., $(-1, 0.3)$ for HR(α)). While the credible interval for the age regression parameter includes 0 in the R(1) model ($(-0.3, 0.0)$), they are negative and are on the cusp of including 0 in the HR(1), HR(α) and R(α) models (e.g., $(-0.22, -0.01)$ for R(α)). The same result is obtained for the number of CHSLD beds' parameter, where the posterior credible intervals almost include 0 (e.g., (-0.18, -0.00) for HR(1) and (-0.12, 0.01) for HR(α)).

		R(1)	R(α)	HR(1)	HR(α)	HR-LPC(1)	HR-LPC(α)
Model fit performance measures							
WAIC		6921.9	6870.3	6803.3	6764.6	6793.3	6778.6
p_W		468.8	458.8	437.6	425.6	435.7	430.1
MSE	Outliers	46.3	39.2	11.7	11.4	14.2	14.5
	Not outliers	17.9	15.1	29.4	23.0	25.8	22.5
	Overall	23.1	19.5	26.2	20.9	23.7	21.1
Posterior summaries for the parameters: Mean (95% CI)							
β_0		-2.56 (-3.13,-1.99)	-1.49 (-2.27,-0.78)	-2.25 (-2.81,-1.64)	-1.37 (-1.96,-0.86)	-2.55 (-2.96,-2.11)	-2.45 (-2.90,-2.04)
β_{diploma}		0.16 (-0.68,1.01)	-0.46 (-1.09,0.24)	0.04 (-0.86,0.90)	-0.40 (-1.05,0.30)	0.08 (-0.79,0.89)	-0.08 (-0.88,0.75)
β_{age}		-0.14 (-0.29,0.00)	-0.11 (-0.22,-0.01)	-0.14 (-0.27,-0.00)	-0.11 (-0.21,-0.02)	-0.12 (-0.26,0.02)	-0.12 (-0.25,0.00)
β_{beds}		-0.11 (-0.22,-0.00)	-0.05 (-0.13,0.02)	-0.09 (-0.18,-0.00)	-0.05 (-0.12,0.01)	-0.09 (-0.20,0.00)	-0.06 (-0.16,0.02)
α		-	0.87 (0.82,0.92)	–	0.87 (0.81,0.93)	–	0.92 (0.87,0.97)
λ		0.90 (0.85,0.94)	0.94 (0.91,0.97)	0.55 (0.32,0.88)	0.59 (0.37,0.89)	0.41 (0.18,0.76)	0.50 (0.25,0.83)
σ		0.43 (0.40,0.47)	0.45 (0.42,0.49)	0.31 (0.27,0.36)	0.33 (0.29,0.38)	0.32 (0.28,0.36)	0.34 (0.29,0.38)
ν		-	-	3.58 (2.03,5.93)	4.14 (2.22,6.91)	2.20 (1.28,3.43)	1.88 (1.02,3.06)

Table 4.4: Results from the analysis of COVID-19 reported cases in Montreal during the second wave (23/08/2020 – 20/03/2021). Model assessment (WAIC and MSE) and parameter posterior summaries: posterior mean and 95% credible interval (CI).

The spatial dependence parameter λ is estimated closer to 1 by the Rushworth models than by the four versions of the proposed model (e.g., posterior means of 0.94 and 0.59, for

R(α) and HR(α), respectively). This means that at each time point when modelling the latent effects, the Rushworth model gives more weight to the spatial structure compared to the proposed model. On the other hand, the four versions of the Heavy Rushworth model estimate smaller conditional standard deviations of the random effects than the Rushworth models (e.g., posterior means of 0.33 and 0.43, for HR(α) and R(α), respectively). This result is sensible as the inclusion of the scaling mixture components $\boldsymbol{\kappa}$ in the proposed model allows the variances to be raised in the boroughs that need it, without increasing the variability in the entire city of Montreal.

Figure 4.4 displays maps of the posterior means, at different time points, of the relative risks and the latent effects resulting from the proposed HR(α) model, which is the one that performed best in terms of WAIC, among all the models considered. The circles indicate which boroughs are identified as potential outliers. It is worth mentioning that the same 6 boroughs are indicated as potential outliers by the HR(1), HR-LPC(1), and HR-LPC(α) models. Similar to Section 4.3.1, borough i is found to be a potential outlier if $\kappa_{u,i} < 1$, where $\kappa_{u,i}$ is the upper limit of the 95% posterior credible for κ_i. It is interesting to see that boroughs may be found to be potential outliers without always presenting extreme risk or latent effect. For example, Dorval (red circle) and Mont-Royal (black circle) appear to have high estimated relative risks and latent effects during the week of October 18th, 2020, compared to the other areas, without it being the case 3 weeks earlier or during the week of January 3rd, 2021, which is the second wave peak as shown on the right-hand side of Figure 4.4. Similarly, Outremont (gray) shows higher values during the week of September 27th, 2020, but not during the following weeks. This is the aim of the proposed model, to identify areas that may have shown unexpected behaviours at some time points, to better understand the spread of the disease and prioritise future interventions.

It is also interesting to note that some areas might sometimes have high estimated risks without being identified as potential outliers, e.g. Saint-Leonard, the darkest borough during

the week of January 3rd, 2021. The map of the posterior means of the latent effects for that week shows that after accounting for the fixed effects, the remaining latent effect for Saint-Leonard behaves like the rest of Montreal. On the other hand, le Plateau Mont-Royal (green circle), which does not appear to have extreme estimated risks over time, has an estimated latent effect that differs from its neighbours during the October 18th week. This behaviour is found to be potentially outlying by the proposed model and further investigation might help understand it.

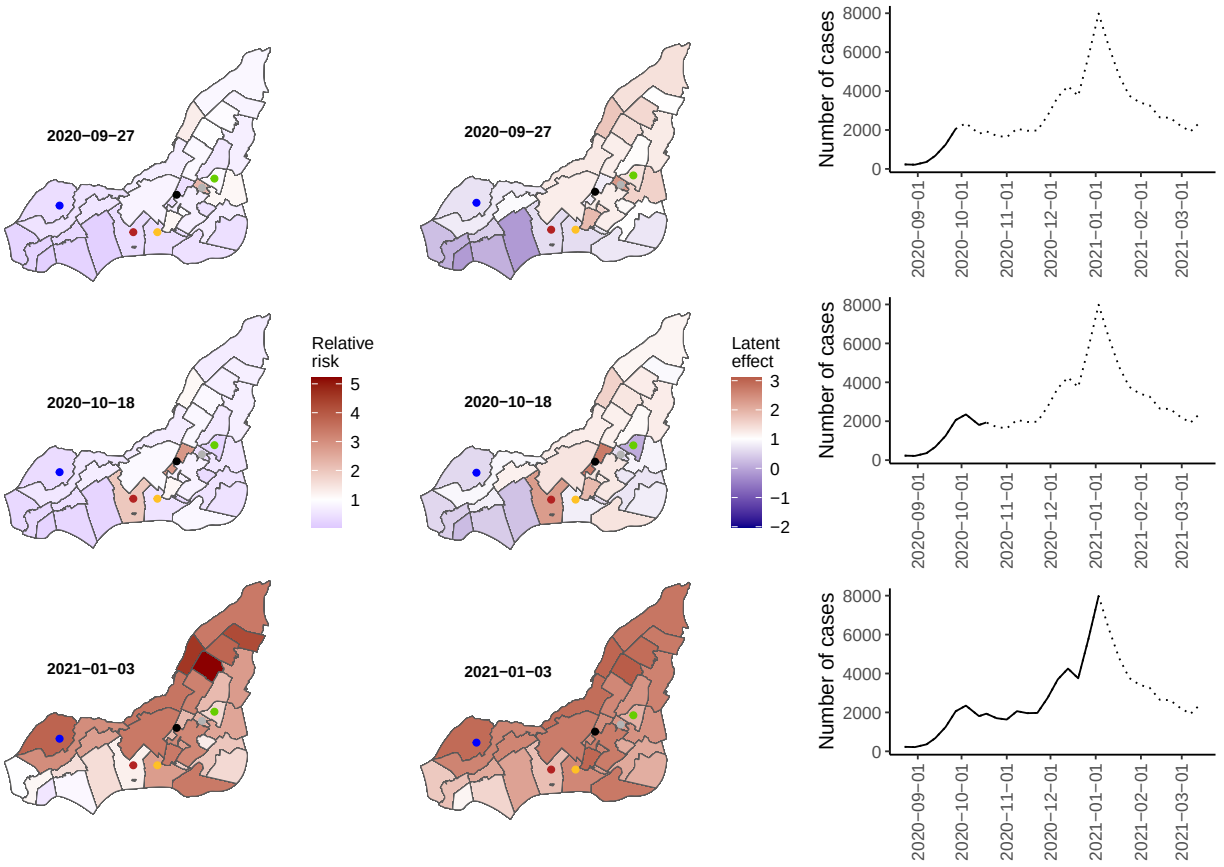


Figure 4.4: Maps of the COVID-19 relative risks (left) and latent effects (centre) estimated by the Heavy Rushworth model for three different time points across the boroughs of Montreal and total number of cases recorded over time in Montreal (right). Solid coloured circles: boroughs identified as potential outliers by the Heavy Rushworth model, the colours distinguish the boroughs to help discuss the results. The outliers are pointed out when $\kappa_u < 1$, where κ_u is the upper bound of the posterior 95% credible interval of κ.

4.3.3 Analysis of COVID-19 hospitalisations in France during the second wave

In this section, the weekly hospitalisation counts due to COVID-19 are studied across the $n = 96$ French departments, during the second wave. In France, the second wave consisted of $T = 26$ weeks, from July 2020 to December 2020, included (Costemalle et al., 2021). Let Y_{it} be the hospitalisation count recorded in department $i = 1, \ldots, 96$ at time $t = 1, \ldots, 26$. Similar to Section 4.3.2, the hospitalisation counts are modelled according to a Poisson distribution whose offsets E are computed from the weekly counts and the departments' population sizes. Figure 4.2 in Section 4.1.1 maps the evolution of the SMRs (top) and of the total number of hospitalisations (bottom). Some departments seem to have extreme values, compared to their neighbours or the rest of France, at some time points. For example, during the peak hospitalisation week (November 16th, 2020), Pyrénées-Atlantiques and Hautes-Pyrénées (two south-west departments) seem to have recorded higher SMRs than their neighbours, while Haute-Corse (northern Corsica) appears to be on the lower tail of the SMR distribution that week. To further study these potentially outlying behaviours, the proposed model is fitted to the data and compared to the Rushworth model. It is worth mentioning that since Corsica is an island, its two departments do not share a border with any other departments. There are however daily ferries travelling between these two departments and the three south-east ones on the Mediterranean Sea (Bouches-du-Rhône, Var, and Alpes-Maritimes), hence a spatial weight $w_{ij} = 1$ is assigned between all of them. Additionally, the proportion of population aged 75 or older is included in the models as a covariate that is fixed through time, x_i. The same priors as the ones described in the previous sections are considered for the model parameters and are again denoted HR(1), HR(α), HR-LPC(1), HR-LPC(α), R(1), and R(α).

The MCMC procedure that included two chains converged for all six models after 10,000 iterations, a burn-in period of 5,000, and a thinning factor of 5, as assessed through trace plots,

effective sample sizes and $\widehat{R}$ statistics (Gelman and Rubin, 1992; Vehtari et al., 2021).

Table 4.5 summarises the performances of the models and parameter posterior distributions. In terms of WAIC, the proposal yielded smaller values than the Rushworth model (e.g., 23,450 and 23,263 for R(α) and HR(α), respectively). The proposed HR-LPC(α) performs the best, among the six models considered, which may indicate a need for spatially structured scaling mixture components, κ's. Note that the difference is small between the proposed models that estimate α and the ones with fixed $\alpha = 1$ (e.g., 23,263 and 23,261 for HR(1) and HR(α), respectively). This result is sensible as both HR(α) and HR-LPC(α) estimate (0.96, 0.99) as the posterior 95% credible interval for α, which is very close to $\alpha = 1$. On the other hand, R(α) yields a smaller WAIC value than R(1) (23,450 vs 23,468), which corresponds to α estimated smaller than 1, with a posterior 95% credible interval of (0.93, 0.96).

In terms of MSE, similar to Sections 4.3.1 and 4.3.2, all versions of the proposed model result in smaller values among the departments identified as potential outliers, compared to R(1) and R(α) (e.g., 8.6 vs 27.3 for HR-LPC(α) and R(α), respectively). On the other hand, R(1) and R(α) perform better in terms of MSE among the departments that are not identified as potential outliers (e.g., 12.1 vs 31.3 for R(1) and HR(1), respectively).

Regarding the regression coefficient, all models agree on a significantly negative relationship between the proportion of the population aged 75+ and the hospitalisation counts (e.g., posterior mean of -2.8 with 95% credible interval (-4.5, -0.9), for HR(α)), which seems counter-intuitive. This may be due to the fact that the available hospitalisation data do not include COVID-19 counts recorded in long-term medical care centres (*Établissement d'hébergement pour personnes âgées dépendantes*, EHPAD), whereas EHPADs are populated by the elderly and were the epicentre of COVID-19 cases during the second wave (Lehot-Couette, 2020).

Similar to the results obtained for the COVID-19 cases in Montreal, the Rushworth models estimate a higher spatial dependence parameter (e.g., posterior mean of 0.93 for R(α)) than

the proposed models (e.g., posterior means of 0.52 and 0.35 for HR(α) and HR-LPC(α), respectively), indicating that a higher weight is allocated to the spatial structure when the model does not accommodate for potential outliers. Finally, the variance parameter is greater in the Rushworth model than in the proposed model, with 95% posterior credible intervals that do not overlap (e.g., (0.50, 0.54) vs (0.31, 0.36), for R(α) and HR(α), respectively).

		R(1)	R(α)	HR(1)	HR(α)	HR-LPC(1)	HR-LPC(α)
Model fit performance measures							
	WAIC	23,468.5	23,450.5	23,261.5	23,262.8	23,256.6	23,249.1
	p_W	1249.0	1246.7	1233.9	1236.4	1225.1	1222.8
MSE	Outliers	27.1	27.3	8.7	9.0	8.9	8.6
	Not outliers	12.1	12.3	31.3	29.6	28.9	27.1
	Overall	15.8	15.6	25.7	24.4	24.3	23.0
Posterior summaries for the parameters: Mean (95% CI)							
β_0		-1.47 (-1.67,-1.18)	-1.50 (-1.65,-1.33)	-1.52 (-1.74,-1.38)	-1.50 (-1.70,-1.32)	-1.50 (-1.66,-1.33)	-1.52 (-1.70,-1.36)
β		-2.77 (-5.37,-0.75)	-2.48 (-4.15,-1.03)	-2.71 (-4.07,-0.63)	-2.84 (-4.52,-0.96)	-2.88 (-4.47,-1.40)	-2.63 (-4.17,-1.07)
α		-	0.94 (0.93,0.96)	-	0.98 (0.96,0.99)	-	0.98 (0.96,0.99)
λ		0.91 (0.87,0.95)	0.93 (0.90,0.96)	0.51 (0.38,0.68)	0.52 (0.39,0.68)	0.34 (0.23,0.49)	0.35 (0.24,0.49)
σ		0.52 (0.50,0.54)	0.52 (0.50,0.54)	0.33 (0.31,0.35)	0.33 (0.31,0.36)	0.32 (0.30,0.33)	0.32 (0.30,0.34)
ν		-	-	3.03 (2.22,4.03)	3.20 (2.37,4.26)	2.06 (1.52,2.79)	1.94 (1.41,2.58)

Table 4.5: Results from the analysis of COVID-19 hospitalisations in France during the second wave (06/07/2020 – 03/01/2021). Model assessment (WAIC and MSE) and parameter posterior summaries: posterior mean and 95% credible interval (CI).

Figure 4.5 shows the distribution of the relative risk posterior means (left) and the latent effect posterior means (middle) at three time points, alongside the evolution of the total number of hospitalisations in France (right). The displayed posterior means correspond to the ones estimated by the proposed HR-LPC(α) model, which performed the best in terms of WAIC among the ones considered. Further, the detected potential outliers are indicated via coloured circles, where the colours correspond to different French regions, to help discuss

the results. A map of the French regions is available in Figure B.6 in Appendix B.4. The proposed model identifies 21 departments as potential outliers, after accounting for the fixed effects. In particular, groups of neighbouring outliers seem to be identified, corresponding to different French regions. For instance, out of the 12 departments within the French region Auvergne-Rhône-Alpes (red circles), four departments are found to be potential outliers. Similarly, half of the departments in Centre-Val de Loire (yellow circles) are found to be potential outliers. Conversely, some French regions seem to gather none of the potentially outlying departments (e.g., Hauts-de-France, Bretagne).

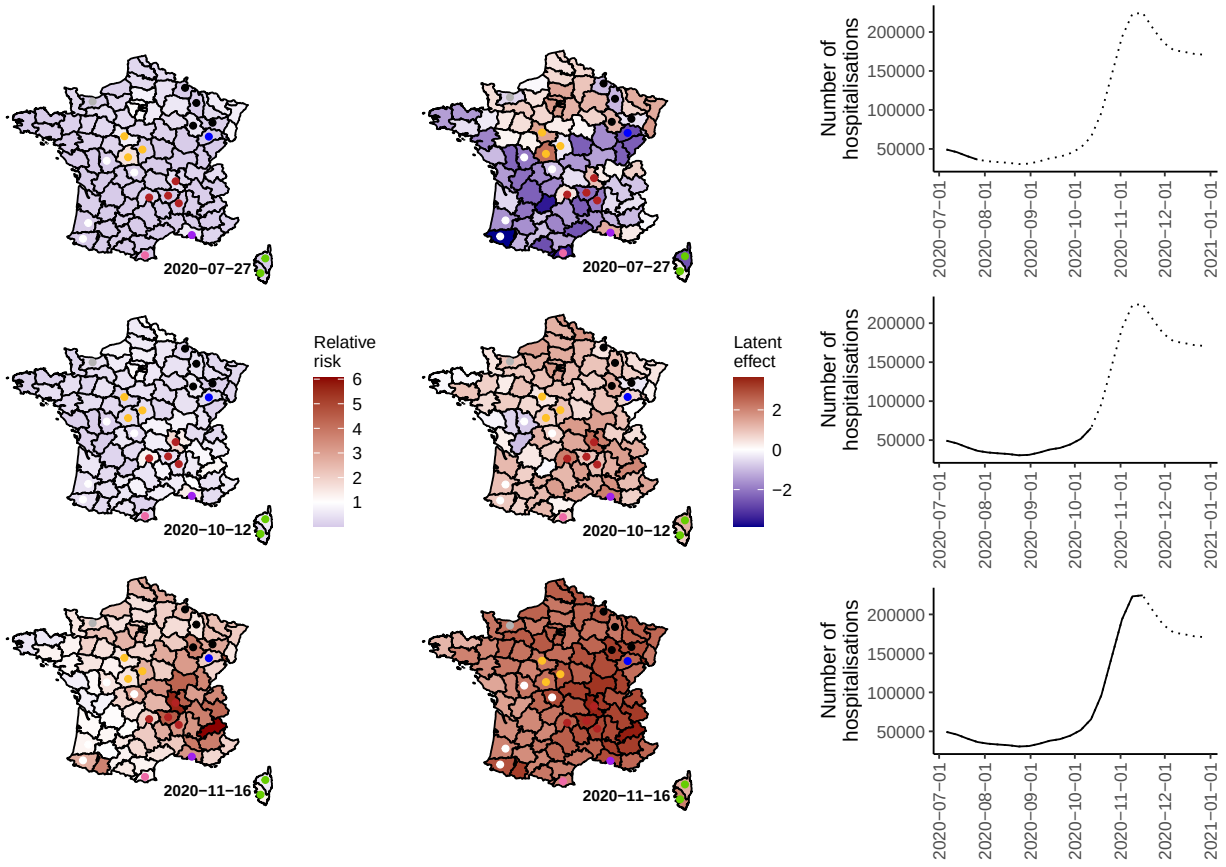


Figure 4.5: Maps of the COVID-19 relative risks (left) and latent effects (centre) estimated by the Heavy Rushworth model with spatially structured outlier indicators for three different time points across the French departments and total number of cases recorded over time in France (right). Solid coloured circles: departments identified as potential outliers by the HR-LPC(α) model, the colours correspond to the French regions to help discuss the results. The outliers are pointed out when $\kappa_u < 1$, where κ_u is the upper bound of the posterior 95% credible interval of κ.

Similarly to the results for the analysis of COVID-19 cases in Montreal, it is interesting to

note that the departments identified as potential outliers do not appear to have outlying estimated latent effects or relative risks at all time points. For instance, the Bouches-du-Rhône department (purple circle) is identified as a potential outlier and has a high estimated latent effect compared to its neighbours during the week of July 27th, 2020, but not later in the year. Similarly, at the beginning of the wave, northern Corsica (green circle) differs from its neighbours (e.g., Bouches-du-Rhône, due to the daily ferries) with a negative estimated latent effect.

Finally, it can again be noted that relative risks estimated on the tails of the distribution do not necessarily coincide with an outlier identification. For example, during the wave peak, the estimated risk of hospitalisation for the Loire department is extreme and that area is identified as a potential outlier (red circle). On the other hand, still during the wave peak, Hautes-Alpes corresponds to the other extreme estimated risk, but is not identified as a potential outlier.

4.4 Discussion

This paper proposes a spatio-temporal disease mapping model that allows for the identification of potential outliers, after accounting for fixed effects. The proposed model extends the Rushworth model (Rushworth et al., 2014) by including a scale mixture component in the conditional variance of the spatio-temporal latent effects similarly to Congdon (2017). Two prior specifications are investigated for the scaling mixture components, namely one that assumes the mixing components to follow independent gamma distributions with mean 1, and another one that allows for a spatial structure for the mixing components. It is expected (Michal et al., 2024) that the independent prior specification would perform better when the potentially outlying areas are far from each other, while situations where potential outliers are neighbours should favour the proposed model with a spatially structured prior specification. We suggest exploring both prior specifications and using WAIC to compare

which model fits best.

The results from two simulation studies help assess the ability of the proposed model to identify outlying areas in a spatio-temporal setting. The 33 boroughs of Montreal and their spatial structure were used to generate weekly data following the Rushworth model over a year (52 time points). Some areas were contaminated into being outliers. In one case, distant areas were selected, while neighbouring ones were contaminated in a second case. In both scenarios, the proposed model performed well in terms of outlier identification. In particular, when the offsets were not small, the correct areas were pointed out 100% of the time. This result agrees with the literature on disease mapping that suggests that models perform better when the offsets are large (see, e.g., Richardson et al. (2004)). The two simulation studies further showed that when outliers are neighbours, the proposed model with spatially structured scaling parameters tended to perform better than the one that assumed independent components.

This fact was further observed in the analyses conducted on COVID-19 cases and hospitalisations in Montreal and France, respectively. In the analysis of weekly COVID-19 cases recorded across the 33 boroughs of Montreal during the second wave, the proposed model with independent scaling parameters performed better in terms of WAIC than the one assuming a spatial structure, and non-neighbouring boroughs were found to be potential outliers. On the other hand, in the analysis of weekly COVID-19 hospitalisations observed during the second wave across the French departments, the proposal with spatially structured scaling parameters performed the best among the fitted ones, and groups of neighbouring departments were pointed out as potential outliers. Finally, 6 Montreal boroughs and 21 French departments are identified as potential outliers during the second wave, after accounting for the fixed effects. Further investigation of these areas might help understand why these presented potentially outlying behaviours, when compared to the rest of the regions of interest.

It is worth mentioning that the proposed model allows for spatial heteroscedasticity, but assumes the overall variability to be constant over time. Napier et al. (2016) extended the Leroux prior (Leroux et al., 1999) into the spatio-temporal setting by allowing the variance parameter to vary with time. It would therefore be interesting to allow for the scaling mixture components of the proposed model to vary with space and time, similarly to the spatio-temporal dynamic linear model of Fonseca et al. (2023).

We believe that the proposed model can be useful to decision makers during an epidemic. Identifying areas that behave differently over time compared to the rest of the region may help to understand the spread of a disease better. Additionally, the proposed model may help policymakers implement localised policies or decide where to prioritise interventions.

Funding

Michal was partially supported from an award from the Fonds de Recherche Nature et Technologies [B2X - 314857], Quebec. Schmidt is grateful for financial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada (Discovery Grant RGPIN-2017-04999) and IVADO [Fundamental Research Project, PRF-2019-6839748021].

Appendices

Supplementary material is available in Appendix B:

- B.1: Stan code for the proposed Heavy Rushworth model
- B.2: Simulation study: data generated from the proposed model
- B.3: Supplementary material for the simulation study shown in Section 4.3.1
- B.4: French regions

Chapter 5

Model-based prediction for small domains using covariates: a comparison of four methods

Preamble to Manuscript 3. When few areas are sampled and abundant auxiliary variables are available, model-based small area estimation (SAE) heavily relies on the association between the variable of interest and covariates. Model-based SAE may be performed in three different frameworks, namely the frequentist, Bayesian, and machine learning settings. However, there is no consensus on how to perform variable selection in the SAE literature. Further, inference for point estimates that result from the LASSO or random forest approaches is not straightforward, in particular when data are non-exchangeable, which is common in a SAE context. These gaps in the SAE and machine learning literatures motivated the work conducted in this manuscript.

This chapter proposes a comparison, in the context of SAE, of four popular modelling methods that handle high-dimensional auxiliary information. Further, a procedure which extends that of Lei et al. (2018) into relaxing the assumption of exchangeable data is proposed to

compute prediction intervals for complex estimates (e.g. LASSO and random forests).

The contributions of this manuscript include (i) a simulation study to compare model-based area-level frequentist, Bayesian, and machine learning approaches in a SAE context with a high number of out-of-sample areas, (ii) a new procedure to provide uncertainty quantification for complex estimates (e.g., LASSO, random forests) when data are not exchangeable, (iii) a simulation study to assess the performance of the proposed procedure to compute prediction intervals compared to the original SC method, (iv) the estimation of the mean household log consumption across all enumeration areas (EAs) in the Greater Accra Metropolitan Area (GAMA), in Ghana.

The notation in this chapter changes slightly from the previous chapters. In this thesis, thus far, we assumed that a region of interest was divided into n non-overlapping areas, indexed by $i = 1, \ldots, n$. In this chapter, we assume that a region (e.g., GAMA) is divided into M areas (e.g., EAs), which are indexed by $c = 1, \ldots, M$.

This manuscript has been accepted for publication in the *Journal of Survey Statistics and Methodology*.

Model-based prediction for small domains using covariates: a comparison of four methods

Victoire Michal[1], Jon Wakefield[2], Alexandra M. Schmidt[1], Alicia Cavanaugh[3], Brian E. Robinson[3], Jill Baumgartner[1,4].

[1]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada*

[2]*Department of Biostatistics, University of Washington, Seattle, USA*

[3]*Department of Geography, McGill University, Montreal, Canada*

[4]*Institute for Health and Social Policy, McGill University, Montreal, Canada*

Abstract

We consider methods for model-based small area estimation when the number of areas with sampled data is a small fraction of the total areas for which estimates are required. Abundant auxiliary information is available from the survey for all the sampled areas. Further, through an external source, there is information for all areas. The goal is to use auxiliary variables to predict the outcome of interest for all areas. We compare areal-level random forests and LASSO approaches to a frequentist forward variable selection approach and a Bayesian shrinkage method using a horseshoe prior. Further, to measure the uncertainty of estimates obtained from random forests and the LASSO, we propose a modification of the split conformal procedure that relaxes the assumption of exchangeable data. We show that the proposed method yields intervals with the correct coverage rate and this is confirmed through a simulation study.

This work is motivated by Ghanaian data available from the sixth Living Standard Survey (GLSS) and the 2010 Population and Housing Census, in the Greater Accra Metropolitan Area (GAMA) region, which comprises 8 districts that are further divided into enumeration areas (EAs). We estimate the areal mean household log consumption using both datasets. The outcome variable is measured only in the GLSS for 3% of all the EAs (136 out of 5019) and 174 potential covariates are available in both datasets. In the application, among the four modelling methods considered, the Bayesian shrinkage performed the best in terms of bias, MSE and prediction interval coverages and scores, as assessed through a cross-validation study. We find substantial between-area variation with the estimated log consumption showing a 1.3-fold variation across the GAMA region. The western areas are the poorest while the Accra Metropolitan Area district has the richest areas.

5.1 Motivation

In 2015, the United Nations (UN) released their 2030 agenda for sustainable development goals (SDGs) consisting of 17 goals, the first of which was to end poverty worldwide (United Nations, 2015). For their first SDG, the UN made seven guidelines explicit, including the implementation of "poverty eradication policies" at a disaggregated level. To that end, producing reliable and high resolution spatial estimates of socioeconomic status and income inequality is fundamental to help decision makers prioritise and target certain areas for decentralised interventions (Elbers et al., 2002). These detailed maps empower local communities to understand their situation compared to their neighbours, which also helps when planning interventions (Bedi et al., 2007).

In Ghana, household surveys are collected every 5 to 7 years to measure the living conditions of households across Ghanaian regions and districts and to monitor poverty. To keep track of the Ghanaian population wealth, the equivalised consumption is recorded for the sampled households. Although the household income is not directly measured, the equivalised expenditure is an alternative that allows decision makers to assess a household's standard of living (Johnson et al., 2005). This measure corresponds to the household consumption scaled by a weight based on the number of members in the household. We aim to estimate the equivalised consumption at a disaggregated level, to help communities, civil society organisations, and policymakers better understand the distribution of the households' living standard in Ghana, in order to prioritise certain areas when implementing poverty eradication policies. The sixth Ghana Living Standards Survey (GLSS), conducted in 2012-2013, was the last household survey carried out prior to the new UN SDGs agenda. The fifth GLSS had shown that inequalities had increased since 2006. In particular, although the overall poverty decreased nation-wide, the wealthiest decile of the population consumed 6.8 times more than the poorest (Cooke et al., 2016). A downside of these household surveys is that the sampling design is stratified two-stage cluster sampling, which only allows for

reliable survey sampling estimates at the district level, at best. Ghana is divided into 10 regions, which were comprised of 170 districts in 2010 or, at a finer level, around 38,000 enumeration areas (EAs). Producing reliable estimates at the EA level would further help the authorities in their policy decisions (Corral et al., 2022).

We analyse data from the sixth GLSS for the Greater Accra Metropolitan Area (GAMA), which consists of 8 districts. The GLSS used a stratified two-stage cluster sample in which strata are defined by an urban or rural indicator. Then, the clusters, which correspond to the EAs, were sampled following proportional to size sampling. Within the sampled EAs, 15 households were systematically sampled. For each sampled household, we have detailed assessment of consumption and their level of education, employment, assets, totalling 174 auxiliary variables. This gives a sample of 136 EAs out of the 5019, in this Ghanaian region. This issue of observing a small proportion of all the areas implies the need to adopt a model-based prediction approach (Pfeffermann, 2013; Tzavidis et al., 2018; Ghosh, 2020; Erciulescu and Opsomer, 2022; Hogg et al., 2023). Additionally, the sampled EAs are anonymised, which means it is unknown which 136 EAs of the 5019 EAs are represented in the survey. Finally, we have data available from the 2010 Ghanaian census for all EAs in the GAMA. Among others, the same 174 variables are measured in this census and in the sixth GLSS. The aim of this work is to produce estimates with uncertainty of the mean log household consumption at the EA level in the GAMA. Note that, as is common in the literature on poverty mapping, we focus on the equivalised consumption in the log scale to model symmetrical data (Elbers et al., 2003; Nguyen et al., 2017). If needed, the estimates could be transformed back into the original scale.

In this paper, to deal with the higher number of auxiliary variables compared to the number of sampled EAs, we assess the performance of random forests and the LASSO (which performs variable selection) to estimate the mean household log consumption at the EA level in the GAMA. For the sake of comparison, we also consider a forward variable selection

approach in the frequentist framework and a Bayesian shrinkage method using a horseshoe prior. For all four approaches, we adopt EA-level models. Due to the nature of the motivating Ghanaian datasets, where only a small proportion of the areas are sampled and are anonymised, synthetic small area estimators are of interest. Further, we propose a modification of the split conformal procedure to compute prediction intervals for the random forest and LASSO predictions while relaxing the assumption of exchangeable responses, which is necessary due to the complex sampling design.

This paper is organised as follows. Section 5.1.1 briefly reviews the literature on small area estimation (SAE) and variable selection in the frequentist, Bayesian and machine learning frameworks. Section 5.2 describes the four methods that will be compared and the proposed procedure to produce prediction intervals for estimates obtained through random forests and the LASSO. Section 5.3 shows the results from two simulation studies. First, Section 5.3.1 presents a comparison between the proposed modified split conformal and the original split conformal procedures. Then, Section 5.3.2 provides a comparison between the four methods that perform variable selection. Section 5.4 discusses the results from the four methods applied to the Ghanaian datasets. Finally, Section 5.5 concludes the paper with a discussion.

5.1.1 Literature review

SAE concerns estimation of area-level summaries when data are sparse or non-existent in the areas (Rao and Molina, 2015). This area of research in survey sampling has greatly evolved in the last 50 years (Pfeffermann, 2002; Pfeffermann, 2013; Rao and Molina, 2015; Ghosh, 2020). Tzavidis et al. (2018) points out that the use of SAE by national statistical institutes (NSIs) and other organisations to produce official statistics exhibits this increasing popularity; e.g., the `povmap` software developed by the World Bank (Elbers et al., 2003; World Bank, 2015) and the Small Area Income and Poverty Estimates project carried out by the US Census Bureau (Census Bureau, 2018).

In survey sampling, the design-based framework may be distinguished from the model-based framework. Design-based methods, also called randomisation methods, assume the variable of interest to be fixed in the finite population while the randomness comes from the sampling process. Direct (weighted) estimators have favourable design-based properties in large samples and rely only on the sampling weights and the recorded responses within each sampled area to produce areal estimates. Hence, estimates for non-sampled areas are missing. Additionally, data sparsity will yield imprecise direct estimates at the areal level. Similarly, data sparsity within areas may lead to imprecise model-assisted estimates. These latter approaches also fall under the umbrella of design-based inference. Model-assisted methods are design-based approaches which model the responses to gain precision but are still design consistent (Särndal et al., 2003). An alternative is to use model-based approaches, where the responses are no longer assumed fixed but treated as random variables which are modelled using auxiliary information and/or random effects. In model-based methods for SAE, it is common to use exterior sources of information to augment the auxiliary information from the sample to the entire finite population; for example, using information obtained from censuses. Tzavidis et al. (2018) describe a two-step approach to produce model-based small area estimates. First, a model is fitted using the survey responses and survey auxiliary variables. Then, the outcome is predicted for the entire finite population according to the estimated model parameters and finite population auxiliary information.

Abundant auxiliary information may be measured in the sample, for the sampled areas, and through exterior sources, for all the areas of the region of interest. It may therefore be necessary to select a subset of covariates to model the response variable, in the presence of high-dimensional auxiliary information. In this way, precision can be increased as unnecessary auxiliary variables are not included. The inference procedure for model-based approaches can be performed under the frequentist or Bayesian frameworks, or with flexible parametric models via machine learning techniques.

Machine learning methods are becoming more popular in the survey sampling community; see for example, Wang et al. (2014) and Breidt and Opsomer (2017). However, it is not straightforward to perform inference and assess the estimates' uncertainty under these approaches. For example, the bootstrap does not work for non-smooth targets such as LASSO estimates (Dezeure et al., 2015). Among machine learning methods, random forests (Breiman, 2001) can be fitted to unit-level or area-level data for a flexible approach. Random forests are a collection of regression trees that recursively partition the responses into increasingly homogeneous subgroups (nodes), based on covariate splits. Random forests potentially have the benefit of accommodating non-linear relationships and complex interactions, and naturally select variables through these covariate splits. Each individual regression tree is fitted on a bootstrap sample of the original dataset. There is a growing literature on methods to measure the uncertainty of random forest point estimates, for example using different Jackknife approaches (Steinberger and Leeb, 2016; Wager et al., 2014; Wager and Athey, 2018) or V-statistics (Zhou et al., 2021). However, the subsampling procedures have drawbacks, including their computational overheads and their unclear application to survey data. Recently, Zhang et al. (2019) proposed the so-called out-of-bag (OOB) prediction intervals, which are computed based on quantiles of the random forest out-of-bag prediction errors. These denote the difference between a data point's outcome and its point estimate, obtained from a random forest grown without said data point. In simulation studies, Zhang et al. (2019) show that their proposed method performs similarly to the split conformal (SC) approach proposed by Lei et al. (2018). The SC approach may be used to compute prediction intervals for any modelling method (e.g., linear models or random forests) and is a novelty in the literature on survey sampling (Bersson and Hoff, 2022; Wieczorek, 2023). To compute prediction intervals for random forest estimates through the SC method, the original dataset is first split into two datasets. A random forest is trained on one subsample, and point estimates and their associated prediction errors are obtained for the other subsample. Then, the intervals are computed based on the empirical quantiles of the prediction errors from the

second subsample. Note that while the OOB method proposed by Zhang et al. (2019) only estimates prediction intervals for random forests, the SC method can potentially be applied to any modelling procedure used to obtain point estimates. A common feature of all these prediction interval methods is that the data are assumed to be exchangeable (Angelopoulos and Bates, 2021). This is a strong assumption and is not usually true for data gathered from a complex survey design.

Inference procedures for model-based approaches can also follow the frequentist or the Bayesian paradigms. In these frameworks, variable selection is an important yet contentious research topic. In the frequentist framework, two-step procedures are common. Variables are first iteratively selected (forward selection) or removed (backward elimination) to model the outcome, based on the optimisation of some criterion (e.g., AIC, BIC, R^2). Then, a final model that includes only the selected covariates is fitted to the data. In SAE, it is common to select variables by comparing models through some criterion (for example, AIC or BIC, or survey sampling adjusted versions); see e.g., Han (2013); Rao and Molina (2015) and Lahiri and Suntornchost (2015). In the frequentist framework, regularisation methods have also been proposed in the literature, such as ridge regression and the LASSO (Tibshirani, 1996, 2011; McConville et al., 2017). These methods apply constraints to the regression parameters. However, in the case of the LASSO, these constraints yield estimates of the model parameters whose uncertainty estimation is difficult, especially in a survey setting. In a simulation study, Lei et al. (2018) show that their proposed SC method performs well in computing prediction intervals for predictions obtained through the LASSO, when the data are exchangeable.

In the Bayesian framework, variable selection is conducted by imposing informative priors on the model parameters. Multiple shrinkage priors have been proposed in the literature, for example, Bayesian ridge regression and the Bayesian LASSO (Hans, 2010). In the former, a Gaussian prior is assigned to the regression parameters, while a double-exponential

distribution is used for the latter. It can be shown that, under the respective priors, computing the maxima *a posteriori* to estimate the parameters results exactly in ridge-type and LASSO-type estimators (Reich and Ghosh, 2019). A more recent popular approach is the use of the horseshoe prior (Carvalho et al., 2010), which imposes *a priori* a heavier weight towards 0 than a normal or double-exponential distribution (Datta and Ghosh, 2013; Porwal and Raftery, 2022).

5.2 Methods

Let a region be divided into M non-overlapping areas, A_c, $c = 1, \ldots, M$. Denote by N_c the number of units in A_c, with outcomes y_{ck}, $k = 1, \ldots, N_c$. The main goal is to estimate the areal mean $\overline{y}_c = (1/N_c) \sum_{k=1}^{N_c} y_{ck}$ for all areas $c = 1, \ldots, M$, using a sample of n_c units taken from $c = 1, \ldots, m$ areas. Denote by s the set of area and household indices included in the sample and denote by s_c, $c = 1, \ldots, M$, the set of sampled units in the c-th area. Let $f_c = n_c/N_c$ be the sampling fraction within each area. For any variable a, let $\overline{a}_c = (1/N_c) \sum_{k=1}^{N_c} a_{ck}$, $c = 1, \ldots, M$, be the population areal mean, and $\overline{a}_c^{(s)} = (1/n_c) \sum_{k \in s_c} a_{ck}$, $c = 1, \ldots, m$ and $\overline{a}_c^{(ns)} = (1/(N_c - n_c)) \sum_{k \notin s_c} a_{ck} = \left(\overline{a}_c - f_c \overline{a}_c^{(s)}\right) / (1 - f_c)$, $c = 1, \ldots, M$, the areal means for the sampled (subscript (s)) and non-sampled (subscript (ns)) units, respectively. For all M areas, the estimation target may be decomposed as follows

$$\overline{y}_c = \frac{1}{N_c} \left(\sum_{k \in s_c} y_{ck} + \sum_{k \notin s_c} y_{ck} \right) = f_c \overline{y}_c^{(s)} + (1 - f_c) \overline{y}_c^{(ns)}, \; c = 1, \ldots, M. \tag{5.1}$$

To estimate $\overline{y}_c$, the non-sampled mean, $\overline{y}_c^{(ns)}$, remains to be estimated for all M areas. Let $\widehat{\overline{Y}}_c^{(ns)}$, $c = 1, \ldots, M$, be the estimator of $\overline{y}_c^{(ns)}$, $c = 1, \ldots, M$. The prediction approach estimator (Lohr, 2021) for the target of inference is

$$\widehat{\overline{Y}}_c = f_c \overline{y}_c^{(s)} + (1 - f_c) \widehat{\overline{Y}}_c^{(ns)}, \; c = 1, \ldots, M. \tag{5.2}$$

The uncertainty of $\widehat{\overline{Y}}_c$ may be measured using prediction intervals of level $(1-\alpha)\%$, $\text{PI}_{(1-\alpha)}$, of the form

$$\text{PI}_{(1-\alpha)\%}\left[\widehat{\overline{Y}}_c\right] = f_c\overline{y}_c^{(s)} + (1-f_c)\text{PI}_{(1-\alpha)\%}\left[\widehat{\overline{Y}}_c^{(ns)}\right],\ c=1,\ldots,M. \tag{5.3}$$

Note that for a non-sampled area c', $f_{c'}=0$ and the estimator reduces to $\widehat{\overline{Y}}_{c'} = \widehat{\overline{Y}}_{c'}^{(ns)}$, with prediction interval, $\text{PI}_{(1-\alpha)\%}\left[\widehat{\overline{Y}}_{c'}\right] = \text{PI}_{(1-\alpha)\%}\left[\widehat{\overline{Y}}_{c'}^{(ns)}\right]$.

Random forests and the LASSO are considered to estimate $\widehat{\overline{Y}}_c^{(ns)}$ in the model-based framework. For the sake of comparison, we also consider a forward variable selection approach in the frequentist paradigm and a Bayesian shrinkage method. The four modelling approaches assume there are p-dimensional covariates available from the sample, $\{\boldsymbol{x}_{ck},\ c,k\in s\}$, as well as areal covariate means, $\overline{\boldsymbol{x}}_c,\ c=1,\ldots,M$, which are known for all the areas of the finite population. Such information may be obtained from a census. Inference is carried out at the areal level in all four methods. In this model-based framework, at the unit level, the finite population response values y are assumed to be an independent and identically distributed (i.i.d.) realisation of super population random variables Y whose sampled and non-sampled moments are, at the area level:

$$\begin{aligned}
\mathbb{E}\left(\overline{Y}_c^{(s)}\right) &= \mu\left(\overline{\boldsymbol{x}}_c^{(s)}\right), & \mathbb{V}\left(\overline{Y}_c^{(s)}\right) &= \sigma^2/n_c,\\
\mathbb{E}\left(\overline{Y}_c^{(ns)}\right) &= \mu\left(\overline{\boldsymbol{x}}_c^{(ns)}\right), & \mathbb{V}\left(\overline{Y}_c^{(ns)}\right) &= \sigma^2/(N_c-n_c).
\end{aligned} \tag{5.4}$$

Therefore, inference is conducted using $\left\{\left(\overline{y}_c^{(s)}, \overline{\boldsymbol{x}}_c^{(s)}\right),\ c=1,\ldots,m\right\}$ and the non-sampled mean predictions, $\widehat{\overline{Y}}_c^{(ns)},\ c=1,\ldots,M$, are computed using the available covariates' non-sampled means, $\overline{\boldsymbol{x}}_c^{(ns)},\ c=1,\ldots,M$.

5.2.1 Random forest and LASSO approaches

First, we consider a random forest prediction approach. This non-parametric method makes no further assumption to Model (5.4). Following Breiman (2001), random forest point estimates are the average over B point estimates obtained from training B independent regression trees on B bootstrap versions of the original sample. Each regression tree partitions the bootstrap response values based on splitting rules applied to covariates. A random forest algorithm is described in appendix C.3. Hence, with a random forest procedure, the estimator (5.2) becomes

$$\widehat{\overline{Y}}_c = f_c \overline{y}_c^{(s)} + (1 - f_c) \left(\sum_{c'=1}^{m} w_{c'}(\overline{\boldsymbol{x}}_c^{(ns)}) \overline{y}_{c'}^{(s)} \right), \tag{5.5}$$

where the weights $w_{c'}(\cdot)$ result from the random forest procedure described in appendix C.3.

Second, we consider the LASSO to predict the areal non-sampled means, assuming a linear relationship between the covariates and the outcome, while performing variable selection. The estimation via the LASSO applies to p-dimensional regression coefficients $\boldsymbol{\beta}$, resulting in $\widehat{\boldsymbol{\beta}}^{\text{LASSO}}$ by solving $\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \left\| \overline{\boldsymbol{y}}^{(s)} - \overline{\boldsymbol{x}}^{(s)} \boldsymbol{\beta} \right\|_2^2 / (2m) + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\}$, $\lambda \geq 0$, where $\overline{\boldsymbol{y}}^{(s)} = \left[\overline{y}_1^{(s)}, \ldots, \overline{y}_m^{(s)} \right]^\top$ and $\overline{\boldsymbol{x}}^{(s)} = \left[\overline{\boldsymbol{x}}_1^{(s)\top}, \ldots, \overline{\boldsymbol{x}}_m^{(s)\top} \right]^\top$. Note that the shrinkage penalty parameter λ is fixed after a 10-fold cross-validation, seeking the smallest test MSE. Therefore, the estimator (5.2) becomes

$$\widehat{\overline{Y}}_c = f_c \overline{y}_c^{(s)} + (1 - f_c) \left[\overline{\boldsymbol{x}}_c^{(ns)\top} \widehat{\boldsymbol{\beta}}^{\text{LASSO}} \right]. \tag{5.6}$$

To measure the uncertainty associated to the random forest and LASSO predictions (5.5) and (5.6), we propose a modification of the SC prediction intervals of Lei et al. (2018) which relaxes the assumption of exchangeable sampled and non-sampled data points. The original SC procedure assumes $\overline{Y}_c^{(s)}$ and $\overline{Y}_c^{(ns)}$ to be exchangeable. However, as shown in (5.4), $\overline{Y}_c^{(s)}$ and $\overline{Y}_c^{(ns)}$ are not exchangeable. Hence, in the proposed modified SC procedure, we assume

the mean structures to be similar and allow the variances to be scaled differently, as is the case in (5.4). Specifically, in this context of a complex sampling design, we assume the variance is independent of the sample strata. The unit-level variance, σ^2, is assumed fixed across the strata and the sampled and non-sampled areal-level variances only vary with the number of sampled and non-sampled units, n_c and $N_c - n_c$, respectively. We propose to scale the residuals computed in the original SC procedure before computing the empirical quantile necessary to the prediction intervals. Said quantile is then scaled when computing the prediction intervals. The proposed scaled SC procedure can be described through the following steps:

1. Randomly split $\left\{\left(\overline{y}_c^{(s)}, \overline{\boldsymbol{x}}_c^{(s)}\right),\ c = 1, \ldots, m\right\}$ into two equal sized datasets. Denote by S_1 and S_2 the resulting two sets of area indices;
2. Train a random forest or a LASSO approach on $\left\{\left(\overline{y}_c^{(s)}, \overline{\boldsymbol{x}}_c^{(s)}\right),\ c \in S_1\right\}$ and predict $\left\{\widehat{\overline{Y}}_c^{(S_2)},\ c \in S_2\right\}$;
3. Compute the scaled absolute residuals $R_c = \sqrt{n_c} \times \left|\overline{y}_c^{(s)} - \widehat{\overline{Y}}_c^{(S_2)}\right|,\ c \in S_2$;
4. Find d_α, the k_αth smallest residual R, for $k_\alpha = \lceil (m/2+1)(1-\alpha) \rceil$;
5. Let the prediction interval be $\text{PI}_{(1-\alpha)\%}\left[\widehat{\overline{Y}}_c^{(ns)}\right] = \widehat{\overline{Y}}_c^{(ns)} \pm d_\alpha / \sqrt{N_c - n_c},\ c = 1, \ldots, M$.

Hence, for random forest or LASSO predictions (5.5) or (5.6), the uncertainty (5.3) becomes

$$\text{PI}_{(1-\alpha)\%}\left[\widehat{\overline{Y}}_c\right] = \widehat{\overline{Y}}_c \pm (1 - f_c)\frac{d_\alpha}{\sqrt{N_c - n_c}}.$$

Appendix C.1 provides a proof of the $(1-\alpha)\%$ marginal coverage of the proposed scaled split conformal prediction intervals.

5.2.2 Forward selection

As a comparison, we consider a frequentist method with the commonly used forward approach with AIC as a variable selection criterion. Model (5.4) is completed by assuming the errors are normally distributed. To predict $\widehat{\overline{Y}}_c^{(ns)}$, the forward approach is a two-step procedure. First, a subset of K covariates $\boldsymbol{z}$ is selected among the available $\boldsymbol{x}$'s. To that end, linear models are iteratively fitted, adding one covariate at a time based on the resulting AIC value. Then, using the selected covariates, a linear model is fitted: $\overline{y}_c^{(s)} \sim \mathcal{N}\left(\overline{\boldsymbol{z}}_c^{(s)\top}\boldsymbol{\eta}, \sigma^2/n_c\right)$, $c = 1, \ldots, m$, to estimate $\widehat{\boldsymbol{\eta}}$, $\widehat{\mathbb{V}}(\widehat{\boldsymbol{\eta}})$ and $\widehat{\sigma}$. The steps required to run this forward approach are detailed in Appendix C.2. The estimator and uncertainty (5.2) and (5.3) become , respectively,

$$\widehat{\overline{Y}}_c = f_c\overline{y}_c^{(s)} + (1-f_c)\left[\overline{\boldsymbol{z}}_c^{(ns)\top}\widehat{\boldsymbol{\eta}}\right],$$
$$\text{PI}_{(1-\alpha)\%}\left[\widehat{\overline{Y}}_c\right] = \widehat{\overline{Y}}_c \pm q_\alpha(1-f_c)\sqrt{\overline{\boldsymbol{z}}_c^{(ns)\top}\widehat{\mathbb{V}}(\widehat{\boldsymbol{\eta}})\,\overline{\boldsymbol{z}}_c^{(ns)} + \frac{\widehat{\sigma}^2}{N_c - n_c}},$$

where q_α denotes the α-level quantile from a $\mathcal{N}(0,1)$ distribution. Note that uncertainty in the covariates selected is not accounted for.

5.2.3 Bayesian shrinkage approach

Finally, a Bayesian approach is considered, where all the available covariates, $\boldsymbol{x}$, are used in a single step to model the outcome while applying the horseshoe prior (Carvalho et al., 2010) to the regression parameters. Similar to the forward approach, a normal distribution is further assumed for Model (5.4). The observed sampled means are modelled through $\overline{y}_c^{(s)} \sim \mathcal{N}(\overline{\boldsymbol{x}}_c^{(s)\top}\boldsymbol{\beta}, \sigma^2/n_c)$, $c = 1, \ldots, m$, with priors $\beta_j \sim \mathcal{N}(0, \lambda_j^2\tau^2)$, $j = 1, \ldots, p$, and σ, τ, $\lambda_1, \ldots, \lambda_p \sim \mathcal{HC}(0,1)$, where $\mathcal{HC}(a,b)$ stands for a half-Cauchy distribution with location a and scale b. In this prior, τ corresponds to the global shrinkage and λ_j, to the local shrinkage. Then, inference is conducted through the posterior distribution, which is

approximated through a Markov Chains Monte Carlo (MCMC) procedure. The estimator and its uncertainty (5.2) and (5.3) become

$$\widehat{\overline{Y}}_c = f_c \overline{y}_c^{(s)} + (1 - f_c) \left(\frac{1}{L} \sum_{\ell=1}^{L} \widehat{\overline{Y}}_c^{(ns)(\ell)} \right),$$
$$\text{PI}_{(1-\alpha)\%} \left[\widehat{\overline{Y}}_c \right] = f_c \overline{y}_c^{(s)} + (1 - f_c) \left[\widehat{\overline{Y}}_{c,\text{lower}_\alpha}^{(ns)}, \widehat{\overline{Y}}_{c,\text{upper}_\alpha}^{(ns)} \right],$$

where $\widehat{\overline{Y}}_c^{(ns)(\ell)} \sim \mathcal{N}\left(\overline{\boldsymbol{x}}_c^{(ns)\top} \boldsymbol{\beta}^{(\ell)}, \sigma^{(\ell)2}/(N_c - n_c) \right)$, $\ell = 1, \ldots, L$, $c = 1, \ldots, M$, is the ℓth element of the MCMC posterior predictive sample, with $\boldsymbol{\beta}^{(\ell)}$ and $\sigma^{(\ell)}$ the ℓth elements in the MCMC samples. The α-level empirical quantiles from the posterior predictive sample are denoted $\widehat{\overline{Y}}_{c,\text{lower}_\alpha}^{(ns)}$ and $\widehat{\overline{Y}}_{c,\text{upper}_\alpha}^{(ns)}$.

5.3 Simulation study

This section presents two simulation studies to assess the performance of the proposed scaled SC procedure and to compare the four modelling methods. Section 5.3.1 focuses on the proposed scaled SC method that computes prediction intervals while relaxing the assumption of exchangeable data points. In Section 5.3.2, different generating models and sampling designs are studied to compare the four model selection methods described in Section 5.2.

Inference is performed in `R`. The random forests of $B = 1000$ trees are trained using the `ranger` package (Wright et al., 2022). For each simulation scenario, the random forest hyperparameters are fixed after a cross-validation study of different values. The code to conduct the proposed scaled SC procedure for random forest estimates is available in appendix C.4. The LASSO method is conducted through the `glmnet` package, using the `cv.glmnet` function to define the optimal shrinkage penalty parameter. Bayesian inference is performed with the `NIMBLE` package (de Valpine et al., 2017). Convergence of the MCMC chains is assessed through trace plots, effective sample sizes and the $\widehat{R}$ statistic (Gelman and Rubin,

1992).

5.3.1 Simulation study: scaled split conformal procedure

To assess the performance of the proposed scaled SC procedure, five model-based simulation scenarios are considered: $R = 500$ finite populations are created, and the different simulation scenarios correspond to the various sampling designs applied to that finite population. Assume that each finite population consists of $M = 500$ areas of sizes $N_c,\ c = 1, \ldots, M$, with $\min_c(N_c) = 50$ and $\max_c(N_c) = 500$. For $c = 1, \ldots, M$, and $k = 1, \ldots, N_c$, the response variable has distribution

$$y_{ck} \overset{i.i.d.}{\sim} \mathcal{N}(9.5 + x_{1,ck} - x_{2,ck} + 2x_{3,ck} - x_{4,ck} + 2x_{5,ck} + x_{6,ck}, 1),$$

with 6 unit-level covariates, $x_1, \ldots, x_6 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, which are the same across the finite populations. In this model-based framework, the same areas and units are sampled across the different finite populations. From each finite population, a sample is drawn according to five sampling designs, which constitute the simulation scenarios:

1. (Stratified) Select all $m = M = 500$ areas and within each area, sample $n_c = 0.5N_c,\ c = 1, \ldots, m$ units;
2. (Stratified) Select all $m = M = 500$ areas and within each area, sample $n_c = 0.7N_c,\ c = 1, \ldots, m$ units;
3. (One-stage) Sample $m = M/2$ areas and within each area, select all $n_c = N_c,\ c = 1, \ldots, m$ units;
4. (Two-stage) Sample $m = M/2$ areas and within each area, sample $n_c = 0.5N_c,\ c = 1, \ldots, m$ units;
5. (Two-stage) Sample $m = M/2$ areas and within each area, sample $n_c = 0.7N_c,\ c = 1, \ldots, m$ units.

The proportion of sampled areas is higher in the stratified sampling designs, as all areas are selected. Hence, the areal-level inference is conducted on more data points in the first two scenarios than in the remaining three, and we expect any modelling method to perform better in these two scenarios. The one-stage and two-stage designs all yield $m = 500$ areal-level responses. The difference between these last three scenarios is in the sampling fraction within areas. Out of the five simulation scenarios considered, the fourth one is the closest to the Ghanaian data analysed in Section 5.4.

For each simulation scenario and in each sample, the estimates described in equation (5.2) are computed using four methods: a linear model that includes the correct six covariates, a linear model that omits x_4, x_5 and x_6, a random forest method that considers all six covariates to grow the trees, and a linear LASSO model. The random forest hyperparameters are set after a cross-validation study as $mtry = 2$ and $nodesize = 5$. For each scenario, in each sample and for each modelling method, 50%, 80% and 95% prediction intervals (5.3) are computed following the SC and the proposed scaled SC procedures. These non-parametric methods may be applied to any modelling approach: a linear regression method as well as a machine learning method. The objective of this simulation study is to assess whether the proposed scaled SC method yields valid coverage rates of the prediction intervals.

All results are shown in Figure 5.1. The observed coverages for each scenario, method and interval level are shown in the left panel and the interval widths, in the right panel. The simulation scenarios are identified by their 1-5 number, as described above. Further, we differentiate the results for the sampled and non-sampled areas (Yes/No, respectively). Scenarios 1 and 2 only present results for sampled areas, as all areas are sampled in a stratified sample. Scenario 3 only shows the results for non-sampled areas because the target is exactly estimated in the sampled areas since all units are selected in a one-stage sample. Therefore, in the third scenario, we measure the predictions' uncertainty only in the non-sampled areas.

In the first scenario and for the sampled areas in the fourth scenario, n_c and $N_c - n_c$ are equal. Therefore, the data points are exchangeable and the SC and proposed scaled SC approaches are the same. In these scenarios, both methods yield the right coverages of the prediction intervals, regardless of the modelling method. In terms of interval widths, the linear model with incorrect set of covariates leads to the widest intervals under both SC procedures. The linear model with correct covariates and the LASSO method yields the narrowest intervals regardless of the SC method.

For all other sampling schemes, $n_c \neq N_c - n_c$. Therefore, the sampled and non-sampled means, $\overline{Y}_c^{(s)}$ and $\overline{Y}_c^{(ns)}$, are not exchangeable, with differently scaled variances, σ^2/n_c and $\sigma^2/(N_c - n_c)$, respectively. In these cases, the original SC intervals obtained for all four modelling methods do not attain the right coverages. The original SC leads to under-coverage of the prediction intervals. On the other hand, the proposed scaled SC procedure produces prediction intervals with the right coverages, regardless of the interval level and modelling method. In particular, when fitting a linear model, with the LASSO constraint or with the right mean structure, the scaled SC intervals have exactly the right coverages. When modelling with a non-parametric random forest approach or with a linear model assuming the wrong mean structure, the scaled SC prediction intervals show a slight error in the coverage rate.

In terms of interval width, the proposed scaled SC intervals tend to be a little wider than the original SC ones, for all interval levels. Regardless of the simulation scenario, the SC intervals and proposed scaled SC intervals obtained for the random forest estimates tend to be narrower than the ones obtained for the linear estimates using the incorrect set of covariates.

Finally, Appendix C.5 shows the results from the equivalent simulation study in the design-based framework. In this case, the finite population is assumed fixed and different samples are taken. Similar results are obtained: when $n_c \neq N_c - n_c$, the proposed scaled SC proce-

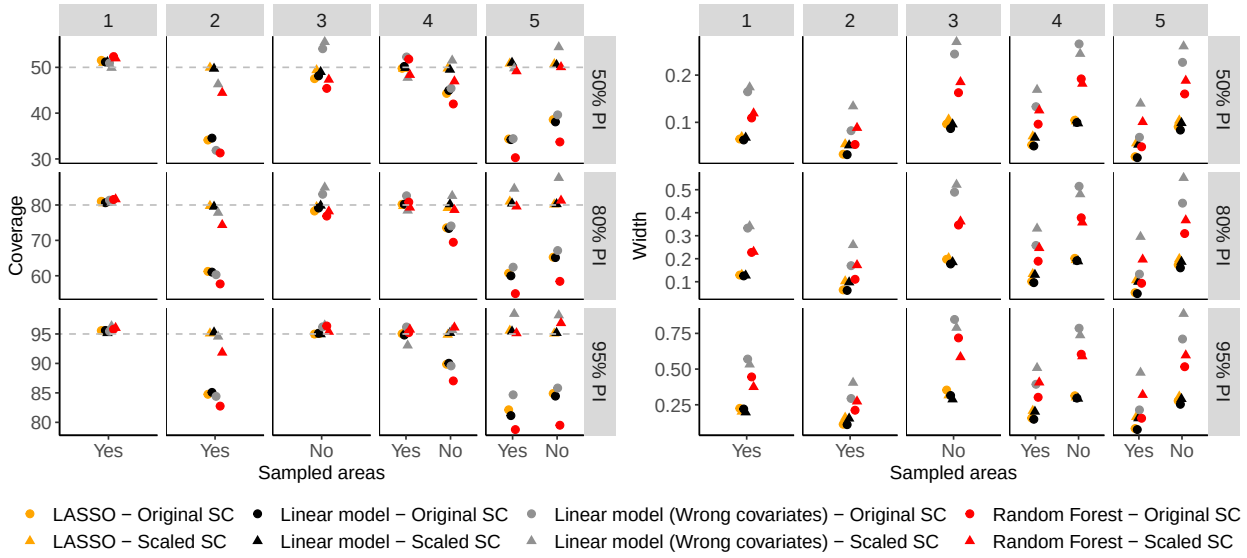


Figure 5.1: Coverages and widths of the prediction intervals (PI) obtained from the proposed scaled and original split conformal (SC) procedures for the four modelling methods and across the five simulation scenarios (1-5). Yes: coverages and widths across the sampled areas; No: coverages and widths across the non-sampled areas.

dure yields the correct coverages, while the original SC approach produces under-covering prediction intervals.

5.3.2 Simulation study: prediction methods comparison

To compare the performance of the random forest and the LASSO methods to the frequentist forward variable selection and the Bayesian shrinkage approaches, as described in Section 5.2, a model-based simulation study is conducted considering three generating models for the outcome and five sampling designs. Similar to Section 5.3.1 and Appendix C.5, a design-based counterpart to this simulation study is shown in Appendix C.6, producing similar results. We replicate $R = 100$ times the creation of three finite populations of $M = 1000$ areas of sizes N_c with $\min_c(N_c) = 50$ and $\max_c(N_c) = 500$ as follows::

A. $y_{ck} \sim \mathcal{N}\left(20 + \boldsymbol{x}_{ck}^{\top}\boldsymbol{\beta}, 0.5^2\right)$, where the covariates are such that $\boldsymbol{x}_{ck} \sim \mathcal{N}_{100}(\mathbf{0}, \boldsymbol{I})$ and with coefficients $\boldsymbol{\beta}^{\top} = (1, -1, 2, -1, 2, 1, 2, 1, -1, 1, 0, \ldots, 0)$;

B. $y_{ck} \sim \mathcal{N}\left(20 + \boldsymbol{x}_{ck}^{\top}\boldsymbol{\beta}, 0.5^2\right)$, where the covariates are such that $\boldsymbol{x}_{ck} \sim \mathcal{N}_{100}(\mathbf{0}, \Sigma_x)$, with

$$\Sigma_x = \begin{bmatrix} 1 & 0.5 & \dots & 0.5 \\ 0.5 & 1 & \dots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & \dots & 1 \end{bmatrix}, \text{ and } \boldsymbol{\beta}^\top = (1, -1, 2, -1, 2, 1, 2, 1, -1, 1, 0, \dots, 0)/10;$$

C. $y_{ck} \sim \mathcal{N}\left(x_{1,ck}^2 + \exp\left(x_{2,ck}^2\right), 0.3\right)$, with covariates $x_{j,c} \sim \mathcal{U}(-1, 1)$, $j = 1, \dots, 100$.

The covariates are the same across the replicated populations and the randomness comes from the y's. Populations A and B assume a linear relationship between the outcome and the first 10 covariates. In scenario B, however, the strength of the association is weak and the covariates are correlated. Population C is inspired by Scornet (2017) and assumes a non-linear relationship between the outcome and covariates. Throughout this simulation study, areas are indiscriminately termed "areas" or "EAs". From each finite population, a sample is drawn following the two sampling schemes:

1. (Stratified) Select all $m = M = 500$ areas and within each area, sample $n_c = 15$, $c = 1, \dots, m$, units;

2. (Two-stage) Sample $m = M/2$ areas and within each area, sample $n_c = 15$, $c = 1, \dots, m$, units.

These simulation scenarios were motivated by the Ghanaian data analysed in Section 5.4, where only 15 households were sampled within the selected areas. Additional sampling designs are considered in Appendix C.7, where a higher number of sampled units within the selected areas, and in Appendix C.8, where a simulation study aims to emulate the Ghanaian data analysed in Section 5.4. For each scenario, the estimates and their prediction intervals are computed as described in Section 5.2. Further, for each scenario, the estimates and their uncertainty are also computed assuming anonymised EAs. In this context, the modelling methods are trained on the sample and predictions are obtained ignoring which areas have been sampled, that is, assuming $f_c = 0$, $c = 1, \dots, M$, at the prediction stage. Once again, this study with anonymised EAs is run because the available Ghanaian data

that is analysed in Section 5.4 does not identify the sampled EAs. The random forest hyperparameters are set following a cross-validation study conducted for each simulation scenario. A random forest with hyperparameters set to $(mtry, nodesize) = (10, 5)$ is found to perform the best for both sampling schemes in populations A and B. In population C, we set $(mtry, nodesize) = (70, 200)$ and $(mtry, nodesize) = (70, 9)$, for the stratified samples and the two-stage samples, respectively. In all scenarios, the Bayesian approach runs through a MCMC procedure with two chains of 5,000 iterations, which include a burn-in period of 2,500 iterations. We find out that with these values, practical convergence was attained, as assessed by the trace plots, effective sample sizes and $\widehat{R}$ statistics (Gelman and Rubin, 1992). For a particular finite population and sampling scheme, running the random forest over 100 replicates takes 55 minutes, while the LASSO takes 4 minutes, the forward method takes 7 minutes and the Bayesian approach, 2.5 hours.

The methods' performances are compared via four measures: the mean absolute bias $\text{AB} = (1/R)\sum_{r=1}^{R}\sum_{c=1}^{M}\left|\widehat{\overline{Y}}_c - \overline{y}_c\right|/M$, mean squared error $\text{MSE} = (1/R)\sum_{r=1}^{R}\sum_{c=1}^{M}\left(\widehat{\overline{Y}}_c - \overline{y}_c\right)^2/M$, coverages of 50%, 80% and 95% prediction intervals $\text{Cov}_{(1-\alpha)\%} = (1/R)\sum_{r=1}^{R}\sum_{c=1}^{M}\mathbb{1}_{\overline{y}_c\in[L,U]}/M \times 100$, with $[L, U]$ the $(1-\alpha)\%$ prediction interval, and their proper interval score $\text{S}_{(1-\alpha)\%} = (1/R)\sum_{r=1}^{R}\sum_{c=1}^{M}\left((U-L) + 2/\alpha\left[(L-\overline{y}_c)\mathbb{1}_{L>\overline{y}_c} + (\overline{y}_c - U)\mathbb{1}_{U<\overline{y}_c}\right]\right)/M$ (Gneiting and Raftery, 2007). Smaller values of the interval proper scores are preferred, indicating narrow intervals and average close to the nominal. Additionally, we extract which covariates have been selected from each method. Note that in the Bayesian framework, a covariate is said to be selected when its coefficient's posterior 95% credible interval does not include 0. For the random forest approach, when the p-value related to a variable's importance (Altmann et al., 2010) is smaller than 0.05, said variable is deemed selected. The variable importance is computed based on results from random forests fitted with permutations of the set of covariates.

Figure 5.2 shows the selected covariates by each method for each model and sampling design.

When the association is linear between the covariates and the outcome (A and B), regardless of the sampling design, the forward approach tends to adequately select the true auxiliary information. However, it also tends to select irrelevant variables. Each unimportant covariate is selected about 20% of the time by the forward method. In scenario A, the LASSO and Bayesian approaches also select the right covariates 100% of the time, while rarely including redundant covariates. When the covariates are correlated, the LASSO and Bayesian methods tend to miss the right covariates 20%– 70% of the time, depending on the sampling design. In scenarios A and B, the random forest method misses the right covariates 10%–50% of the time, while it always captures the correct set when the association is non-linear. The LASSO selects 1 out of 2 correct covariates about 80% of the time in this third population. Both the forward and Bayesian approaches miss the correct set of covariates in scenario C almost 100% of the time and include irrelevant variables.

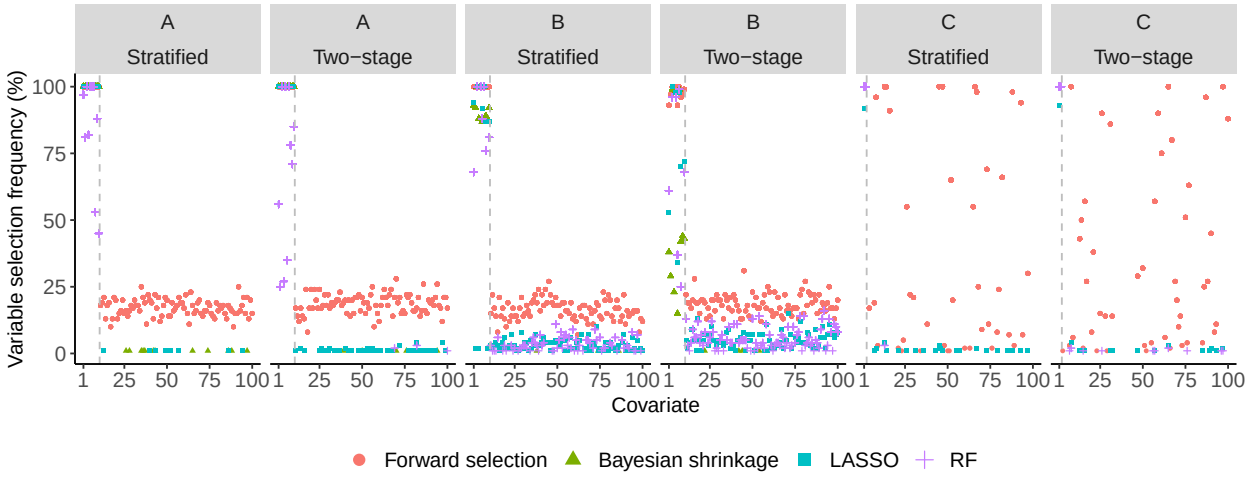


Figure 5.2: Covariate selection frequency for each method across the 6 simulation scenarios. Left of the vertical dashed line: true covariates used in the generating models.

Figure 5.3 shows the absolute biases multiplied by 100, MSEs, prediction intervals' coverages and proper scores for all methods, generating models (A-C) and sampling schemes. The results for the sampled and non-sampled areas are differentiated through the red and black symbols, respectively. Only results for the sampled EAs are produced for the stratified sampling design, as all areas are sampled. The results assuming the anonymised EAs

are distinguished from the ones in which we know which areas have been sampled by the circle and cross symbols, respectively. The results shown in Figure 5.3 are also provided in Appendix C.9, in Tables C.2 – C.9.

For all performance measures, the four modelling approaches yield similar results when it is known and unknown which areas have been sampled. For example, in population C with a two-stage sampling design and regardless of the modelling method, the MSE results over the anonymised sampled EAs are not worse than the results over the non-sampled EAs. This result is reassuring as for analysing the Ghanian data, where the sampled EAs are anonymised.

In terms of bias, all methods are virtually unbiased with mean absolute biases between 0 and 0.08, regardless of the population and sampling design. In the linear scenarios, the random forest tends to yield slightly higher mean absolute biases, compared to the forward, LASSO and Bayesian methods, which is sensible as the correct mean structure cannot be estimated in this non-parametric approach.

In terms of MSE, there does not seem to be a difference between the LASSO, forward and Bayesian methods, for all simulation models and sampling schemes. These three methods, which fit linear models, yield slightly smaller MSEs than the random forest approach when the association between the covariates and the outcome is strongly linear (A). For scenario C, however, the random forest produces smaller MSEs than the three linear modelling approaches. The random forest method divides the other three modelling methods' MSEs by a factor of 3 in population C, regardless of the sampling scheme. This result is explained by the fact that the random forest approach is a non-parametric method that adapts better to the non-linear setting.

The prediction intervals computed for all four methods in each sampling scheme for populations A and B yield the right coverages, with a slight under-coverage for the random forest approach. This may be due to the incorrect mean structure that is fitted to the data. These

intervals are wider for the random forest method in model A, for both sampling designs, as deduced from the proper interval scores. When the relationship between the outcome and the covariates is non-linear (C), we observe that all four modelling methods yield under-coverage in both sampling designs. The random forest method, which accommodates a non-linear relationship, leads to prediction intervals with slightly higher coverage rates than the other three methods, in particular across the sampled areas, but still misses the targetted rates by about 30%. Note, however, that the random forest approach produces prediction intervals with smaller proper scores than the other three modelling methods.

5.4 Areal log consumption prediction in the Greater Accra Metropolitan Area

In this section, the four modelling methods described in Section 5.2 are applied to the data for the Greater Accra Metropolitan Area (GAMA) in Ghana. Using the sixth GLSS and the 2010 Ghanaian census, a complete map of the mean equivalised consumption (in the log scale) is produced across the $M = 5019$ enumeration areas (EAs), for each method. Note that in the household survey, only $m = 136$ EAs have been sampled. To provide estimates in the missing areas, the response values are modelled using the $p = 174$ auxiliary variables which are measured in both available datasets and have been scaled for computational efficiency.

For the random forest approach, due to the small sample size m and following Hastie et al. (2009), a cross-validation study on the survey data was run to set the hyperparameters to $B = 1000$ trees grown with $mtry = 25$ and $nodesize = 3$. The Bayesian approach required two MCMC chains of 100,000 iterations, including a burn-in period of 50,000, and a thinning factor of 15. Convergence was attained as assessed by the trace plots, effective sample sizes and $\widehat{R}$ statistics.

Wakefield et al. (2020) point out the importance of including the design variables in model-

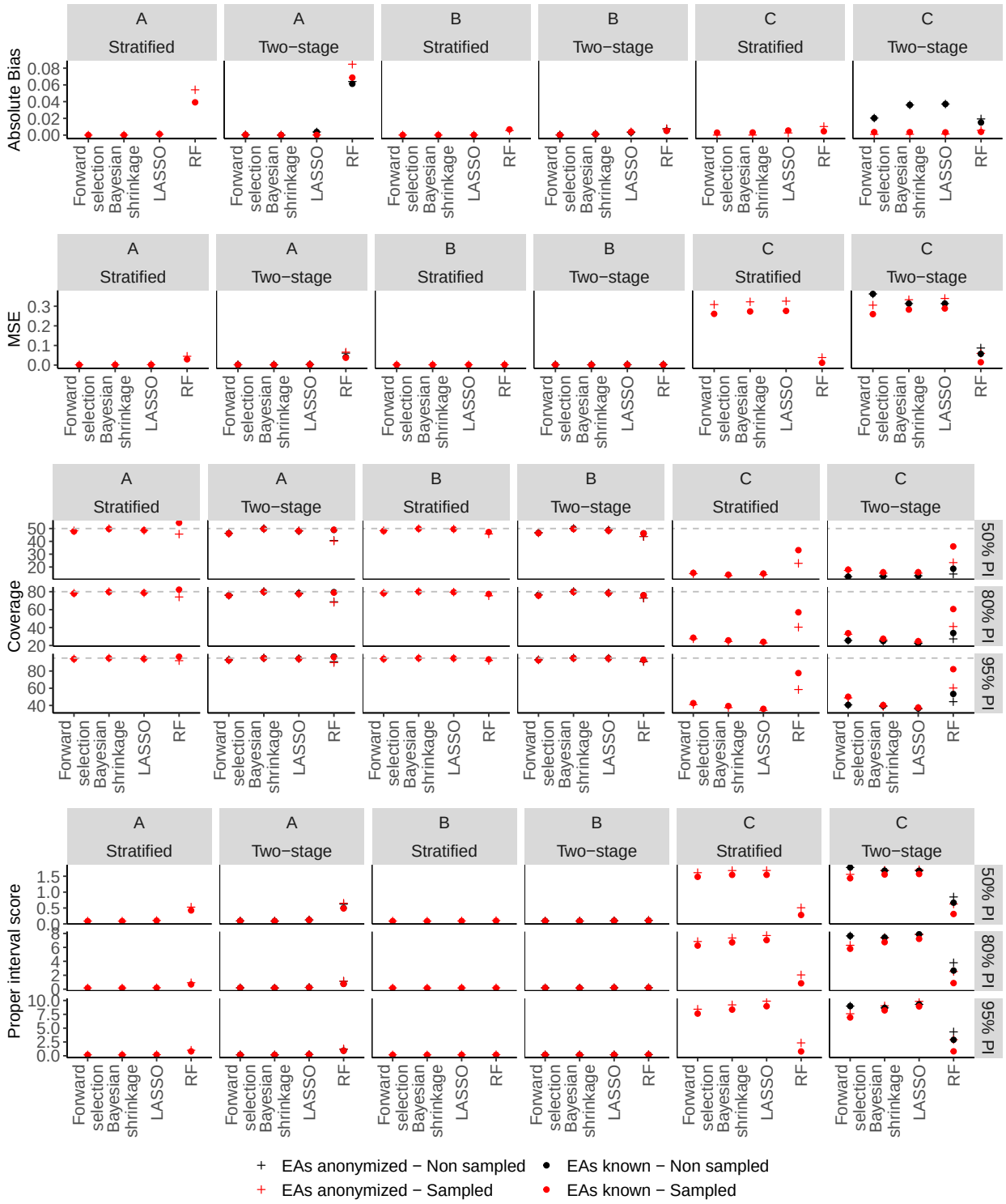


Figure 5.3: Mean absolute bias, MSE, coverages and proper scores of the prediction intervals, obtained for each method across the 6 simulation scenarios. RF: Random forest approach.

based small area estimation methods. To that end, the urban indicator, which corresponds to the sample strata, is added to all four modelling methods. In the forward selection approach, this inclusion means that the urban indicator is added to the vector of selected covariates, even if it was not selected in the first step. In the Bayesian shrinkage and LASSO methods, it means there is no shrinkage applied to the regression coefficient that corresponds to the urban indicator. Finally, in the random forest approach, it means that the urban indicator is part of the variables considered for each covariate split. Figure 5.4 presents the covariates that were selected by each method. Despite all methods including the urban indicator, only the random forest finds it relevant with a p-value for its variable importance smaller than 0.05. Additionally, Figure 5.4 shows that the horseshoe prior leads to only one variable whose coefficient's posterior 95% credible interval does not include 0. The LASSO approach selects about 6% of the available covariates (11 variables), while the forward method and random forest methods select more than 12% of the variables (21 and 22, respectively). The variable indicating whether a household's floor is made of cement or concrete is selected by all four methods.

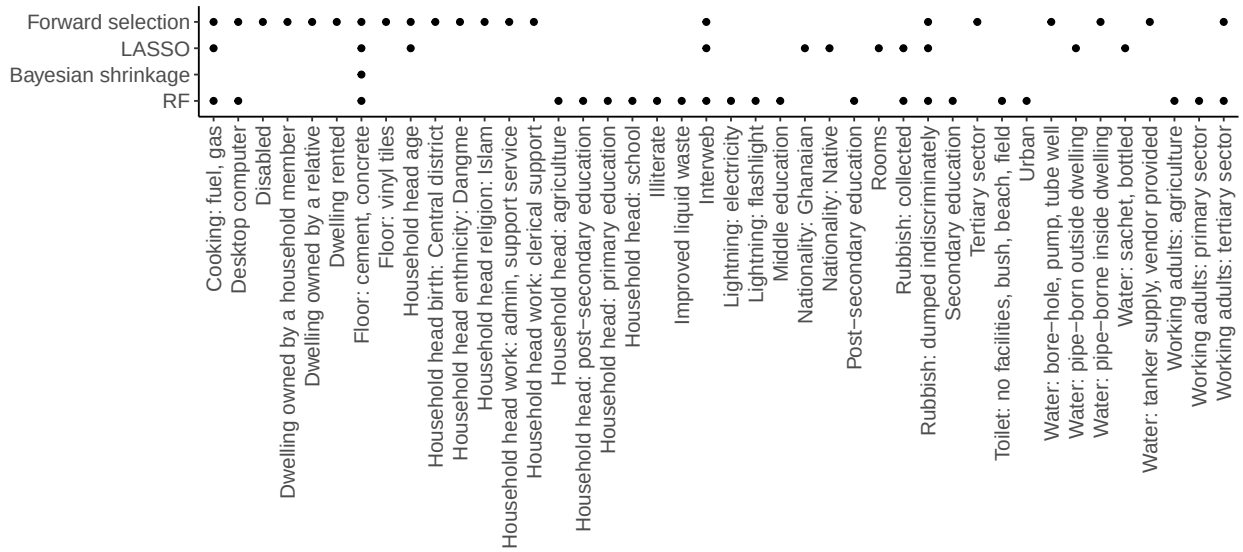


Figure 5.4: Selected covariates for each method when modelling the log equivalised consumption in GAMA.

Figure 5.5 shows the mean log consumption areal estimates and their 95% prediction in-

tervals' widths for each of the four methods. Among the four methods, the random forest approach yields the most homogeneous point estimates across the EAs. This can further be seen in Figure 5.6 which compares the predictions obtained using each method for each EA. The prediction interval widths are shown across the EAs in Figure 5.5 and compared between the modelling methods in pairwise scatter plots in Figure 5.7. The prediction intervals computed for the linear approach with forward variable selection are the narrowest. As expected, the widths of the intervals obtained through the proposed scaled SC approach for the random forest and LASSO predictions behave similarly. The widths are of the form $(1 - f_c) \times 2 \times d_\alpha / \sqrt{N_c - n_c}$, where d_α is the only quantity that differs between the LASSO and random forest approaches. Note that in this analysis, the scaled SC procedure divides the dataset into two halves, consequently computing the necessary residuals and quantile based on only $m/2 = 68$ data points.

Finally, to determine which method performs the best in this particular data application, an 8-fold cross-validation study is conducted. The 136 sampled EAs are divided into 8 rural EAs and 128 urban EAs. Hence, in this 8-fold cross-validation study, 17 EAs are removed from the sample at a time (1 rural and 16 urban EAs), the four methods are fitted on the remaining 119 EAs and predictions are obtained for the 17 removed ones. The four methods are compared in terms of mean absolute bias, MSE, coverages and proper scores of the 50%, 80% and 95% prediction intervals in Table 5.1. These performance measures are computed as described in Section 5.3.2, with the number of replicates R corresponding to the 8 folds and the total number of areas M becoming the number of sampled EAs. The Bayesian shrinkage approach performs the best among the four methods we consider, yielding the smallest bias, MSE and interval scores and reaching the right coverage rates of the prediction intervals. On the other hand, the prediction intervals obtained for the forward selection approach lead to significant undercoverage. A cross-validation study was also conducted where the four methods were fitted without forcing the inclusion of the urban indicator. The results are not shown in this paper as the performance of the four methods in terms of bias, MSE, coverage

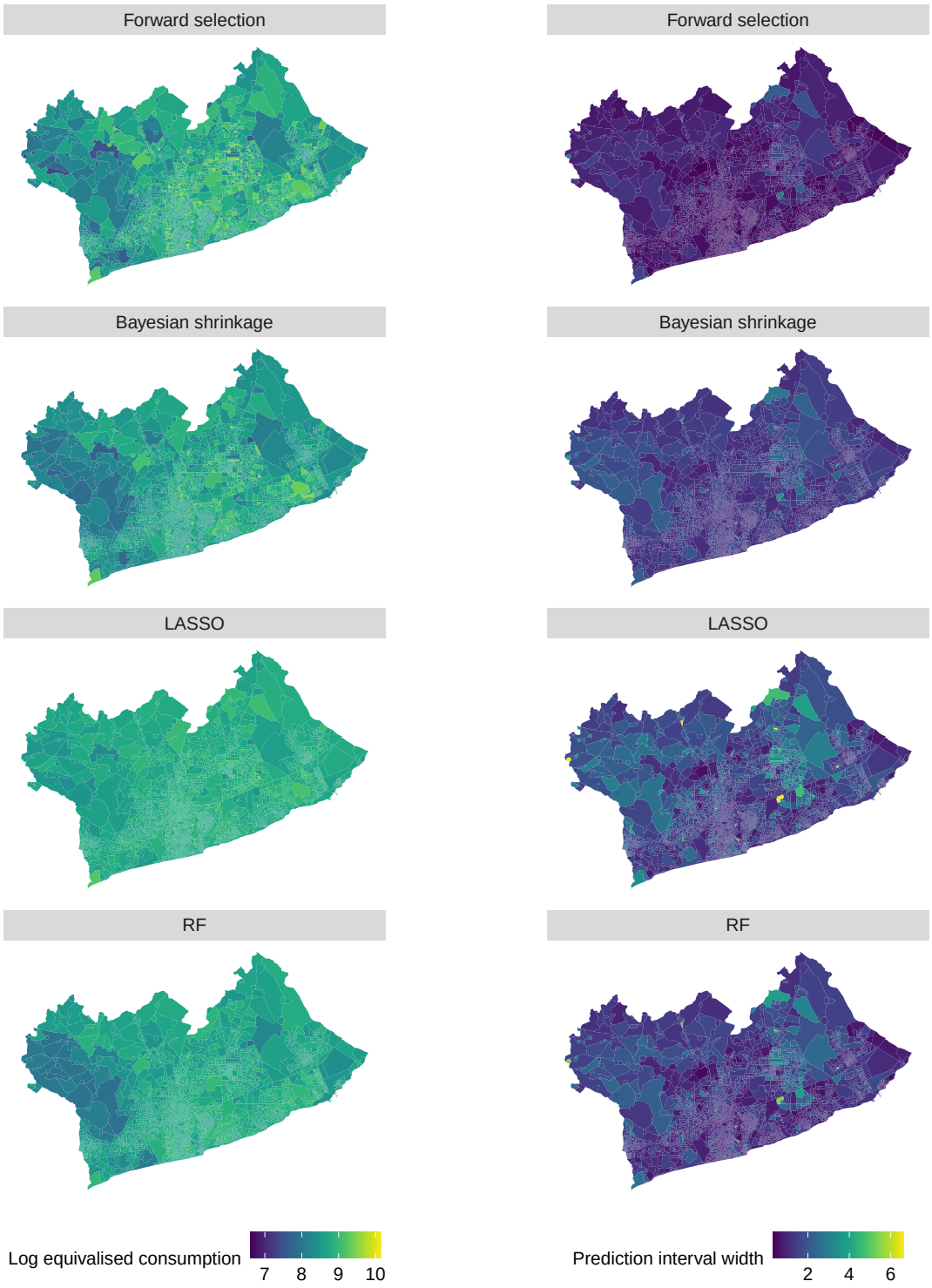


Figure 5.5: Estimated mean log equivalised consumption in the GAMA EAs (Left) and widths of the corresponding 95% prediction intervals (Right) obtained from each modelling method. RF: Random forest.

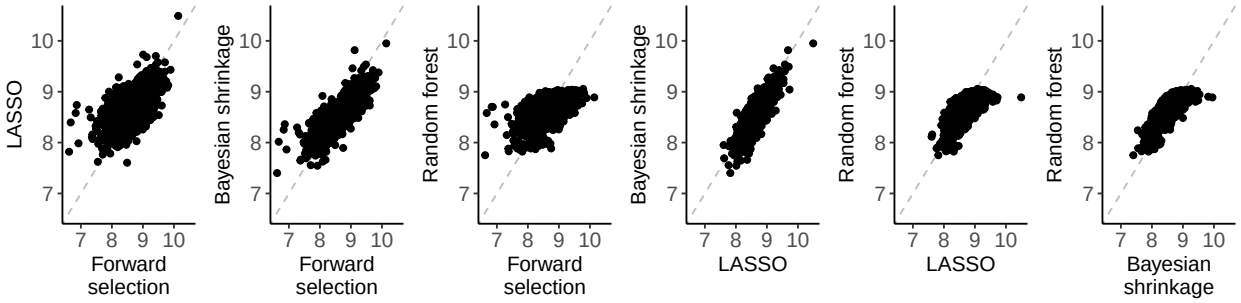


Figure 5.6: Pairwise comparison of the areal estimates obtained from each of the four methods: forward selection, LASSO, Bayesian shrinkage and random forest.

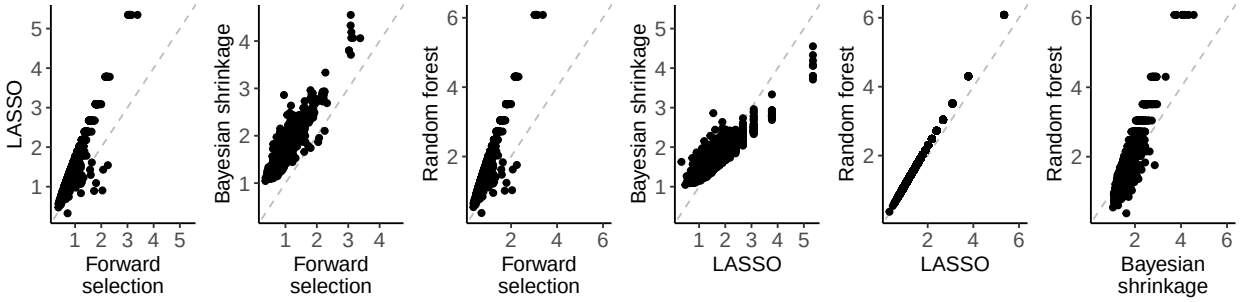


Figure 5.7: Pairwise comparison of the areal prediction interval widths obtained from each of the four methods: forward selection, LASSO, Bayesian shrinkage and random forest.

and proper score of the prediction intervals were similar to the ones shown in Table 5.1, obtained including the urban indicator, for each modelling method. The Bayesian shrinkage method considered in this paper consists in applying the horseshoe prior to the regression coefficients. Other priors could have been considered, such as a Bayesian ridge prior. A cross-validation study was conducted with the Bayesian ridge approach for the GAMA sample. Because the results were similar to the ones shown in Table 5.1, obtained with the horseshoe prior, in terms of bias, MSE, coverage and proper score of the prediction intervals, they are not presented in this paper.

In the original scale, on average, we find that the Bayesian shrinkage consumption estimates among the richest 10% are 2.3 times bigger than the ones among the poorest 10%. We also find that the 92% urban EAs are not uniformly distributed across the estimated consumption deciles: there are only 79% urban EAs among the poorest 10%, versus 91% among the richest

	Absolute Bias	MSE	PI Coverage			Proper interval score		
			95%	80%	50%	95%	80%	50%
Bayesian shrinkage	0.244	0.086	94.1	80.9	48.5	1.33	1.91	3.86
Forward selection	0.975	0.168	72.1	52.2	28.7	3.55	5.35	8.03
LASSO	0.965	0.133	91.9	76.5	49.3	1.75	2.48	4.65
Random forest	0.516	0.097	91.9	79.4	50.0	1.61	2.08	4.16

Table 5.1: Mean absolute bias, MSE, coverages and proper scores of the 50%, 80% and 95% prediction intervals, obtained for each method in the 8-fold cross-validation study on the GAMA sample.

10%. Following Dong and Wakefield (2021), to identify the EAs where interventions should be prioritised, we rank the EAs from poorest to richest, based on the Bayesian shrinkage point estimates. In particular, in this Bayesian framework, we obtain each EAs posterior ranking distribution, by ranking the point estimates at each MCMC iteration. Figure 5.8 shows the posterior ranking distributions for 5 of the 10% poorest EAs and 5 of the 10% richest EAs. Additionally, the right-hand side of Figure 5.8 maps the 10% poorest and richest EAs. We find that the Greater Accra South district, which corresponds to the western EAs in Figure 5.8, gathers most of the poorest EAs, while the Accra Metropolitan Area district, which corresponds to the southern EAs in Figure 5.8, is the richest. Figure 5.8 further shows that the 500 poorest EAs' ranking distributions overlap, which seems to indicate that there is a need to intervene in the poorest 500 EAs.

5.5 Discussion

In this paper, approaches based on random forests and the LASSO are compared with a frequentist forward variable selection procedure and a Bayesian shrinkage method to estimate area-level means of a variable of interest when abundant auxiliary variables are available. Throughout, the areas correspond to the sampling clusters. The methods are area-level model-based small area prediction procedures used to obtain areal estimates and their uncertainties. First, a random forest approach models the outcome values. By construction, auxiliary variables are selected when partitioning the response values through covariate splits.

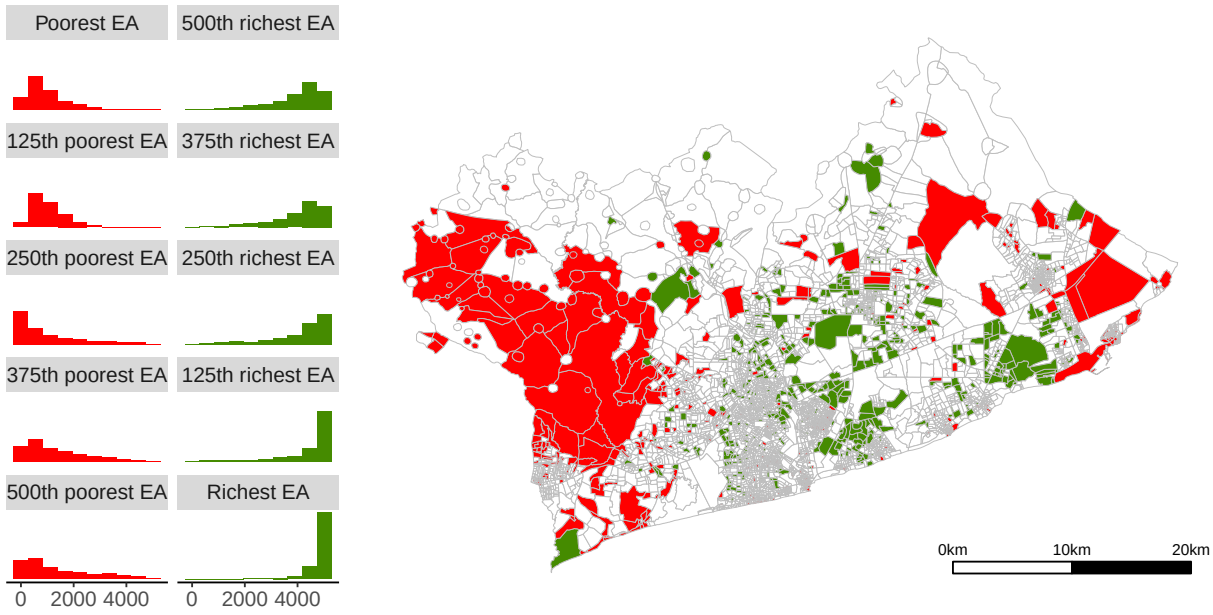


Figure 5.8: Left: Histograms of the posterior ranking distributions of 5 of the 10% poorest EA's (left column, red) and 5 of the 10% richest EA's (right column, green), as estimated from the MCMC samples obtained for the Bayesian shrinkage approach. Right: Map of the Greater Accra Metropolitan Area highlighting the 500 poorest EA's (red) and the 500 richest EA's (green). There are a total of 5019 EAs in the study region.

Then, in the frequentist framework, a LASSO method selects covariates by shrinking irrelevant regression coefficients towards 0. To measure the uncertainty of estimates obtained from random forests and the LASSO methods, a modification of the split conformal (SC) procedure is proposed. The SC algorithm (Lei et al., 2018) estimates prediction intervals with no specific distribution assumption for the data. However, the data are assumed to be exchangeable. The proposed scaled SC procedure relaxes the assumption that the data are exchangeable. Specifically, the proposed algorithm allows the data points to have variances of different scales. This proposed scaled SC procedure allows inference to be conducted for the random forest and the LASSO estimates.

A first simulation study assesses the performance of the proposed scaled SC method compared to the original SC procedure. It is found that when the data points are exchangeable, both procedures perform similarly, regardless of the modelling method. In the simulation scenarios where the number of sampled units is not equal to the number of non-sampled

units ($n_c \neq N_c - n_c$), the variances are scaled differently, $\sigma^2/n_c \neq \sigma^2/(N_c - n_c)$. Hence, the SC procedure does not yield the appropriate coverage rates for the prediction intervals in these scenarios. The proposed scaled SC method corrects the under-coverage in all the simulation scenarios that were considered.

The random forest and LASSO approaches are compared with the frequentist forward selection and Bayesian shrinkage methods in an additional simulation study. When data are generated from a linear model, the methods that assume normality yield smaller biases and MSEs than the random forest approach. All modelling methods, however, lead to adequate prediction interval coverages. The random forest method performs better in terms of MSE when the data are generated from a non-linear model. All methods yield under-coverage when few units are selected within the sampled areas in this complex population.

In the sixth Ghana Living Standards Survey, from 2012–2013, the log equivalised consumption is measured at the household level in a small fraction of the areas (EAs) within the Greater Accra Metropolitan Area (GAMA), alongside 174 auxiliary variables. The same auxiliary information is recorded for all the GAMA EAs in the 2010 Ghanaian census. Using both datasets and the four EA-level method-based approaches, areal estimates of the mean log equivalised consumption are computed for all EAs in the GAMA. Additionally, prediction intervals are computed for all EA estimates to measure their uncertainties. The LASSO and forward variable selection methods select more than 10% of the auxiliary variables, while the Bayesian horseshoe model yields posterior credible intervals that do not include 0 for only one coefficient. The random forest procedure estimates a smoother map of the mean log consumption than the other three approaches. A cross-validation study conducted on the sample data shows that the Bayesian shrinkage method performs the best, among the four methods considered, on this particular dataset.

Finally, in this paper, before fitting random forests to the different datasets, cross-validation studies were run to help set the hyperparameters. These hyperparameters are the number

of regression trees included in the forest, the number of variables considered at each step when growing the trees, and the final node sizes. This step remains to be improved: as other hyperparameters could have led to better performing random forests. For further discussion on the selection of random forest hyperparameters; see e.g., McConville and Toth (2019); Dagdoug et al. (2023). On the other hand, the proposed scaled SC procedure used to compute prediction intervals for the random forest and LASSO estimates relies on an equal split of the data points to grow a forest and compute prediction errors. In the data application of this paper, this implies that the prediction interval limits are based on 68 data points. This partition, suggested by Lei et al. (2018) for the original SC algorithm, could be revisited to attempt to narrow down the resulting intervals.

Appendices

Supplementary material is available in Appendix C:

- C.1: Proof of the proposed scaled split conformal prediction interval coverage
- C.2: Forward approach
- C.3: Random Forest algorithm
- C.4: `R` code: proposed scaled split conformal procedure
- C.5: Design-based simulation study: scaled split conformal procedure
- C.6: Design-based simulation study: prediction methods comparison
- C.7: Extra design-based simulation study: prediction methods comparison
- C.8: Model-based simulation study using the Ghanaian data
- C.9: Detailed results for the model-based simulation study summarised in Section 5.3.2

Chapter 6

Conclusion

The three manuscripts of this thesis developed methods to analyse areal data under three different settings: purely spatial disease mapping, spatio-temporal disease mapping, and SAE. This thesis has contributed to the literature on disease mapping, small area estimation, and machine learning.

Chapter 3 introduced a novel disease mapping model to identify potentially outlying areas with respect to the disease risk, after accounting for the effect of covariates. Two different prior specifications were investigated for the scaling mixture components. I discussed the differences between the proposed heavy-tailed BYM2 model and the prior introduced by Congdon (2017). In particular, I showed how we expect the proposed model to perform better in identifying outliers, compared to Congdon's prior, when the outlying areas are neighbours, while we expect both models to perform similarly when outliers are distant. The simulation studies summarised in this manuscript confirmed this theoretical comparison. Further, the analysis of Zika cases recorded in 2015-2016 across the neighbourhoods of Rio de Janeiro led to the identification of potential outliers. This data analysis showed how the proposed model may be useful to better understand the spread of the disease, identify potential issues in the recording of cases, and prioritise interventions.

In Chapter 4, I proposed a new spatio-temporal disease mapping model that aims to identify potentially outlying areas regarding the evolution of the disease risk, after accounting for covariates. The proposed model is a scaling mixture extension of the spatio-temporal model proposed by Rushworth et al. (2014). Two prior specifications for the scaling components were discussed, where we expect the proposed model to perform better in identifying potential neighbouring outliers when the scaling parameters are spatially structured. This was confirmed through simulation studies. Finally, two data applications showed that the proposed model may be useful, for example in the midst of a pandemic, to prioritise interventions and implement localised policies.

Chapter 5 was motivated by Ghanaian survey data, where the goal was to estimate the average household log consumption at the EA level, using two datasets with many covariates: a survey, where the outcome was measured for 3% of all EAs in the GAMA, and a census, where the outcome was not measured. Hence, this manuscript focused on SAE when few areas are sampled and abundant auxiliary information is available. In this manuscript, I compared four area-level model-based approaches in the frequentist, Bayesian and machine learning frameworks. Further, I proposed a new procedure to measure the uncertainty of complex point estimates, such as ones computed through the LASSO and random forest methods. The proposed procedure is an extension of the SC method proposed by Lei et al. (2018), where the assumption of exchangeable data is relaxed as it is a strong assumption in the context of SAE. I proved that the proposed procedure yields prediction intervals of the right coverage. This result was confirmed through simulation studies. Further, the four modelling approaches were compared through simulation studies that aimed to mimic SAE data, and the household log consumption was estimated for all EAs in the GAMA.

6.1 Avenues for future research

This thesis provides advancements of various methods to analyse areal data. There are interesting extensions that can be explored in the future.

For instance, the spatio-temporal disease mapping model proposed in Chapter 4 extends that of Rushworth et al. (2014) to accommodate and identify outlying areas. The prior introduced by Rushworth et al. (2014) is itself an extension of Leroux et al. (1999) to the temporal framework. However, Riebler et al. (2016) discuss how latent effects that follow the Leroux model cannot be scaled such that the model parameters lie in the marginal distribution and do not depend on the spatial structure under study. Therefore, the parameters in the spatio-temporal model proposed in Chapter 4 lie in the *conditional* distribution and do depend on the spatial structure under study, which means that interpretation should be done with care. Hence, it would be interesting to extend the model proposed in Chapter 3 to the spatio-temporal framework. One may think of two directions: first, to allow the unstructured and spatially structured effects to evolve through time, or, second, to allow the variance and mixing parameters to vary with time. This second case would be along the lines of what was proposed by Nobre et al. (2005); Napier et al. (2016). In both cases, however, there might be identifiability issues that should be investigated. Another potential research avenue regarding the proposal in the second manuscript, is to allow the scaling parameters to vary across space and time, similar to what Fonseca et al. (2023) proposed in a dynamic linear model setting. Therefore, in this case, an area would be identified as a potential outlier for a particular time point.

The work conducted in Chapter 5 was motivated by the available Ghanaian data, where the EAs sampled in the GLSS are anonymised. Therefore, we did not know which of the EAs in the GAMA were the observed 3%. With this information, one might extend the different modelling approaches considered to include random effects, and, in particular, spatially structured random effects. In the Bayesian framework, this would be similar to what was

done by Wakefield et al. (2020). Then, it would be interesting to compare this Bayesian approach to a machine learning method with spatially structured latent effects. For instance, Krennmair and Schmid (2022) propose a random forest method for SAE data that allows for random effects, and one may borrow ideas to include spatially structured random effects. Further, Krennmair and Schmid (2022) propose a bootstrap procedure to measure the uncertainty of their areal estimates of non-exchangeable data. It would be interesting to investigate how the SC procedure might be altered to allow for areal random effects. Note that although Dunn et al. (2018) propose various conformal prediction procedures for models with random effects, the authors assume the random effects to be independent, which would not be the case were they spatially structured. Further investigation is needed in this setting of random forests with spatially structured latent effects.

Appendices

APPENDIX A

Appendix to Manuscript 1

A.1 Stan code for the proposed model

The stan code used to fit the proposed BYM2-Gamma model in the simulation studies (section 3.3.1, Appendices A.3, A.4, A.5, A.6 and A.7) and in the analysis of the Zika epidemic in Rio de Janeiro (section 3.3.2) is presented below.

Listing A.1: Stan code for the BYM2-Gamma proposed model

```
data {
  int<lower=1> N; //Number or areas
  int<lower=1> N_edges; // Total number of neighbours in the region
  int<lower=1> p; // General case where there are p covariates, excluding the intercept
  matrix[N,p] X;
  int<lower=1, upper=N> node1[N_edges]; // vectors of neighbourhood
  int<lower=1, upper=N> node2[N_edges]; // structure
  int<lower=0> y[N]; // Zika counts
  vector<lower=0>[N] log_E; // offset
  real<lower=0> scaling_factor; // to scale the variance of the latent effects
```

```
11 }
12
13 parameters {
14   real beta0; // intercept
15   vector[p] beta; // Fixed effects
16   real<lower=0> sigma; // marginal standard deviation
17   real<lower=0, upper=1> lambda; // mixing parameter
18   vector[N] theta; // unstructured components
19   vector[N] s; // spatially structured components
20   vector<lower=0>[N] kappa; // outlier indicator
21   real<lower=0> nu; // parameter included in the prior for each kappa_i
22 }
23
24 transformed parameters {
25   vector[N] convolved_re; // complete latent effect
26   for(i in 1:N){ convolved_re[i] = sqrt(1 - lambda) * theta[i] +
         sqrt(lambda/scaling_factor) * s[i]; }
27 }
28
29 model {
30   for(i in 1:N)
31    y[i] ~ poisson_log(log_E[i] + beta0 + X[i,]*beta + convolved_re[i] *
         (sigma/sqrt(kappa[i])) );
32
33   target += -0.5 * dot_self(s[node1] - s[node2]); // Prior for the spatially structured
       components
34   sum(s) ~ normal(0, 0.001 * N); // Soft sum-to-zero constraint to be able to have an
```

```
        intercept

35
36    for(j in 1:p){
37      beta[j] ~ normal(0.0, 10.0);
38    }
39
40    beta0 ~ normal(0.0, 10.0);
41    theta ~ normal(0.0, 1.0);
42    sigma ~ normal(0.0, 1.0);
43    lambda ~ uniform(0.0, 1.0);
44    kappa ~ gamma(nu/2.0,nu/2.0);
45    nu ~ exponential(1.0/4.0);
46  }
47
48  generated quantities {
49    vector[N] mu_log;
50    vector[N] lik;
51    for(i in 1:N){
52      mu_log[i]=log_E[i] + beta0 + X[i,]*beta + sigma*convolved_re[i]/sqrt(kappa[i]);
53      lik[i] = exp(poisson_log_lpmf(y[i] | mu_log[i])); // likelihood to compute the WAIC
54    }
55  }
```

A.2 Convergence diagnostics for the proposed model

In this section, we present the trace plots, effective sample sizes and $\widehat{R}$ statistics for a few selected parameters of the two parametrisations of the proposed model, when fitted to the

data application in Section 3.3.2. For the mixture components, κ's, we select the ones that produced the best and the worst convergence diagnostics.

	BYM-Gamma		BYM2-logCAR	
	ESS	$\widehat{R}$	ESS	$\widehat{R}$
κ_{92}	1305	1.000	1614	0.999
κ_{13}	2000	0.999	2838	0.999
λ	1958	1.009	1211	1.007
ν	2000	1.000	1817	1.000
σ	1912	1.001	1987	1.004

Table A.1: Effective sample sizes (ESS) and $\widehat{R}$ statistics for some parameters when fitting the two parametrisations of the proposed model to the Zika data. κ_{13} and κ_{92} were chosen because they produced the best and the worst convergence diagnostics.

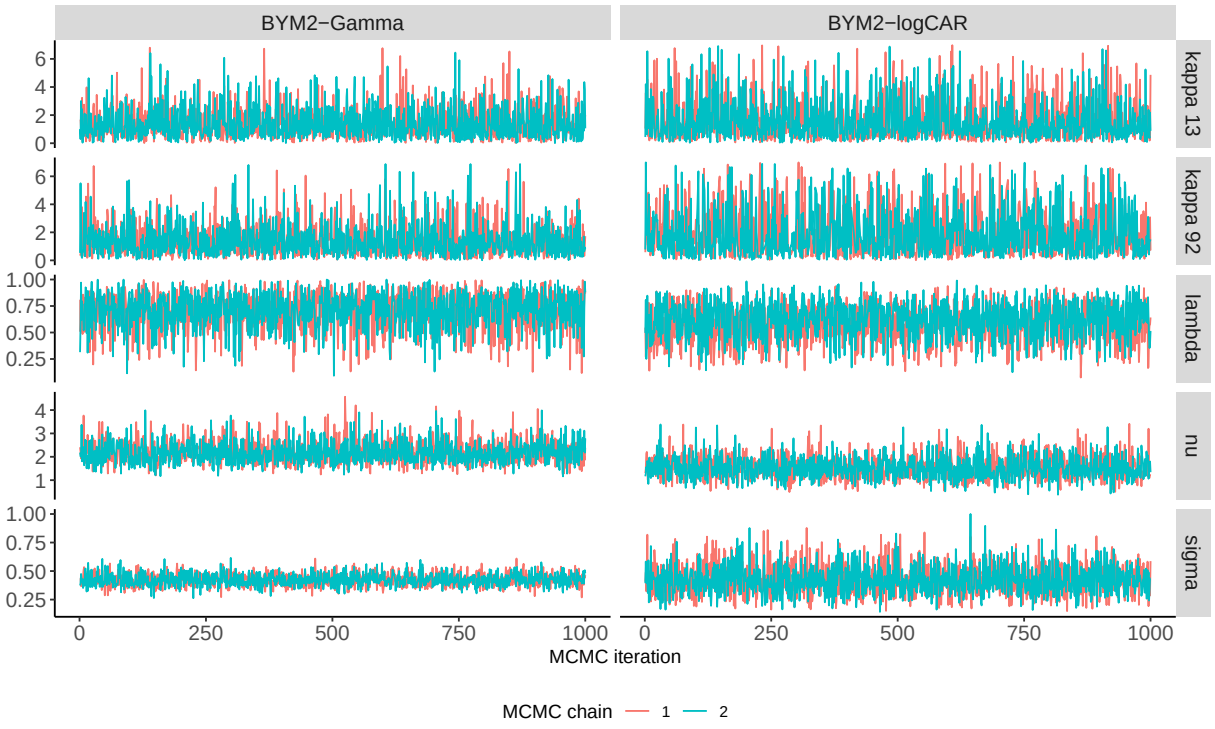


Figure A.1: Trace plots for some parameters when fitting the two parametrisations of the proposed model to the Zika data. κ_{13} and κ_{92} were chosen because they produced the best and the worst convergence diagnostics.

A.3 Simulation study: generating data from the proposed BYM2-Gamma model

To assess the proposed BYM2-Gamma model's ability to recover the truth, a simulation study is conducted wherein data are generated from the proposed BYM2-Gamma model for 100 replicates. The $n = 160$ districts of Rio de Janeiro and their neighbourhood structure are used. The latent effects' unstructured and scaled spatially structured components are generated as follows:

$$\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad \text{and} \quad \boldsymbol{u}^{\star} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_{\star}^{-}),$$

where $\boldsymbol{Q}_{\star} = h(\boldsymbol{D} - \boldsymbol{W})$, with h, the scaling factor, entirely defined by the spatial structure of Rio de Janeiro. An algorithm to generate from the ICAR prior is presented in Chapter 2 of Rue and Held (2005). The mixing components that induce the marginal heavier tails, $\boldsymbol{\kappa}$, are independently generated from a Gamma$(\nu/2, \nu/2)$, with ν fixed at 4 to allow for fairly heavy tails. The latent effects are then computed as

$$b_i = \left[\sqrt{1-\lambda}\boldsymbol{\theta}_i + \sqrt{\lambda}u_i^{\star}\right] \times \sigma/\sqrt{\kappa_i}, \ i = 1, \ldots, n,$$

where $\lambda = 0.8$ and $\sigma = 0.3$. Finally, a population of size $n = 160$ is generated from the Poisson model

$$Y_i \sim \mathcal{P}\left(E_i \exp\left[\beta_0 + b_i\right]\right),$$

with $\beta_0 = -0.1$ and the offsets, $[E_1, \ldots, E_n]^{\top}$, taken from the analysis of the Zika counts. Then, models BYM2-Gamma and Congdon are fitted to each of the 100 replicates, using the same inference procedure as in section 3.3.1. The goal is to check if we recover the true values used to generate the data, and to check if the WAIC is able to distinguish between the proposed model and Congdon's.

Figure A.2 shows that the WAIC is able to always choose the model that generated the data,

namely the BYM2-Gamma model. Figure A.3 presents the posterior summaries obtained from the BYM2-Gamma model across the 100 replicates for the intercept, β_0, the mixing parameter, λ, the hyperparameter, ν, and the overall standard deviation, σ. For all samples, the 95% posterior credible intervals of all parameters contain the true values used to generate the data. The interest lies particularly on the main parameters of the model, such as the outlier indicators, $\boldsymbol{\kappa}$. Figure A.4 plots the posterior summaries, for one replicate, of the κ's across all districts in Rio de Janeiro. Most of the 95% posterior credible intervals for κ contain the true value used to generate the data. Moreover, for those neighbourhoods that have outlying observations, the estimate for κ is quite concentrated around its true value. This suggests that the model is able to point out the neighbourhoods with outlying observations. Similarly, the true latent effects, $\boldsymbol{b}$, are shown to be recovered by the 95% posterior credible intervals in Figure A.5.

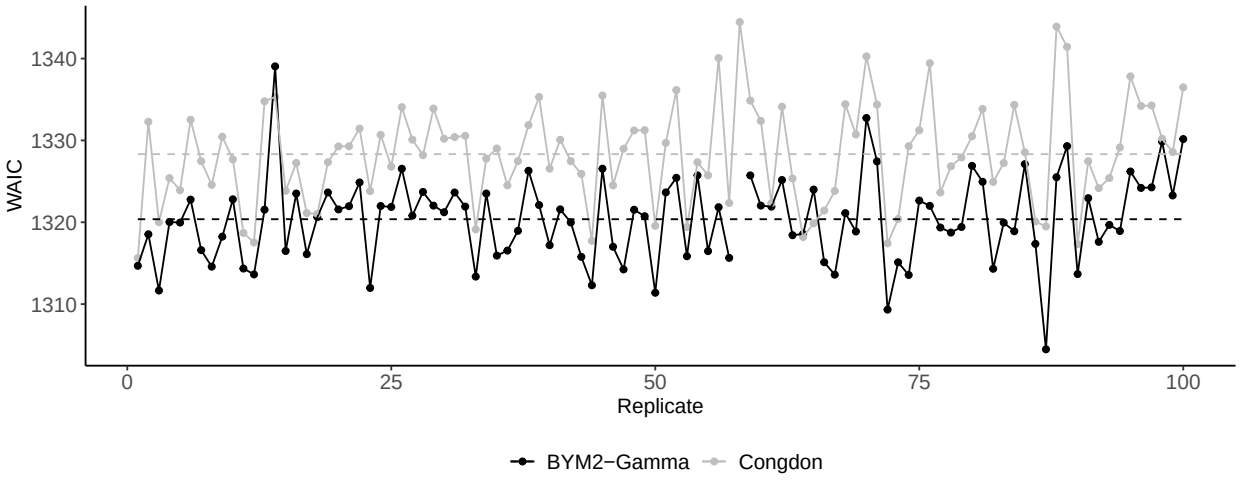


Figure A.2: WAIC across the 100 replicates for the proposed BYM2-Gamma model and Congdon's regarding the simulated data from the proposed BYM2-Gamma model. Dashed lines: mean WAIC for each model

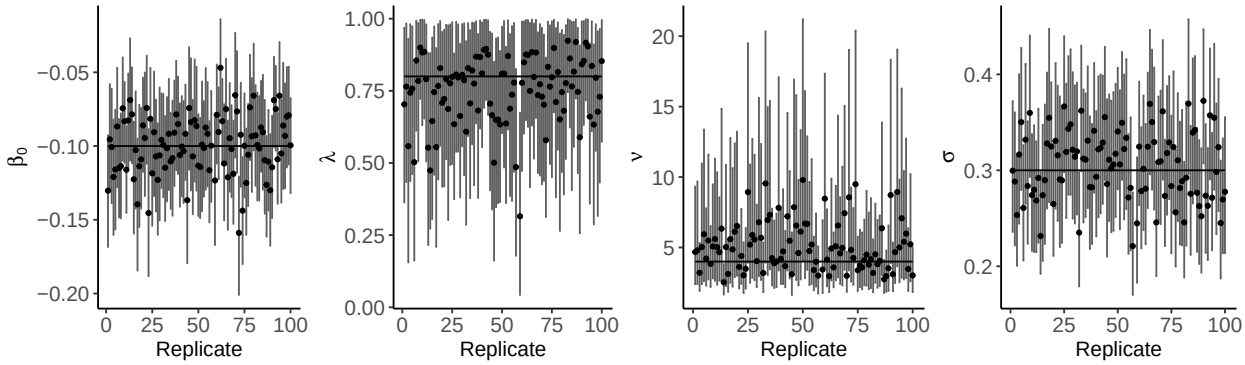


Figure A.3: Posterior summaries of the parameters for the proposed BYM2-Gamma model across the 100 replicates regarding the simulated data from the proposed BYM2-Gamma model.
Solid circle: posterior mean; Vertical lines: 95% posterior credible interval; Solid horizontal line: true value.

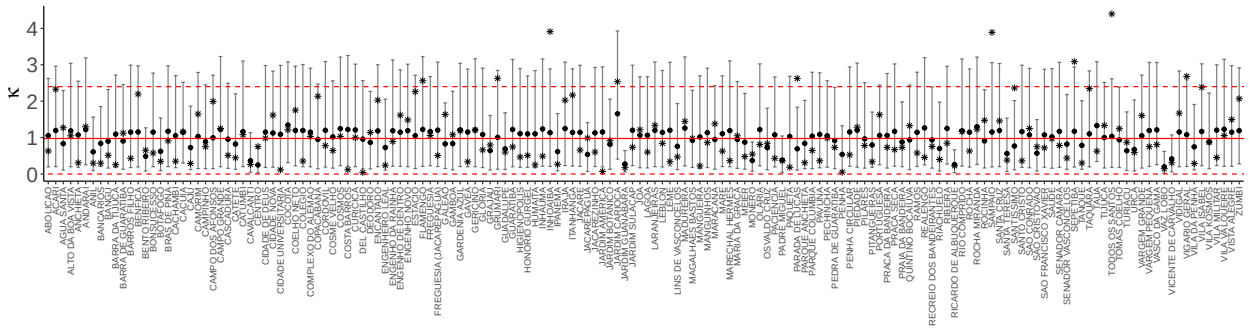


Figure A.4: Posterior summaries (mean and 95% credible interval) of the κ parameters across all the districts of Rio de Janeiro for one replicate when fitting the BYM2-Gamma model. The stars correspond to the true generated κ's and the red horizontal lines correspond to the prior summary (solid line: prior mean, dashed lines: prior 95% credible interval).

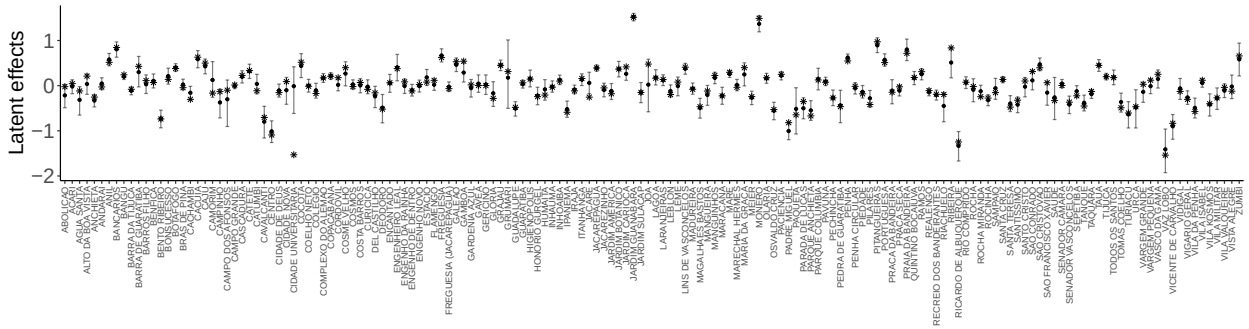


Figure A.5: Posterior summaries (mean and 95% credible interval) of the latent effects across all the districts of Rio de Janeiro for one replicate when fitting the BYM2-Gamma model. The stars correspond to the true generated latent effects.

A.4 Simulation study: generating data from the proposed BYM2-logCAR model

We now assess the proposed BYM2-logCAR model's ability to recover the truth. Similar to Appendix A.3, a simulation study is conducted wherein 100 replicated datasets are generated from the proposed BYM2-logCAR model using the $n = 160$ districts of Rio de Janeiro. The unstructured and spatially structured components, $\boldsymbol{\theta}$ and $\boldsymbol{u}^\star$ respectively, are independently generated, like in Appendix A.3. The scaling mixture components, $\boldsymbol{\kappa}$, are generated using the spatial structure as follows:

$$\boldsymbol{z} \mid \nu_\kappa \sim \mathcal{N}\left(\mathbf{0}, \nu_\kappa \boldsymbol{Q}_{\alpha,\star}^{-1}\right) \quad \text{and} \quad \kappa_i = \exp\left(-\frac{\nu_\kappa}{2} + z_i\right), \; i = 1, \ldots, n,$$

where $\boldsymbol{Q}_{\alpha,\star} = h\boldsymbol{Q}_\alpha = h_\alpha[\boldsymbol{D} - \alpha\boldsymbol{W}]$ is again the valid precision matrix that is scaled by h_α, which is computed based on $\boldsymbol{D} - \alpha\boldsymbol{W}$. We impose $\alpha = 0.99$ and define an arbitrary $\nu_\kappa = 0.3$ to allow the κ's to depart from 1. Like in Appendix A.3, the latent effects are then computed as $b_i = \left[\sqrt{1-\lambda}\boldsymbol{\theta}_i + \sqrt{\lambda}u_i^\star\right] \times \sigma/\sqrt{\kappa_i}$, $i = 1, \ldots, n$, where $\lambda = 0.8$ and $\sigma = 0.3$. Finally, the population of size $n = 160$ is generated from the Poisson model, $Y_i \sim \mathcal{P}\left(E_i \exp\left[\beta_0 + b_i\right]\right)$, with $\beta_0 = -0.1$ and the offsets, $[E_1, \ldots, E_n]^\top$, taken from the analysis of the Zika counts. The proposed BYM2-logCAR model and Congdon's are both fitted on the 100 replicated datasets using the same inference procedure as in section 3.3.1.

Figure A.6 shows that the WAIC always favours the proposed BYM2-logCAR model, which generated the data. Figure A.7 shows how well the proposed BYM2-logCAR model is able to recover the true values of the model parameters through the posterior summaries across the 100 replicates for the intercept, β_0, the mixing parameter, λ, the hyperparameter, ν_κ, and the overall standard deviation, σ. Across the 100 replicates, the proposed BYM2-logCAR model always captures the truth, as the posterior 95% credible intervals (vertical lines) always cover the true values of the parameters (solid horizontal lines). Regarding the scaling

mixture components, $\boldsymbol{\kappa}$, Figure A.8 shows the posterior summaries for one replicate and generated values, across all districts. The κ's generated following this structured prior seem to vary less than the ones generated from the independent gamma priors in Appendix A.3. Therefore, the posterior credible intervals are narrower than the ones from the simulation study presented in Appendix A.3. Regardless, the posterior 95% credible intervals almost always cover the true mixture components. Similarly, the generated latent effects plotted in Figure A.9 are recovered by the posterior 95% credible intervals.

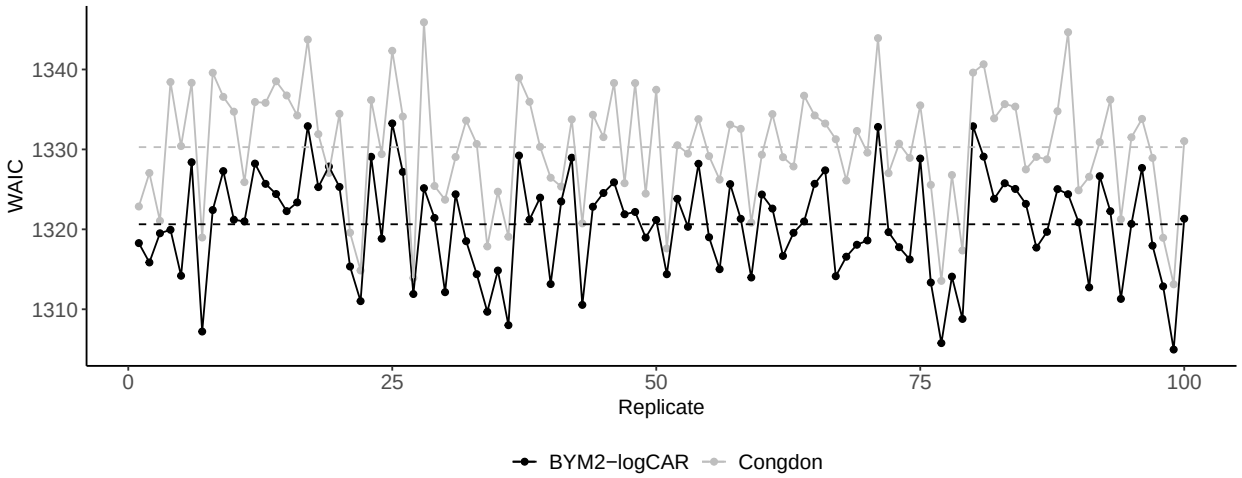


Figure A.6: WAIC across the 100 replicates for the proposed BYM2-logCAR model and Congdon's regarding the simulated data from the proposed BYM2-logCAR model. Dashed lines: mean WAIC for each model

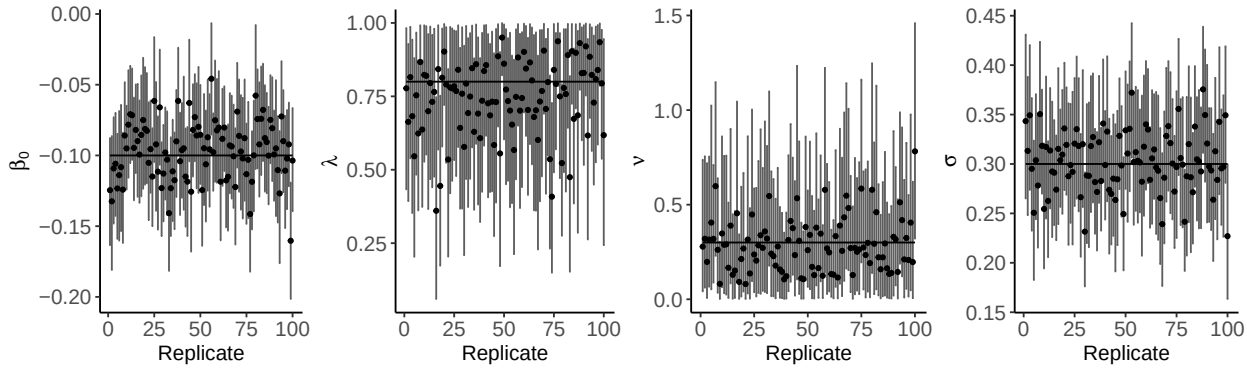


Figure A.7: Posterior summaries of the parameters for the proposed BYM2-logCAR model across the 100 replicates regarding the simulated data from the proposed BYM2-logCAR model.
Solid circle: posterior mean; Vertical lines: 95% posterior credible interval; Solid horizontal line: true value.

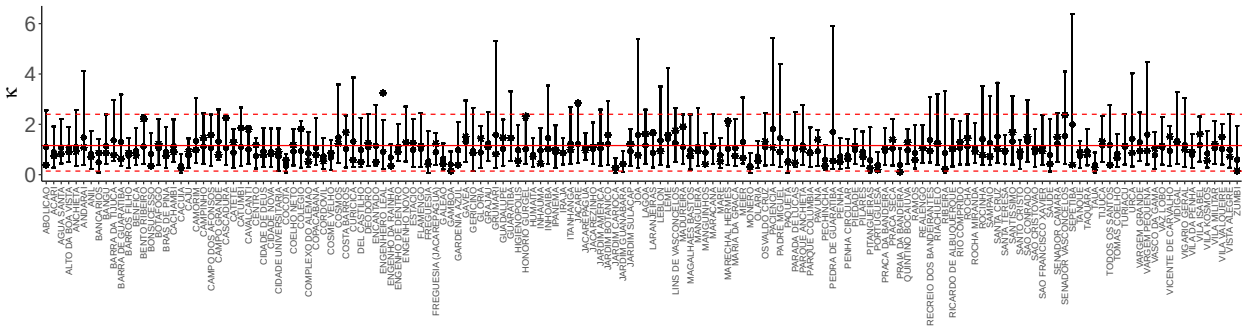


Figure A.8: Posterior summaries (mean and 95% credible interval) of the κ parameters across all the districts of Rio de Janeiro for one replicate when fitting the BYM2-logCAR model. The stars correspond to the true generated κ's and the red horizontal lines correspond to the prior summary (solid line: prior mean, dashed lines: prior 95% credible interval).

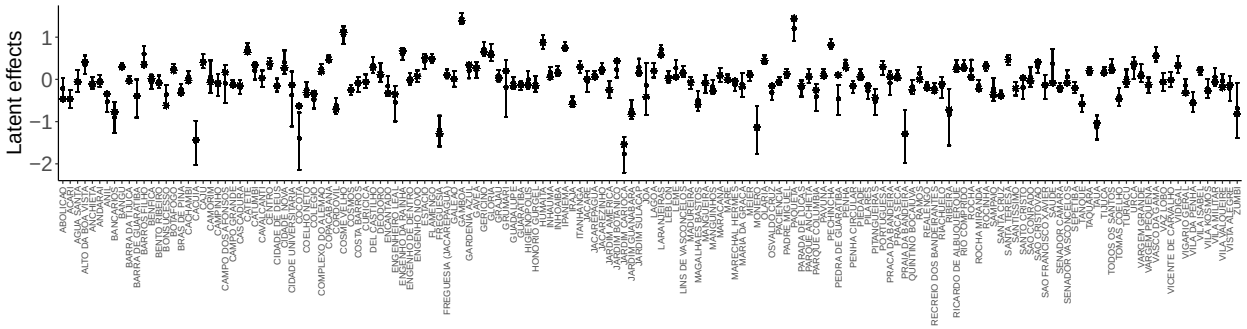


Figure A.9: Posterior summaries (mean and 95% credible interval) of the latent effects across all the districts of Rio de Janeiro for one replicate when fitting the BYM2-logCAR model. The stars correspond to the true generated latent effects.

A.5 Simulation study: no outlying areas

To confirm that the proposed model does not detect outliers when unnecessary, a simulation study is again conducted on the map of Rio de Janeiro without contaminating any district. Data are generated 100 times as follows:

$$Y_i \sim \mathcal{P}\left(E_i \exp[\beta_0 + b_i]\right), \ i = 1, \ldots, n,$$

with $n = 160$, $\beta_0 = -0.1$, $\boldsymbol{E} = [E_1, \ldots, E_n]^\top$ taken from the Zika data analysis. The latent effects, $\boldsymbol{b} = [b_1, \ldots, b_n]^\top$, are simulated once from a PCAR distribution:

$$\boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_b^2 \left[\boldsymbol{D} - \alpha \boldsymbol{W}\right]^{-1}\right),$$

with $\sigma_b = \sqrt{0.2}$ and $\alpha = 0.7$. Figure A.10 shows the map of the 50th replicate of the simulated dataset, where no district seems to be an outlier with respect to the whole city. Again, the two parametrisations of the proposed model are compared to Congdon's, using the same prior distributions as described in section 3.3.1.

In terms of WAIC, the proposed models seem to perform best, as shown in Figure A.11. For this simulation study, the interest lies particularly in comparing the outliers detections from the two versions of the proposed model and Congdon's. Figure A.12 presents the districts that are found to be outliers by the BYM2-Gamma proposed model (a), the BYM2-logCAR proposed model (b) and Congdon's (c). The BYM2-Gamma model only identifies one district, Freguesia, to be a potential outlier in 2% of the replicates. The BYM2-logCAR and Congdon's models on the other hand detect Freguesia up to 8% of the times, showing more sensitivity to the neighbourhood structure. Congdon's model further identifies 5 districts as potential outliers although no district was contaminated.

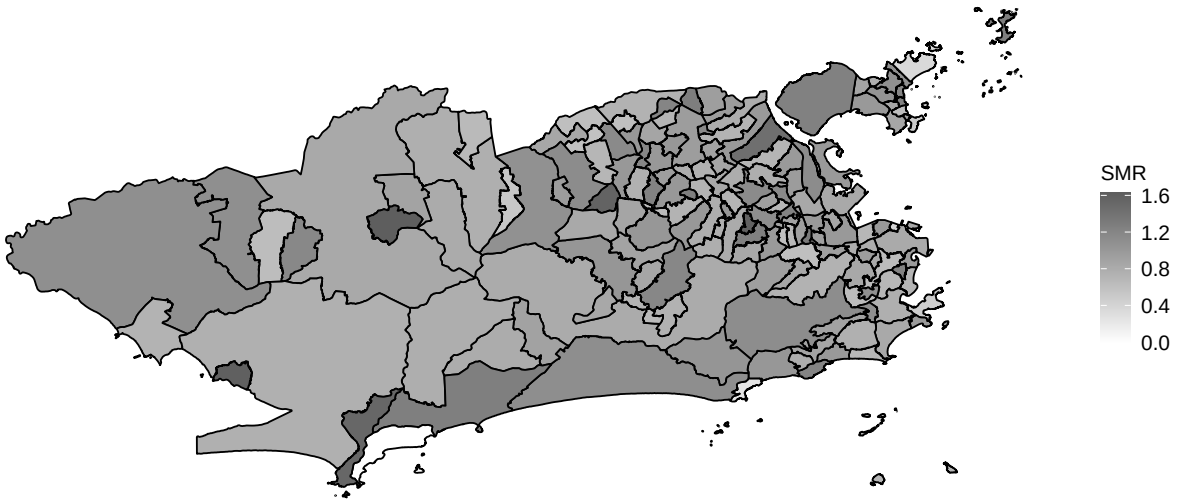


Figure A.10: Standardised morbidity ratio for the 50th simulation without outliers.

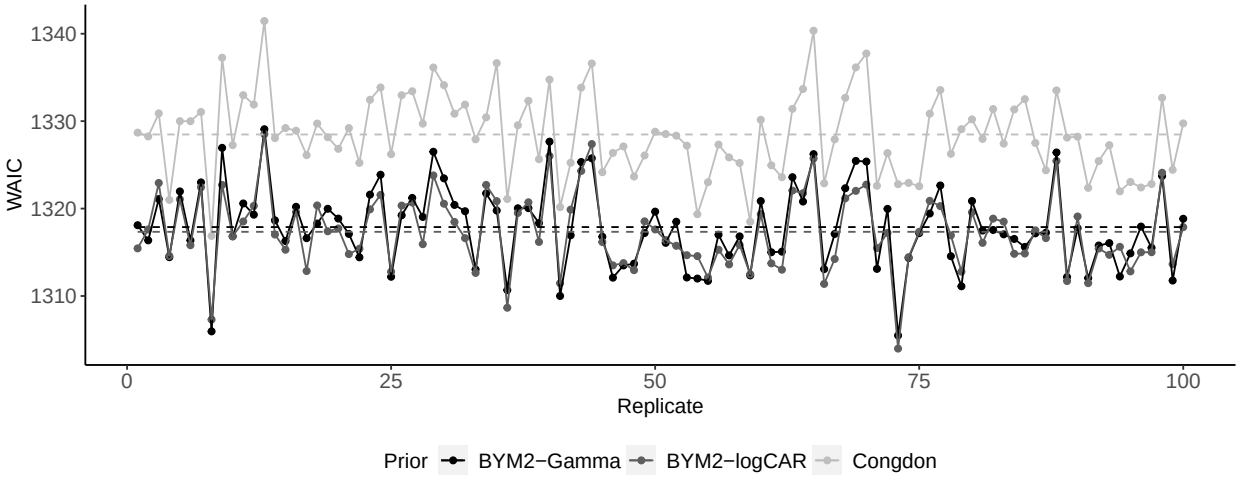
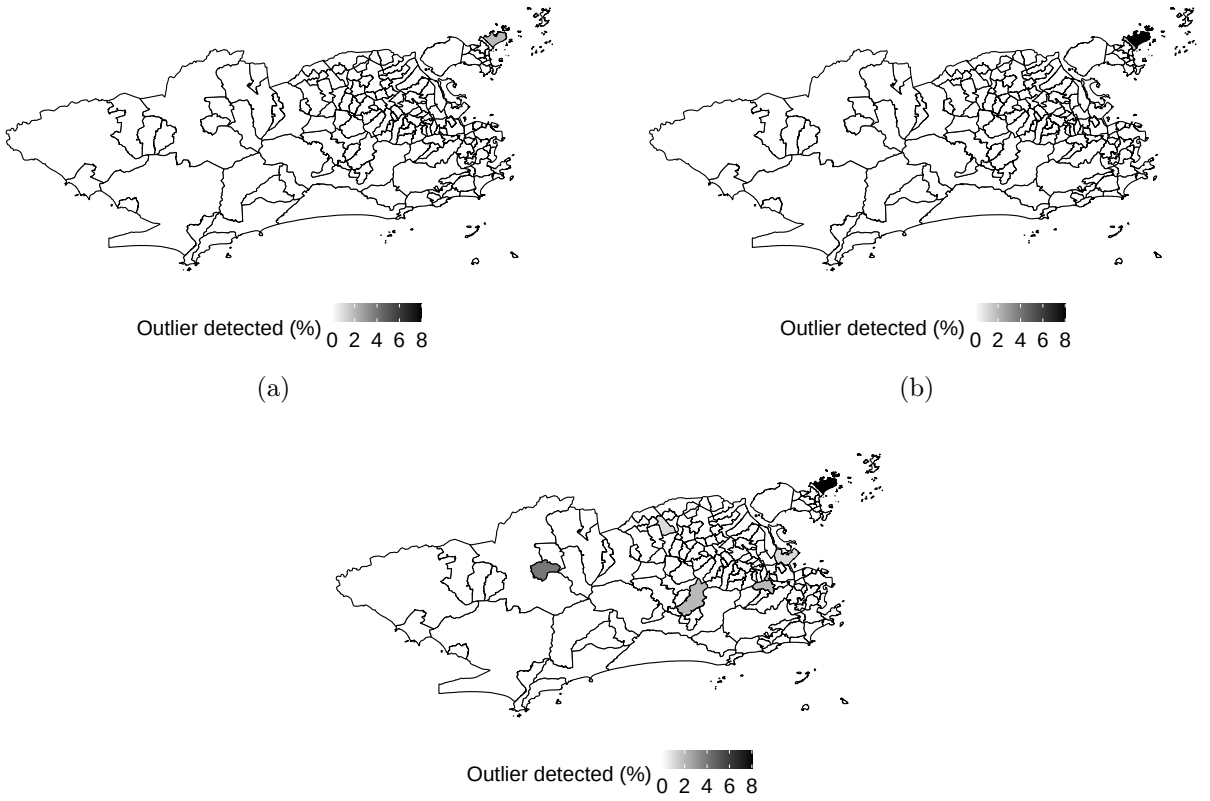


Figure A.11: WAIC across the 100 replicates for the proposed models and Congdon's for the simulation without outliers. Dashed lines: mean WAIC for each model



(a) (b)

(c)

Figure A.12: Maps of the percentages of outliers as indicated by $\kappa_{ur} < 1$ across the $r = 1, \ldots, 100$ replicates, where κ_{ur} is the upper bound of the posterior 95% credible interval of κ in the rth replicate of the simulated dataset without outliers. a) BYM2-Gamma model; b) BYM2-logCAR model; c) Congdon's model.

A.6 Simulation study: distant outliers in France

In this simulation study, 20 distant French departments are contaminated such that outliers are created. Similar to the simulation study presented in Section 3.3.1, there are no covariates in this analysis, and all areas are first imposed a relative risk of 1, $\mu_i = 1$. The same five offset categories are defined. Based on these categories, we select 20 non-neighbouring departments to be outliers. Four departments are chosen from each offset category. That is, there are 4 outliers within the smallest offset group, 4 within the second-to-smallest offset group, and so on. Then, within each group of four departments, the relative risks are contaminated

into outliers by setting the relative risks to be equal to $\mu_{i'} = 0.25$, $\mu_{i''} = 0.5$, $\mu_{i'''} = 1.5$ and $\mu_{i''''} = 2$. The resulting outliers are mapped in Figure A.13, highlighting the offset sizes and imposed relative risks. Again, $R = 100$ populations of size $n = 96$ are created by generating the number of cases $Y_i \sim \mathcal{P}(E_i\mu_i)$. The same four models with priors defined in section 3.3.1 are fitted through `rstan`. After 20,000 iterations with a burn-in period of 10,000 and a thinning factor of 10, the 2 MCMC chains attained convergence as assessed by trace plots, effective sample sizes and $\widehat{R}$ statistics.

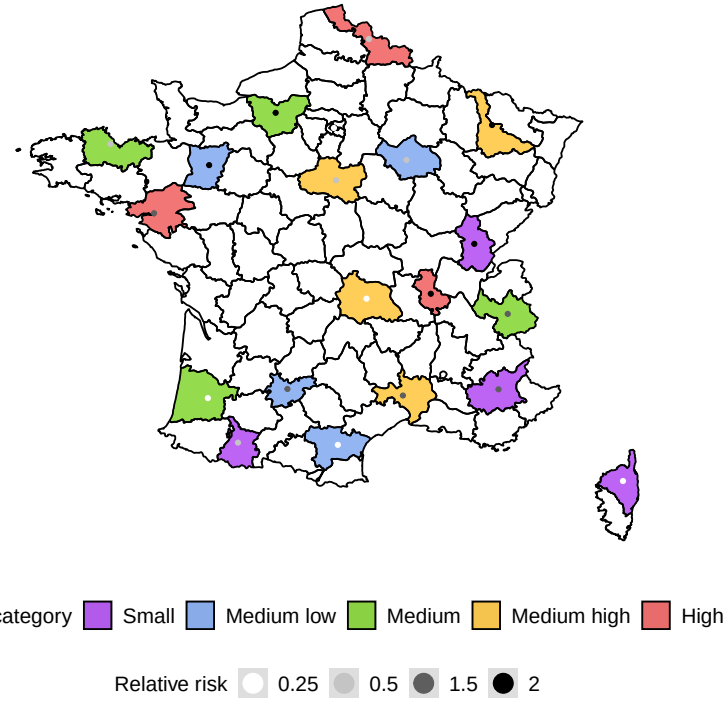


Figure A.13: French departments arbitrary chosen to be outliers in the simulation study with distant outliers. Colours depict the offset category based on the empirical offset quantiles. The points represent the relative risk set to each outlying district.

In terms of WAIC (Watanabe and Opper, 2010), for which smaller values are preferred, the proposed BYM2-Gamma model performs similarly to Congdon's, as shown in Figure A.14. The BYM2-Gamma and original Congdon models always perform better than the models that include spatially structured scaling mixture components. On average, the BYM2-logCAR and Congdon-logCAR models yield a criterion of 983, while the BYM2-Gamma and

Congdon models present a WAIC of 958 and 959, respectively.

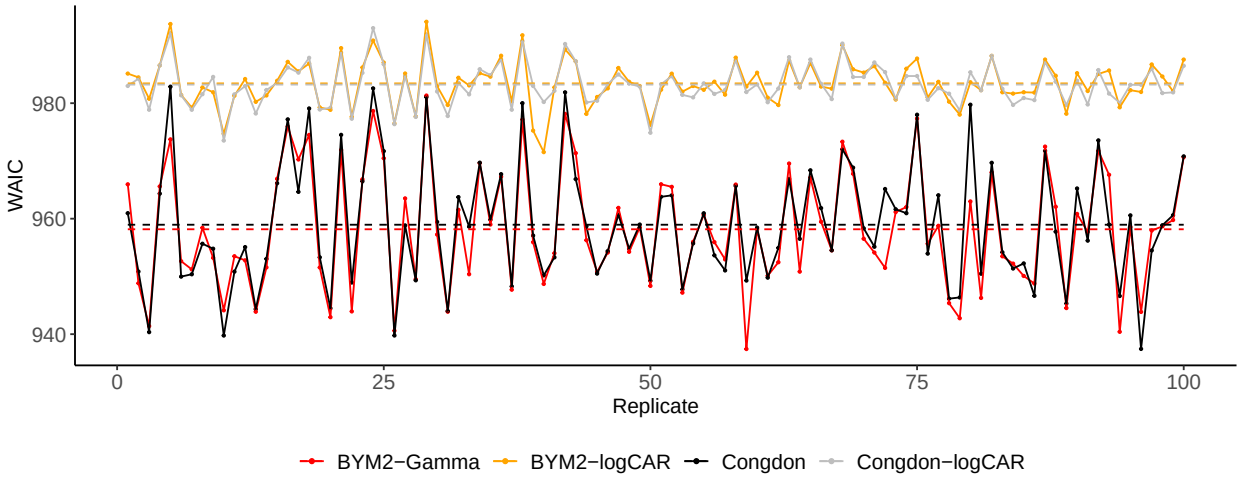


Figure A.14: WAIC across the 100 replicates for the proposed models and Congdon's, in the simulation study with distant outliers. Dashed lines: mean WAIC for each model.

The models' performances are also compared in terms of MSE, as shown in Figure A.15. As expected, all models result in MSEs that are smaller in the areas with large offsets, and MSEs that are larger in the areas with small offsets. Additionally, all models tend to better fit the data in non-outlying areas, that is in the areas with a relative risk of 1. Regarding the outlying areas only, the largest MSEs are observed for extreme risks of 2 whereas the smallest correspond to extreme risks of 0.5. On average over the 100 replicated datasets and across all areas, the MSEs are of 0.0010 for the BYM2-Gamma and Congdon models, and 0.0011 for both log-CAR parametrisations.

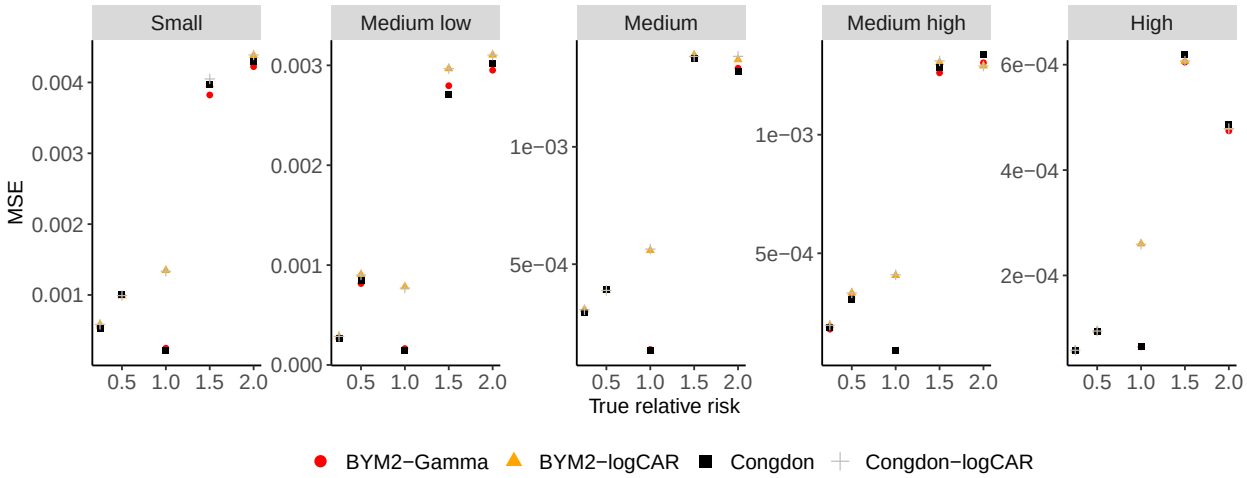


Figure A.15: MSE over the 100 replicates for the proposed models and Congdon's according to the true relative risk and the offset size, in the simulation study with distant outliers.

Regarding the detection of outliers, which is the main focus of this simulation study, Table A.2 shows how often each model accurately detects districts as outliers (sensitivity) and non-outliers (specificity), depending on the offset category. Additionally, Figure A.16 shows how often each district is detected as a potential outlier by the four models, while indicating the offset sizes. Recall, area i is detected as an outlier when $\kappa_{u,i} < 1$, where $\kappa_{u,i}$ is the upper bound of the 95% posterior credible interval of κ_i. Overall, all models are able to find all of the contaminated districts. Additionally, except for Congdon's model with the logCAR parametrisation, none of the models tend to point out as potential outliers too many of the non-contaminated areas (specificity greater than 99%).

	Offset category	BYM2-Gamma	BYM2-logCAR	Congdon	Congdon-logCAR
Sensitivity	Small	100.0	100.0	100.0	100.0
	Medium low	100.0	100.0	100.0	100.0
	Medium	100.0	100.0	100.0	100.0
	Medium high	100.0	100.0	100.0	100.0
	High	100.0	100.0	100.0	100.0
	Overall	100.0	100.0	100.0	100.0
Specificity	Small	99.9	99.2	100.0	88.1
	Medium low	99.9	100.0	99.9	84.0
	Medium	99.8	100.0	99.9	88.7
	Medium high	99.9	99.8	99.9	79.5
	High	99.9	99.8	100.0	87.1
	Overall	99.9	99.7	99.9	85.5

Table A.2: Sensitivity and specificity of the outlier detection for each model depending on the offset size, in the simulation study with distant outliers.

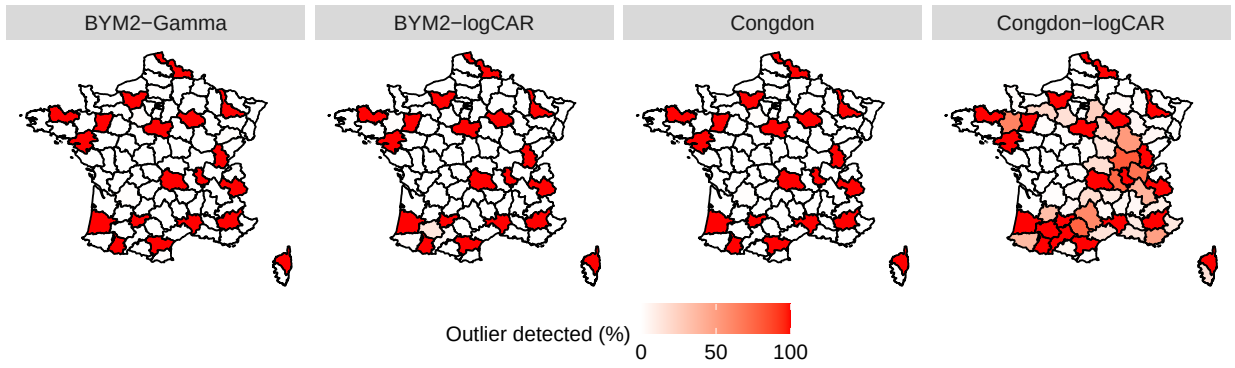


Figure A.16: Percentage of times among 100 replicates that the outliers were identified by each model, in the simulation study with distant outliers. The outliers are pointed out when $\kappa_u < 1$, where κ_u is the upper bound of the posterior 95% credible interval of κ.

A.7 Simulation studies on the map of Rio de Janeiro

In this section, we present the results from simulation studies conducted using the map of Rio de Janeiro wherein some arbitrary areas are contaminated into outlying areas, to assess the performance of the proposed model in comparison to the one proposed by Congdon (2017). Similar to Section 3.3.1, the design of the simulation studies is inspired by Richardson et al.

(2004). The $n = 160$ districts of Rio de Janeiro and their neighbourhood structure are used as the region of study. In the first simulation study (section A.7.1), areas that are far from each other are contaminated into outliers. In the second simulation study (section A.7.2), neighbouring areas are contaminated into outliers. In the third simulation study (section A.7.3), neighbouring areas are contaminated and we include a covariate. In all simulation studies, the goal is to identify the correct districts as outliers.

A.7.1 Distant outliers in Rio

In the first simulation study, 20 districts are arbitrarily chosen to be outliers. The goal is for our proposed model to accurately identify the outliers. Out of simplicity, there are no covariates included in the generating process nor when fitting the models. First, all $n = 160$ latent effects, which correspond to log relative risks in this covariate-free simulation study, are set to 0: $b_i = 0,\ i = 1, \ldots, n$. Then, the offsets $[E_1, \ldots, E_n]^\top$ are taken from the real data application to Zika counts that is presented in section 3.3.2. We define five offset categories based on the empirical offset quantiles. The first category corresponds to the smallest offsets and the fifth category, to the largest ones. The categories are termed "Small" for $E \leq 59.1$, "Medium low" for $E \in (59.1, 112.4]$, "Medium" for $E \in (112.4, 177.2]$, "Medium high" for $E \in (177.2, 267.2)$ and "High" for $E > 267.2$. Based on these categories, we select 20 districts to be outliers. Four districts are chosen from each offset category. That is, there are 4 outliers within the smallest offset group, 4 within the second-to-smallest offset group, and so on. Then, within each group of four districts, the relative risks are contaminated into outliers by setting the log relative risks to be equal to $b_{i'} = \ln(0.25)$, $b_{i''} = \ln(0.5)$, $b_{i'''} = \ln(1.5)$ and $b_{i''''} = \ln(2)$. Figure A.17 maps the 160 districts of Rio de Janeiro, showing which areas are outliers based on the offset category and the contaminated relative risk. Again, all the white areas have a relative risk of 1. Finally, $R = 100$ populations of size $n = 160$ are created according to a hierarchical Poisson model. That is, $Y_i \sim \mathcal{P}(E_i \exp[b_i])$. The only source of randomness across the 100 replicates comes from the repeated sampling

from a Poisson distribution.

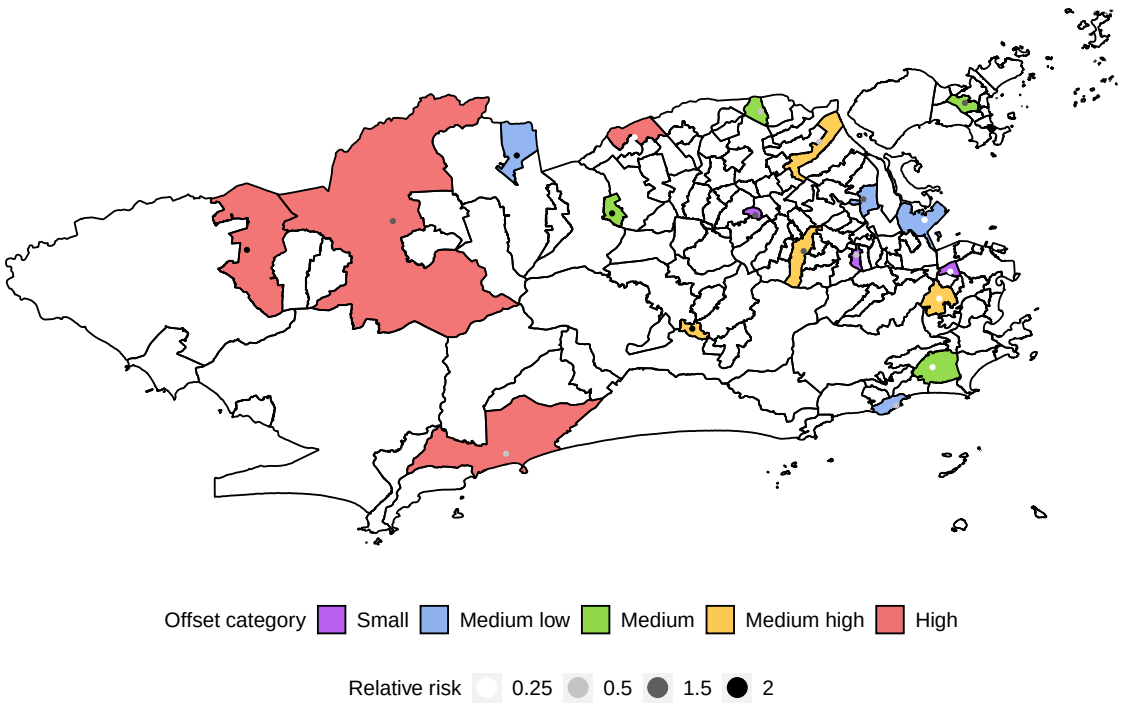


Figure A.17: Districts of Rio de Janeiro city arbitrary chosen to be outliers in the simulation study with distant outliers. Colors depict the offset category based on the empirical offset quantiles. The points represent the relative risk set to each outlying district.

Using the two scale mixtures described in section 3.2.1, the Congdon model is compared to the proposed model. The first version of the proposed model is denoted BYM2-Gamma and the second, BYM2-logCAR. The original Congdon model is termed Congdon, whereas the one with spatially structured scale mixture components is denoted Congdon-logCAR. For the four models, the intercept is given a quite vague prior: $\beta_0 \sim \mathcal{N}(0, 10^2)$ and the mixing parameter, λ, is assigned a uniform, $\mathcal{U}(0, 1)$, prior distribution. The same $\mathcal{N}_+(0, 1)$ prior is considered for σ, which is a *marginal* standard deviation in the proposed model, while it is a *conditional* standard deviation in Congdon's. Finally, in the BYM2-Gamma and Congdon models, the prior distribution for the κ's is described in (3.5) with $\nu \sim \text{Exp}(1/4)$. For the BYM2-logCAR and Congdon-logCAR parametrisations, the κ's follow *a priori* the distribution in (3.6) and we set $\nu \sim \text{Exp}(1/0.3)$.

The models are fitted through the R package `rstan` (Stan Development Team, 2020). For each dataset, the MCMC procedure consists of 2 chains of 20,000 iterations with a 10,000 burn-in period and a thinning factor of 10. Convergence of the chains is assessed through trace plots, effective sample sizes and the $\widehat{R}$ statistic (Gelman and Rubin, 1992; Vehtari et al., 2021).

In terms of WAIC (Watanabe and Opper, 2010), the proposed BYM2-Gamma model performs better than Congdon's, on average, as shown in Figure A.18. The BYM2-Gamma and original Congdon models always perform better than the models that include spatially structured scaling mixture components. On average, the BYM2-logCAR model yields a criterion of 1289.5 versus 1288.6 for the Congdon-logCAR model, while the BYM2-Gamma model presents a WAIC of 1260.9, versus 1263.9 for Congdon's.

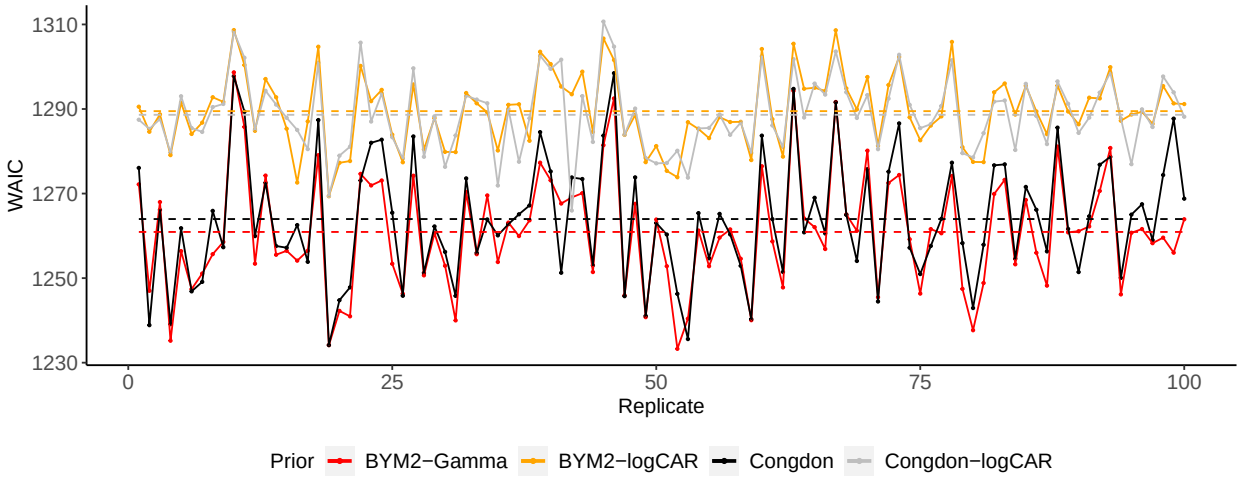


Figure A.18: WAIC across the 100 replicates for the proposed models and Congdon's in the simulation study with distant outliers in Rio de Janeiro. Dashed lines: mean WAIC for each model.

The models' performances are also compared in terms of MSE, as shown in Figure A.19. Again, as expected, all models yield smaller MSEs in areas with larger offsets. Additionally, all models tend to better fit the data in areas that are not outliers, that is in the areas with a relative risk of 1. On average over the 100 replicated datasets and across all areas, the MSEs

are of 0.005 for the BYM2-Gamma model, 0.006 for Congdon and 0.008 for both log-CAR parametrisations.

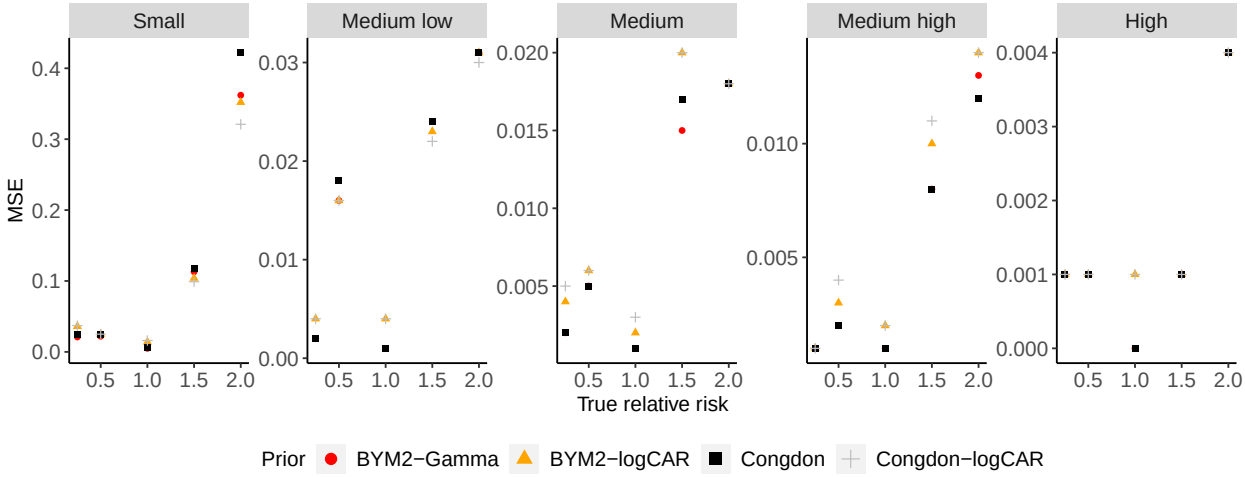


Figure A.19: MSE over the 100 replicates for the proposed models and Congdon's according to the true relative risk and the offset size in the simulation study with distant outliers in Rio de Janeiro.

Regarding the detection of outliers, which is the main focus of this simulation study, Table A.3 shows how often each model accurately detects districts as outliers (sensitivity) and non-outliers (specificity), depending on the offset category. Additionally, Figure A.20 shows how often each district is detected as a potential outlier by the four models, while indicating the offset sizes. Area i is detected as an outlier when $\kappa_{u,i} < 1$, where $\kappa_{u,i}$ is the upper bound of the 95% posterior credible interval of κ_i. Overall, all models are able to find the contaminated districts in the four upper offset categories. When the offsets are the smallest, all models detect the outliers only half of the time, with a slight advantage for the proposed models (e.g. sensitivity of 55.5 for BYM2-Gamma versus 50.25 for Congdon). In this simulation study where outliers are distant, the parametrisations with spatially structured scaling mixture components tend to identify slightly more outliers than are truly present in the data (e.g. specificities of 95.4 versus 90.2 for BYM2-logCAR and Congdon-logCAR, respectively).

	Offset category	BYM2-Gamma	BYM2-logCAR	Congdon	Congdon-logCAR
Sensitivity	Small	55.50	54.00	50.25	49.50
	Medium low	94.25	94.75	91.50	94.75
	Medium	99.50	94.75	98.50	95.00
	Medium high	100.00	99.00	100.00	99.50
	High	100.00	100.00	100.00	98.50
	Overall	89.85	88.50	88.05	87.45
Specificity	Small	99.93	97.82	99.93	95.04
	Medium low	100.00	95.39	100.00	90.21
	Medium	100.00	99.64	100.00	98.93
	Medium high	99.93	99.57	99.93	98.11
	High	99.96	99.89	99.96	99.25
	Overall	99.96	98.46	99.96	96.31

Table A.3: Sensitivity and specificity of the outlier detection for each model depending on the offset size in the simulation study with distant outliers in Rio de Janeiro.

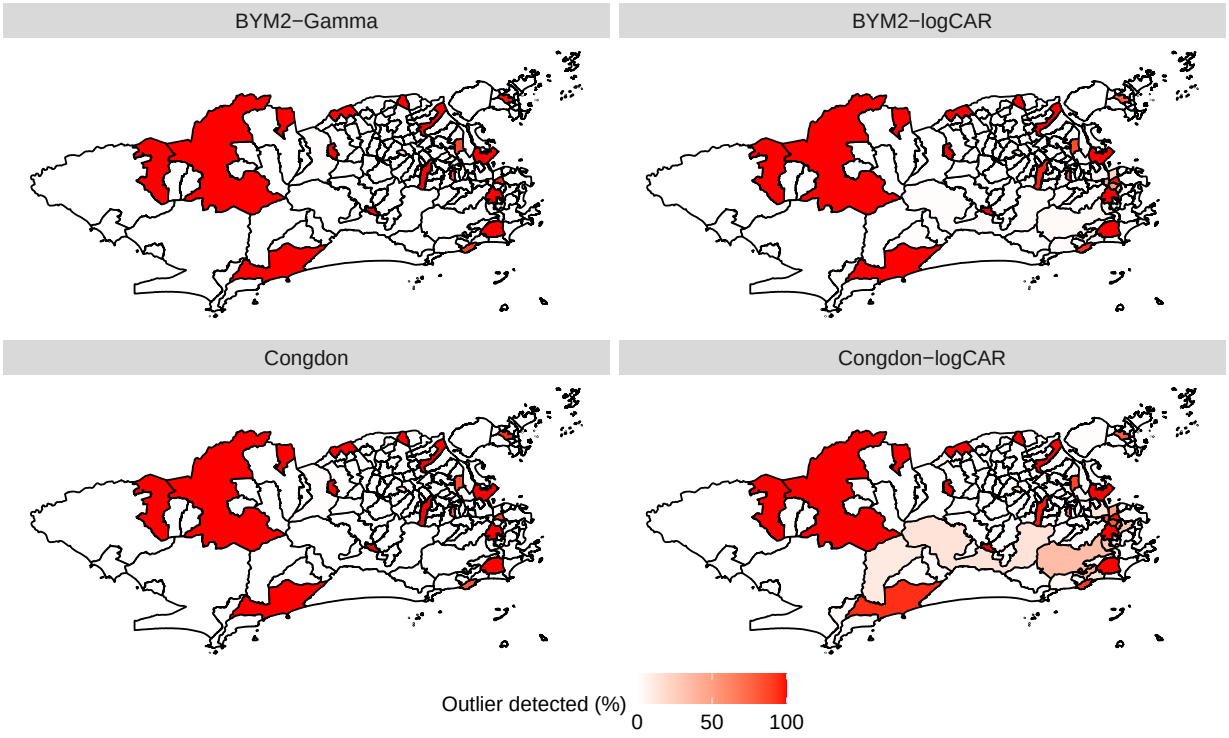


Figure A.20: Percentage of times among 100 replicates that the outliers were identified by each model, in the simulation study with distant outliers in Rio de Janeiro. The outliers are pointed out when $\kappa_u < 1$, where κ_u is the upper bound of the posterior 95% credible interval of κ.

A.7.2 Neighbouring outliers in Rio

In this second simulation study using the map of Rio de Janeiro, 20 districts are contaminated such that 2 groups of 10 neighbouring outliers are created. Once again, there are no covariates in this analysis and all areas are first imposed a relative risk of 1. Similarly to section A.6, the offsets $[E_1, \ldots, E_n]^\top$ are taken from the Zika data analysis from section 3.3.2. Hence, the same five offset categories are defined. Then, 20 districts are selected to be outliers, such that each group of 10 neighbouring outliers contains 2 areas of each offset category. Within each such pair of districts, the relative risks are contaminated into outliers by setting $b_i = \ln(0.5)$ and $b_{i'} = \ln(1.5)$. The resulting outliers are mapped in Figure A.21, highlighting the offset sizes and imposed relative risks. Again, $R = 100$ populations of size $n = 160$ are created by generating the number of cases $Y_i \sim \mathcal{P}\left(E_i \exp[b_i]\right)$. The same four models with priors defined in section A.6 are fitted through `rstan`. After 20,000 iterations with a burn-in period of 10,000 and a thinning factor of 10, the 2 MCMC chains attained convergence as assessed by trace plots, effective sample sizes and $\widehat{R}$ statistics.

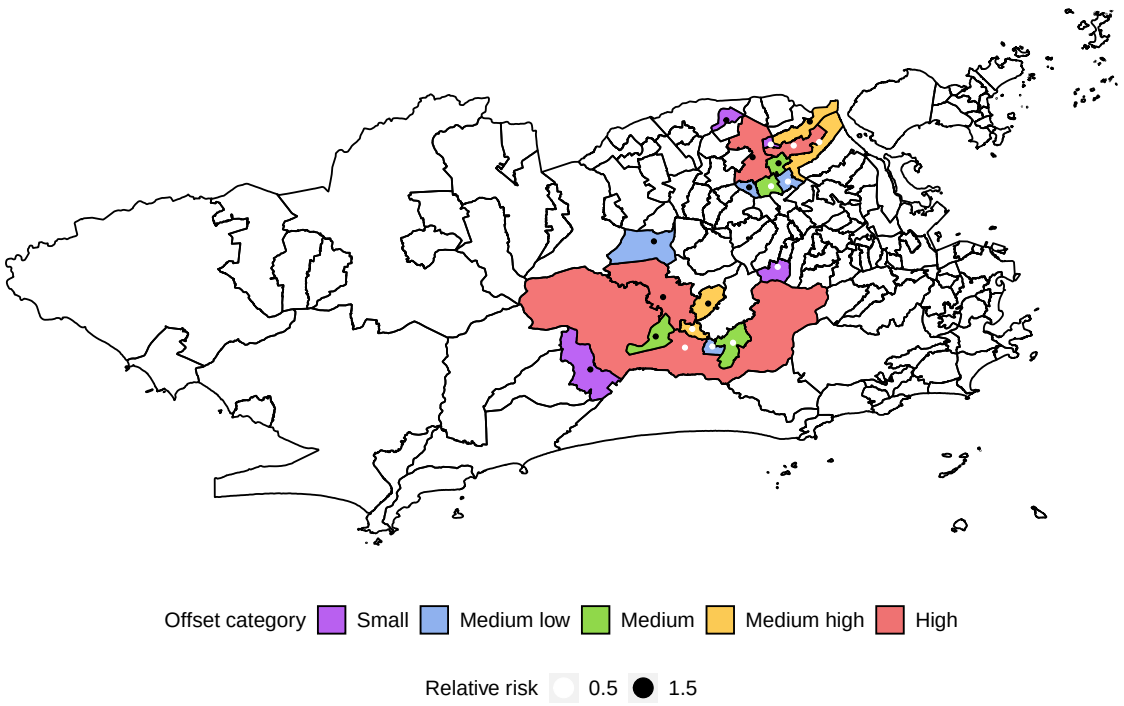


Figure A.21: Districts of Rio de Janeiro city arbitrary chosen to be outliers in the simulation study with neighbouring outliers. Colors depict the offset category based on the empirical offset quantiles. The points represent the relative risk set to each outlying district.

In terms of WAIC, as shown in Figure A.22, Congdon performs slightly worse than the other three models, with an average value of 1275, versus 1270 for Congdon-logCAR the two proposals.

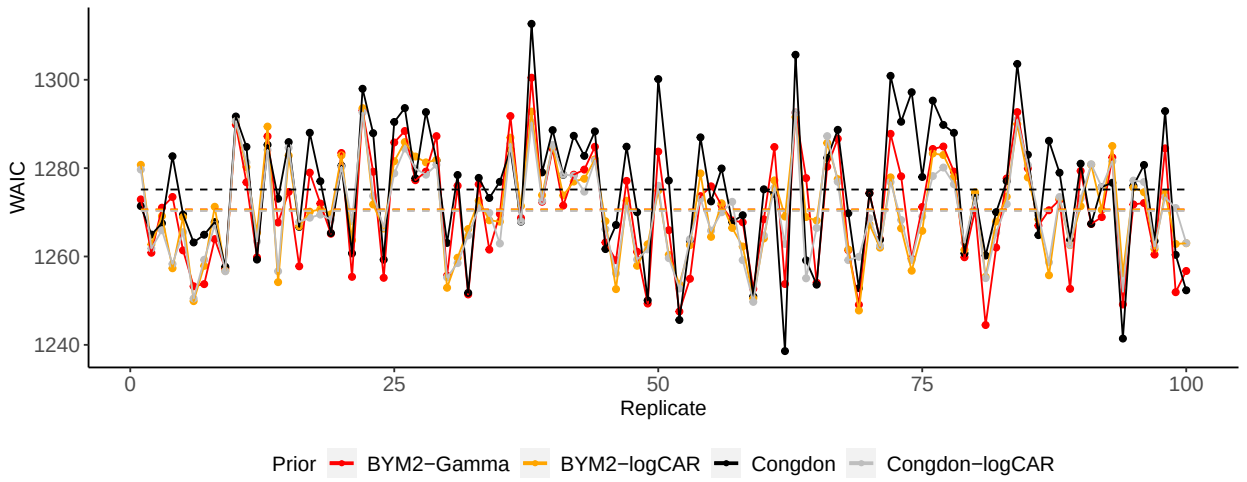


Figure A.22: WAIC across the 100 replicates for the proposed models and Congdon's in the simulation study with neighbouring outliers in Rio de Janeiro. Dashed lines: mean WAIC for each model.

In terms of MSE, as expected, all models fit better the data in areas with higher offsets than in areas with smaller offsets, as shown in Figure A.23. Again, all models better fit the data in areas that are not outliers, areas with a relative risk of 1. Over the 100 replicates and all areas, the four models perform similarly, with an average MSE of 0.004.

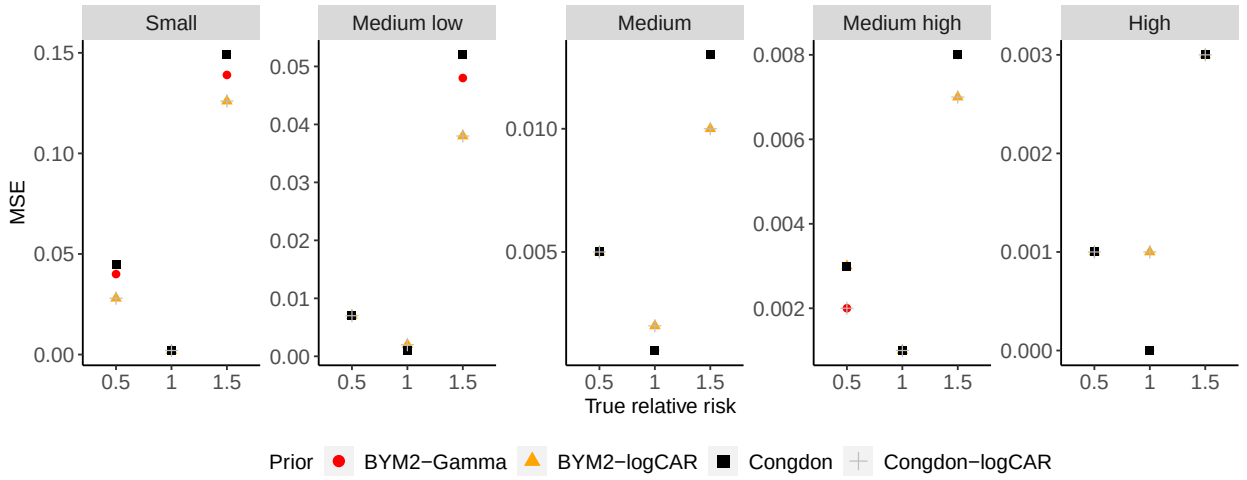


Figure A.23: MSE over the 100 replicates for the proposed models and Congdon's according to the true relative risk and the offset size, in the simulation study with neighbouring outliers in Rio de Janeiro.

Regarding the detection of outliers, the results are summarised in Table A.4 and Figure A.24. Similarly to the previous simulation study with distant outliers, both models with spatially structured κ's tend to identify more outliers than truly present in the data (e.g. overall specificities of 97% and 96.5% for BYM2-logCAR and Congdon-logCAR, respectively, versus 99.9% for both BYM2-Gamma and Congdon). In the smallest offset category, all models often miss the outliers, with a clear advantage for the models with spatially structured κ's (e.g. sensitivity of about 30% for BYM2-Gamma and Congdon versus 64% for BYM2-logCAR and Congdon-logCAR). Regardless of the offset size, the BYM2-Gamma model performs better than Congdon's in terms of detected outliers. In particular, in the third offset category, the BYM2-Gamma model misses outliers only 1.5% of the time versus 18.75% for Congdon's model.

	Offset category	BYM2-Gamma	BYM2-logCAR	Congdon	Congdon-logCAR
Sensitivity	Small	35.25	64.25	30.50	64.00
	Medium low	80.25	93.00	64.25	89.00
	Medium	98.50	100.00	81.25	96.25
	Medium high	100.00	100.00	91.00	100.00
	High	100.00	100.00	93.75	97.75
	Overall	82.80	91.45	72.15	89.40
Specificity	Small	100.00	100.00	99.93	100.00
	Medium low	99.96	98.07	99.96	98.00
	Medium	99.89	93.79	99.93	92.50
	Medium high	99.96	96.50	100.00	95.79
	High	99.96	96.36	99.71	96.21
	Overall	99.96	96.96	99.91	96.51

Table A.4: Sensitivity and specificity of the outlier detection for each model depending on the offset size in the simulation study with neighbouring outliers in Rio de Janeiro.

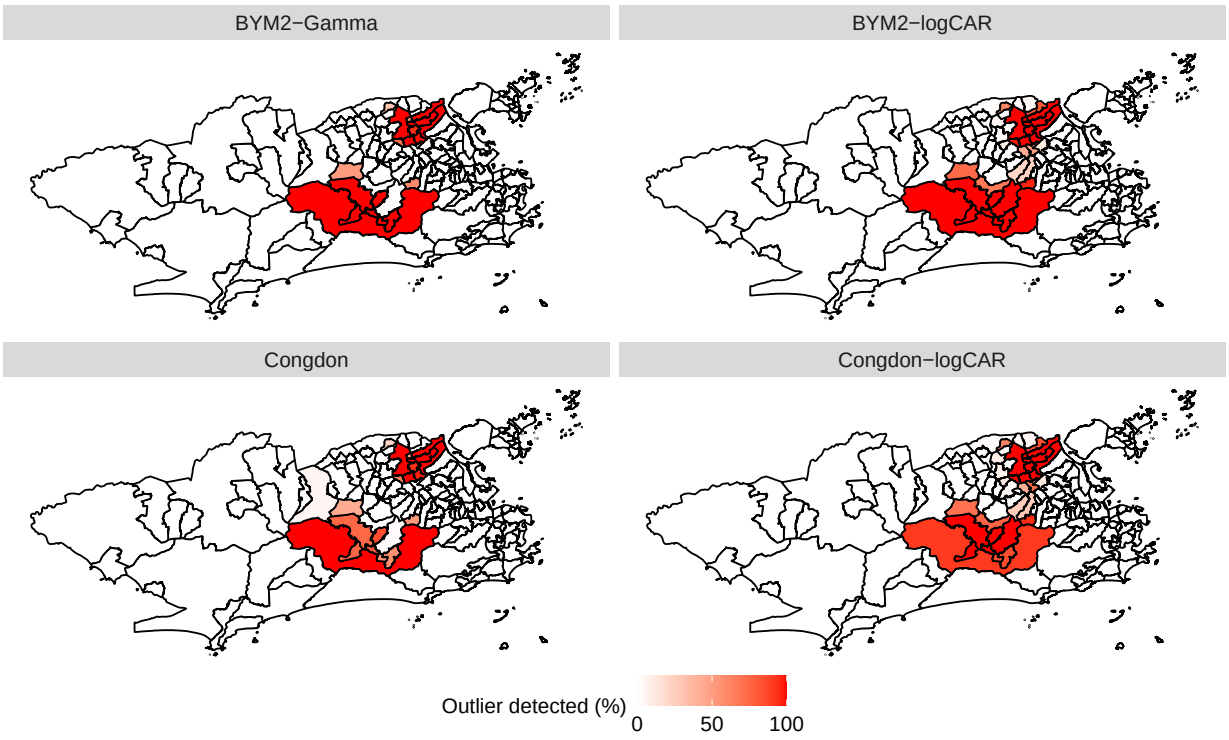
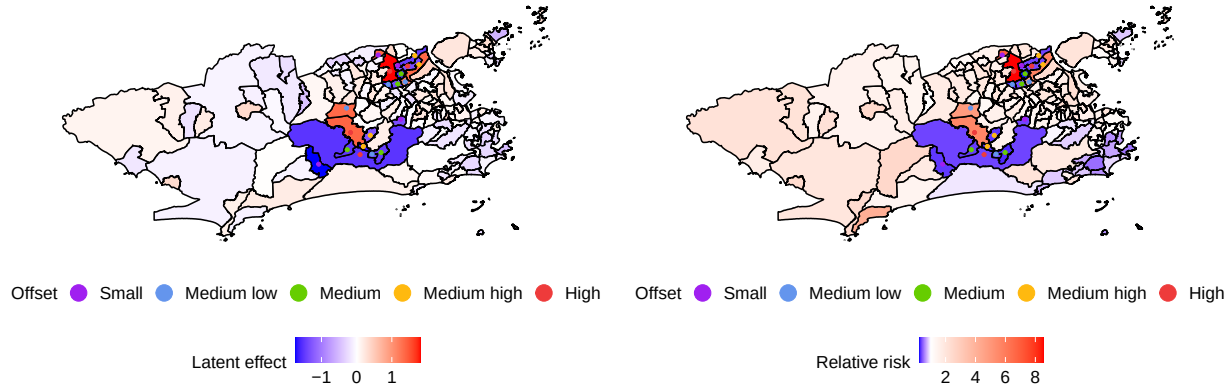


Figure A.24: Percentage of times among 100 replicates that the outliers were identified by each model, in the simulation study with neighbouring outliers in Rio de Janeiro. The outliers are pointed out when $\kappa_u < 1$, where κ_u is the upper bound of the posterior 95% credible interval of κ.

A.7.3 Neighbouring outliers with a covariate in Rio

In this third simulation study using the map of Rio de Janeiro, the same offset categories and 2 groups of 10 neighbouring outliers as in section A.7.2 are chosen. Again, the goal is to identify the outlying areas. First, all $n = 160$ latent effects are generated following a proper CAR (PCAR) distribution: $\boldsymbol{b} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \left[\boldsymbol{D} - \alpha \boldsymbol{W}\right]^{-1}\right)$, where the matrices $\boldsymbol{W}$ and $\boldsymbol{D}$ are computed as defined in Section 3.1.1, using the neighbourhood structure of Rio de Janeiro. We set $\sigma^2 = 0.1$ and $\alpha = 0.99$ such that the proper spatial distribution is close to an ICAR distribution. Following Section A.7.2, four districts are chosen from each offset category and their generated latent effects are contaminated as $b_i^{contam} = b_i + e_i$, with $e_i \sim \mathcal{U}\left(2\max(|b_{(1)}|, |b_{(n)}|), 3\max(|b_{(1)}|, |b_{(n)}|)\right)$, where $b_{(1)}$ and $b_{(n)}$ denote the minimum and maximum generated latent effects, respectively. Figure A.25 (a) maps the resulting

160 latent effects, showing which areas are outliers based on the offset category. Finally, $R = 100$ populations of size $n = 160$ are created according to the hierarchical Poisson model $Y_i \sim \mathcal{P}\left(E_i \exp[\beta_0 + \beta x_i + b_i]\right)$, where $\beta_0 = 2.5$, $\beta = -3.5$ and the covariate x is the development index taken from the real data application to Zika counts presented in Section 3.3.2. The resulting relative risks are mapped in Figure A.25 (b), showing the outlying areas based on the offset category. Once again, the same four models are fitted through `rstan` and convergence of the 2 MCMC chains was attained after 20,000 iterations with a burn-in period of 10,000 and a thinning factor of 10.



(a) (b)

Figure A.25: Rio de Janeiro maps of the latent effects (a) and relative risks (b) after contamination, in the simulation study with a covariate and neighbouring outliers. The coloured points depict the offset category based on the empirical offset quantiles.

In terms of WAIC, the proposed BYM2-Gamma model performed the best, with a mean WAIC of 1383 over the 100 replicates. As shown in Figure A.26, the other three models' performances are similar to each other, with average values of 1389 (Congdon), 1390 (BYM2-logCAR) and 1388 (Congdon-logCAR).

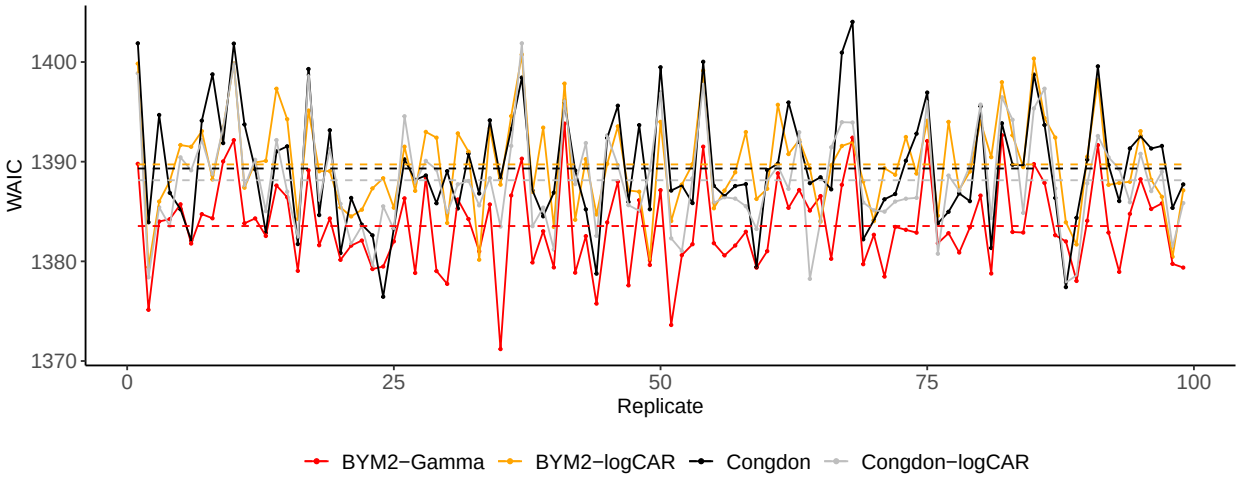


Figure A.26: WAIC across the 100 replicates for the proposed models and Congdon's, in the simulation study with a covariate and neighbouring outliers in Rio de Janeiro. Dashed lines: mean WAIC for each model.

Figure A.27 shows each model's MSE for every districts across the different offset categories. The four models yield again smaller MSEs in districts with relative risks closer to 1, regardless of the offset size. Additionally, regardless of the relative risk size, all models reach smaller MSEs values for larger offset values. Overall, the proposed BYM2-Gamma model performed better with a mean MSE of 0.0189, versus 0.0212, 0.0204 and 0.0202, for the proposed BYM2-logCAR, Congdon and Congdon-logCAR models, respectively.

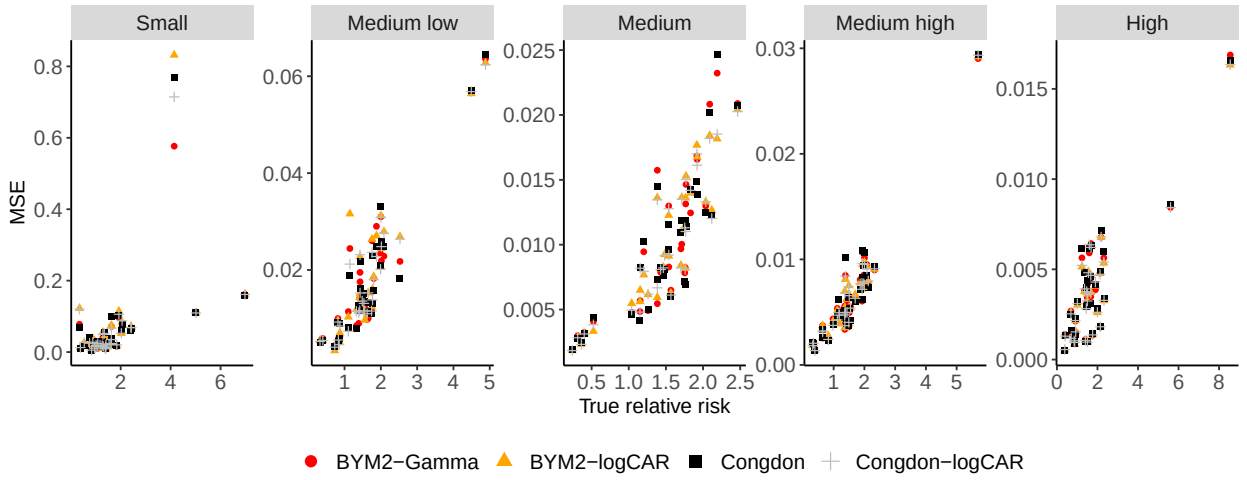


Figure A.27: MSE over the 100 replicates for the proposed models and Congdon's according to the true relative risk and the offset size, in the simulation study with a covariate and neighbouring outliers in Rio de Janeiro.

Table A.5 shows the sensitivities and specificities of outlier identification produced by each model across the five offset categories. The proposed BYM2-Gamma model performs better in both identifying the correct outliers, and not pointing out the non-contaminated areas. Overall, Congdon's model misses some outlying districts 8% of the time, and up to 19% of the time, in the fourth offset category. The proposed spatially structured prior for the mixture components improved Congdon's model performance, where Congdon-logCAR only misses 2% of the contaminated areas, overall. Additionally, Congdon's model tends to capture more outliers than were contaminated, like the western and eastern non-contaminated districts that are detected 75% of the time, as shown in Figure A.28.

	Offset category	BYM2-Gamma	BYM2-logCAR	Congdon	Congdon-logCAR
Sensitivity	Small	98.5	96.0	99.2	97.0
	Medium low	100.0	99.2	90.4	99.0
	Medium	100.0	99.8	90.9	99.0
	Medium high	100.0	99.5	81.8	99.5
	High	100.0	99.8	100.0	99.0
	Overall	99.7	98.8	92.5	98.7
Specificity	Small	99.3	99.9	99.1	99.9
	Medium low	98.4	99.5	96.9	98.5
	Medium	99.4	98.2	95.6	96.8
	Medium high	99.8	99.9	99.4	99.9
	High	98.3	96.3	97.9	96.3
	Overall	99.0	98.8	97.8	98.3

Table A.5: Sensitivity and specificity of the outlier detection for each model depending on the offset size, in the simulation study with a covariate and neighbouring outliers in Rio de Janeiro.

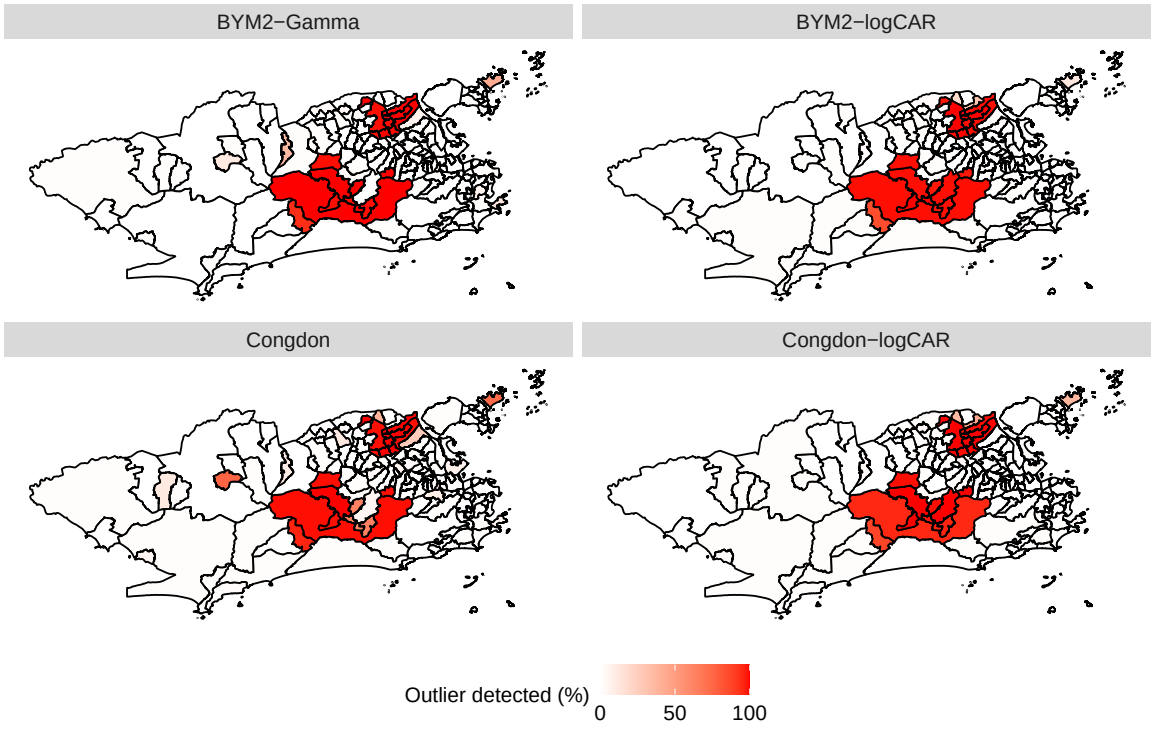


Figure A.28: Percentage of times among 100 replicates that the outliers were identified by each model, in the simulation study with a covariate and neighbouring outliers in Rio de Janeiro. The outliers are pointed out when $\kappa_u < 1$, where κ_u is the upper bound of the posterior 95% credible interval of κ.

A.8 Comparison with the model proposed by Corpas-Burgos and Martinez-Beneito (2020)

Table A.6 compares the definitions of the proposed model, and the ones proposed by Congdon (2017) and Corpas-Burgos and Martinez-Beneito (2020) (CB-MB). Further, Table A.6 provides a comparison of the models depending on the spatial dependence parameter, λ.

Model definitions		
Congdon	$b_i \mid \boldsymbol{b}_{(-i)} \sim \mathcal{N}\left(\frac{\lambda}{1-\lambda+\lambda\left(\sum_{j=1}^{n} w_{ij}\right)} \sum_{j=1}^{n} w_{ij}\kappa_j b_j, \frac{\sigma^2}{\kappa_i\left[1-\lambda+\lambda\left(\sum_{j=1}^{n} w_{ij}\right)\right]}\right)$	
CB-MB	$b_i \mid \boldsymbol{b}_{(-i)} \sim \mathcal{N}\left(\frac{\lambda}{1-\lambda+\lambda\left(\sum_{j=1}^{n} w_{ij}\sqrt{c_j}\right)} \sum_{j=1}^{n} w_{ij}\sqrt{c_j} b_j, \frac{\sigma^2}{\sqrt{c_i}\left[1-\lambda+\lambda\left(\sum_{j=1}^{n} w_{ij}\sqrt{c_j}\right)\right]}\right)$	
Our Proposal	$b_i = \sigma/\sqrt{\kappa_i}\left(\sqrt{1-\lambda}\theta_i + \sqrt{\lambda/h}u_i\right)$	
	$\lambda = 0$	$\lambda = 1$
Congdon	$b_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2/\kappa_i)$	$b_i \mid \boldsymbol{b}_{(-i)} \sim \mathcal{N}\left(\frac{1}{\left(\sum_{j=1}^{n} w_{ij}\right)} \sum_{j=1}^{n} w_{ij}\kappa_j b_j, \frac{\sigma^2}{\kappa_i\left(\sum_{j=1}^{n} w_{ij}\right)}\right)$
CB-MB	$b_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2/\sqrt{c_i}\right)$	$b_i \mid \boldsymbol{b}_{(-i)} \sim \mathcal{N}\left(\frac{1}{\left(\sum_{j=1}^{n} w_{ij}\sqrt{c_j}\right)} \sum_{j=1}^{n} w_{ij}\sqrt{c_j} b_j, \frac{\sigma^2}{\sqrt{c_i}\left(\sum_{j=1}^{n} w_{ij}\sqrt{c_j}\right)}\right)$
Our Proposal	$b_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2/\kappa_i)$	$b_i \mid \boldsymbol{b}_{(-i)} \sim \mathcal{N}\left(\frac{1}{\left(\sum_{j=1}^{n} w_{ij}\right)} \sum_{j=1}^{n} w_{ij}\sqrt{\frac{\kappa_j}{\kappa_i}} b_j, \frac{\sigma^2/h}{\kappa_i\left(\sum_{j=1}^{n} w_{ij}\right)}\right)$

Table A.6: Comparison of the models introduced by Congdon (2017), Corpas-Burgos and Martinez-Beneito (2020) (CB-MB), and the heavy-tailed BYM2 proposal. In our proposal, the unstructured component θ_i is independent of the spatially structured component u_i.

APPENDIX B

Appendix to Manuscript 2

B.1 Stan code for the proposed Heavy Rushworth model

Listing B.1: Stan code for the Heavy Rushworth proposed model

```
data {
  int<lower=1> N; // number of areas
  int<lower=1> TT; // number of time points
  int<lower=1> NT; // number of areas * number of time points
  vector<lower=1, upper=N>[N] d; // vector of the number of neighbours for each area
  matrix<lower=0, upper=1>[N,N] W; // matrix of spatial weights
  vector[N] zeros;
  int<lower=0> y[NT]; // long vector of cases ordered by time: (y_11, ..., y_n1, ...,
      y_1T, ..., y_nT)
  vector[NT] log_E; // Offset in the log scale
  vector[NT] X; // vector of covariates
}

```

```
13 parameters {
14   real beta0; // intercept
15   real beta; // regression parameter
16   real<lower=0> sigma; // conditional std deviation
17   real<lower=0, upper=1> lambda; // spatial dependence parameter
18   real<lower=-1, upper=1> alpha; // temporal dependence parameter
19   vector<lower=0>[N] kappa; // outlier indicator
20   real<lower=0> nu; // hyperparameter for kappa
21   vector[NT] s; // spatial effects
22 }
23
24 transformed parameters {
25  matrix[N,N] PrecMat; // Precision matrix for the proposed model
26  PrecMat = (1/sigma^2)*(diag_matrix(kappa .* (1-lambda + lambda*d)) - lambda * W
        .* (kappa*(kappa')));
27 }
28
29 model {
30   y ~ poisson_log(log_E + beta0 + X*beta + s);
31
32   // Prior for the latent effects at time 1
33   s[1:N] ~ multi_normal_prec(zeros, PrecMat);
34   // soft sum-to-zero constraint to avoid identifiability issues with the intercept:
35   sum(s[1:N]) ~ normal(0, 0.001 * N);
36   for(t in 2:TT){
37     // Prior for the latent effects at time 2, ..., T
```

```
38      s[((t-1)*N+1):((t-1)*N+N)] ~
            multi_normal_prec(alpha*s[((t-2)*N+1):((t-2)*N+N)], PrecMat);
39    }
40
41    beta0 ~ normal(0.0, 1.0);
42    beta ~ normal(0.0, 1.0);
43    nu ~ exponential(1.0/4.0);
44    sigma ~ normal(0.0,0.1);
45    lambda ~ uniform(0.0,1.0);
46    kappa ~ gamma(nu/2.0, nu/2.0);
47    alpha ~ uniform(-1.0,1.0);
48  }
```

B.2 Simulation study: data generated from the proposed model

In this section, we present the results from a simulation study where 100 replicated datasets are generated from the proposed model. The aim is to verify that the proposal is able to recover the true values of all the model parameters. Similar to Section 4.3.1, the $n = 33$ boroughs of Montreal are considered over $T = 52$ time points. The overall log risk is set to $\beta_0 = -1$ and the offsets are taken from a Poisson distribution, $E_i \sim \text{Pois}(40),\ i = 1, \ldots, n$. To generate the $n \times T$ latent effects b_{it}, the scaling mixture components $\kappa_1, \ldots, \kappa_n$ are first generated from a Gamma$(\nu/2, \nu/2)$ distribution, with $\nu = 4$. Then, the latent effects are generated from the proposed model (4.3), with $\lambda = 0.9$, $\sigma = 0.1$, and $\alpha = 1$. Finally, the datasets are created such that $Y_{it} \sim \text{Pois}\left(E_i \exp\left(\beta_0 + b_{it}\right)\right),\ i = 1, \ldots, n,\ t = 1, \ldots, T$.

Again, similar to Section 4.3.1, the proposed model is fitted both assuming a uniform prior

on α and fixing $\alpha = 1$, denoted HR(α) and HR(1), respectively. The `rstan R` package is used (Stan Development Team, 2020) and convergence of the two MCMC chains is attained after 5,000 iterations with a burn-in period of 2,500 and a thinning factor of 5, as assessed through trace plots, $\widehat{R}$ statistics (Gelman and Rubin, 1992; Vehtari et al., 2021) and effective sample sizes.

Figure B.1 shows that across the 100 replicates, the posterior summaries (mean and 95% credible interval) obtained from fitting both versions of the proposed model cover the true values of all the scalar parameters. It is worth mentioning that the prior for α is a uniform distribution over the interval (-1,1). Since this prior does not include 1, the posterior 95% credible intervals for α cannot not recover 1, although the estimation is close to the truth. Interestingly, λ tends to be underestimated on average, with posterior means approximately 0.65. However, the posterior 95% credible intervals do recover the true values for this parameter. Figure B.2 shows the posterior summaries for the scaling mixture components resulting from both parametrisations in the first simulation replicate. The results for the other 99 replicates are similar to the ones presented here and across all the replicates and all boroughs. For these κ parameters, the 95% posterior credible intervals' coverage rates are 94.6% and 94.5% for HR(α) and HR(1), respectively. Further, the prior summaries (horizontal solid line and dashed lines) help visualise that the proposal is able to differ from the prior and learn from the data in order to identify potential outliers (e.g., Pierrefonds-Roxboro, Rosemont-La-Petite-Patrie). Finally, Figure B.3 summarises the latent effects' posterior distribution estimated over time by HR(1) and HR(α) across 5 different replicates (columns) and for 5 different boroughs (rows). Through both models, the posterior means follow the true trend of the latent effects and the credible intervals capture the true values 95.5% and 95.3% of the time, for HR(α) and HR(1), respectively.

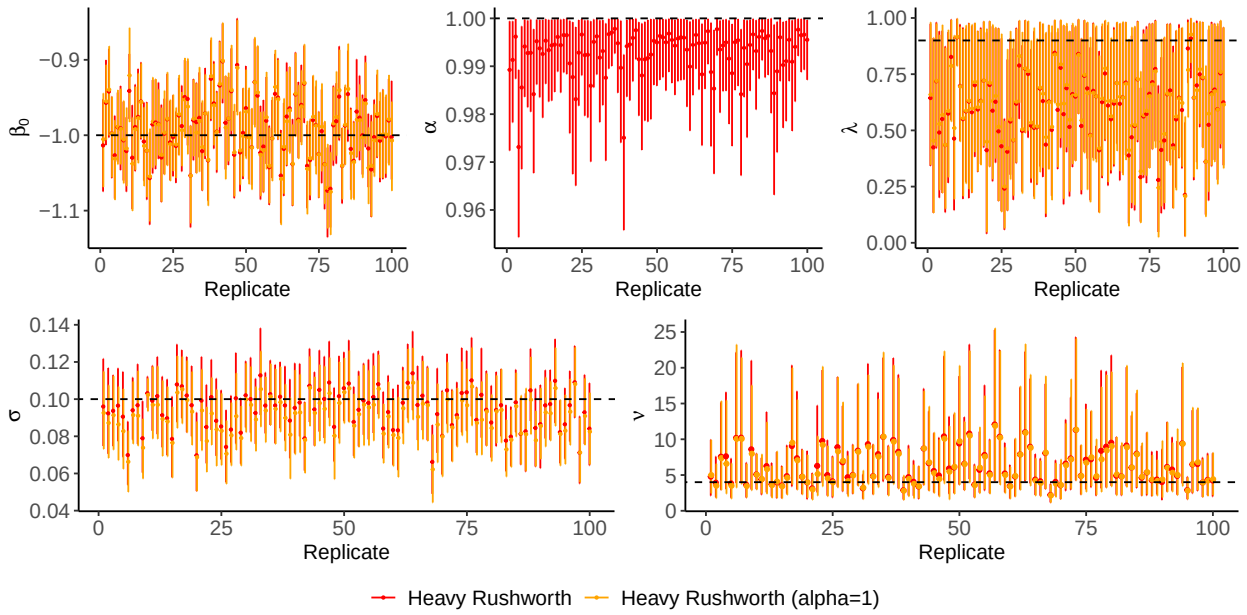


Figure B.1: Parameters' posterior summaries obtained from both versions of the proposed model across the 100 replicates in the simulation study where data are generated from the proposed model. Circles: posterior means; Vertical lines: posterior 95% credible intervals; Dashed lines: true parameter values.

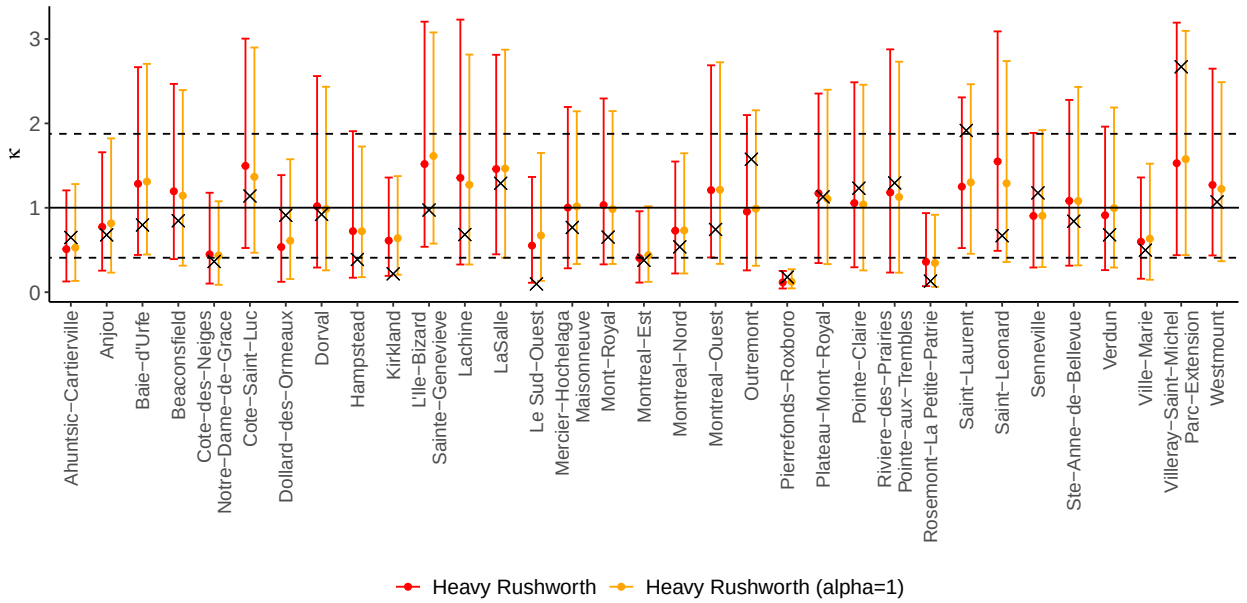


Figure B.2: Posterior summaries for the scaling mixture components obtained from both versions of the proposed model in the first replicate of the simulation study where data are generated from the proposed model. Circles: posterior means; Vertical lines: posterior 95% credible intervals; Crosses: true values; Solid horizontal line: prior mean; Dashed horizontal lines: prior 95% credible interval.

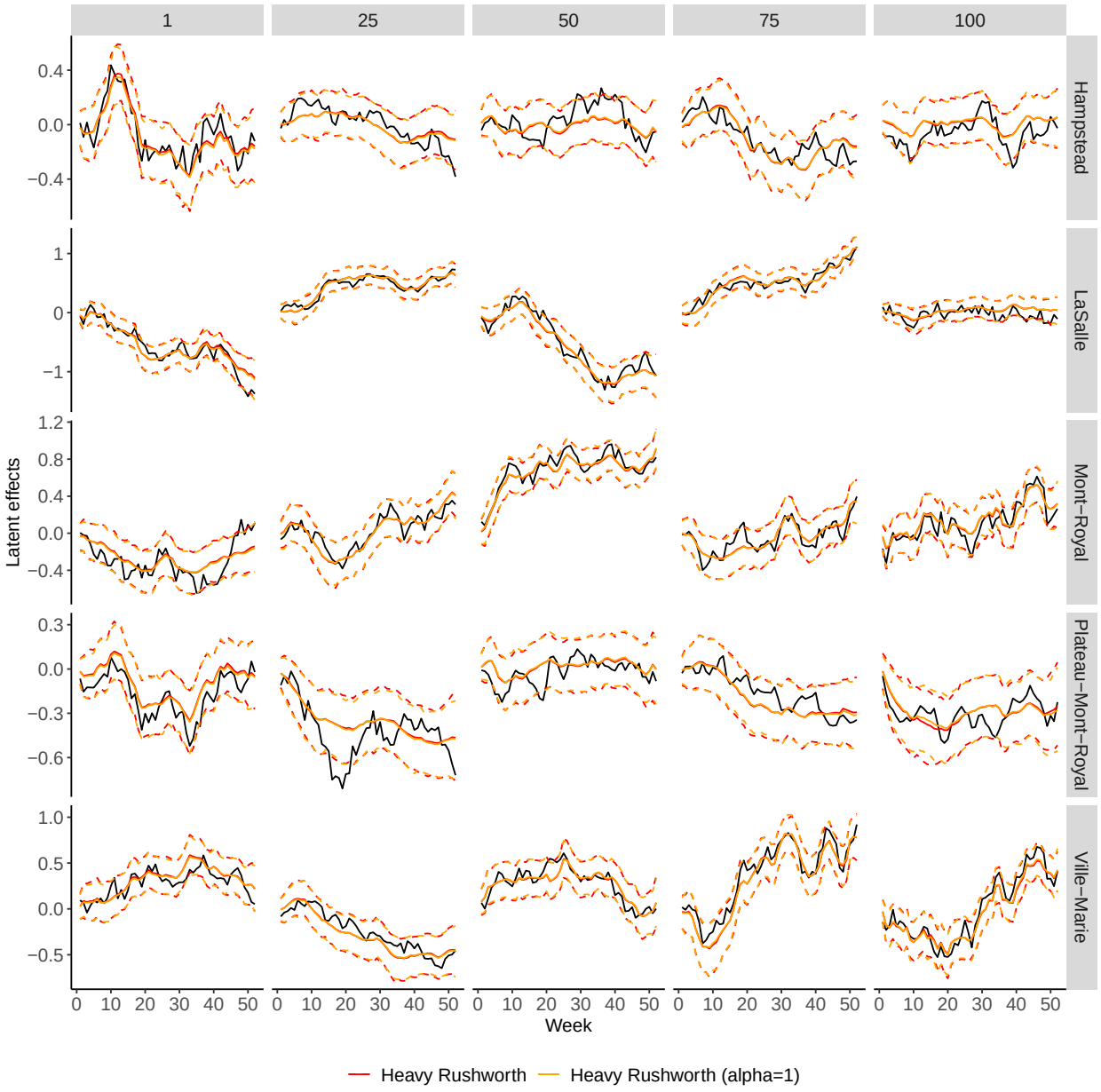


Figure B.3: Posterior summaries for the latent effects obtained over time from both versions of the proposed model in 5 different boroughs (rows) across 5 different replicates (columns) of the simulation study where data are generated from the proposed model. Solid coloured lines: posterior means over time; Dashed coloured lines: 95% posterior credible intervals; Solid black lines: true values over time.

B.3 Supplementary material for the simulation study shown in Section 4.3.1

In this section, additional figures are presented to complete Section 4.3.1. Figure B.4 shows the latent effects over time before (dashed gray line) and after (solid black line) contamination for the 5 outliers in both simulation scenarios. As a comparison, the distribution of the latent effects of a non-contaminated borough (Ahuntsic-Cartierville) is also shown over time under both simulation scenarios. Figure B.5 summarises the WAIC and MSE obtained for each model and each scenario. The WAIC values are shown for each replicate and the MSEs are distinguished between the offset sizes. For each model and each replicate, let $\boldsymbol{Y} = [Y_{11}, \ldots, Y_{n1}, \ldots, Y_{1T}, \ldots, Y_{nT}]^\top$, and the WAIC is computed as follows: $\text{WAIC} = -2\sum_{i,t} \ln\left(\mathbb{E}\left[f\left(Y_{it} \mid \boldsymbol{\theta}\right) \mid \boldsymbol{Y}\right]\right) + 2\sum_{i,t} \mathbb{V}\left[\ln\left(f\left(Y_{it} \mid \boldsymbol{\theta}\right)\right) \mid \boldsymbol{Y}\right]$, where $f(\cdot \mid \boldsymbol{\theta})$ is the likelihood that corresponds to a particular model with set of parameters, $\boldsymbol{\theta}$.

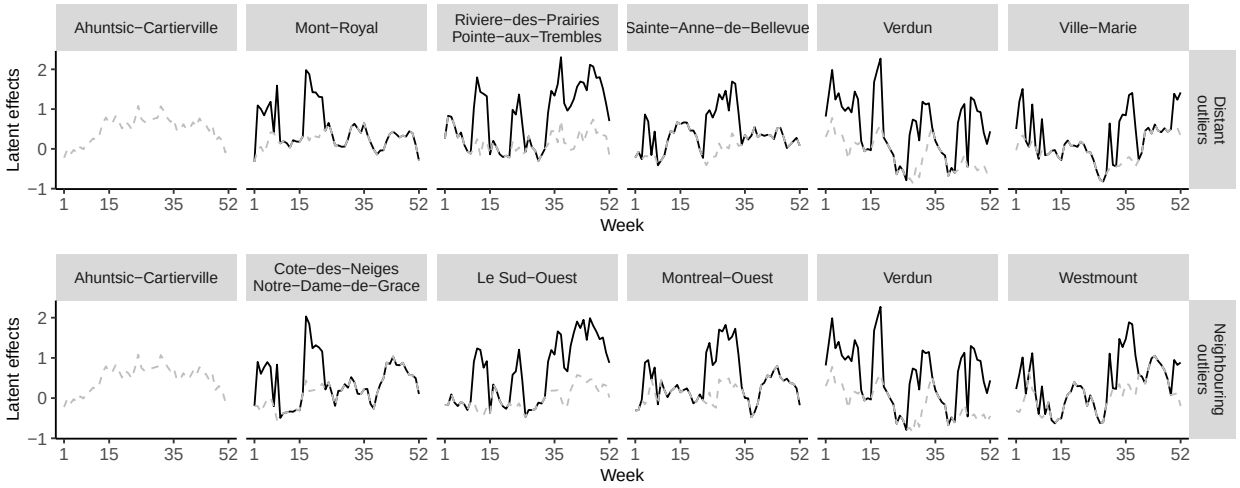


Figure B.4: Generated latent effects over time for the five outliers of each simulation scenario and Ahuntsic-Cartierville. Dashed gray line: latent effects generated from the Rushworth model; Solid black line: latent effects after contamination. The periods where the two lines overlap correspond to periods of non-contamination ($r_{jt} = 0$).

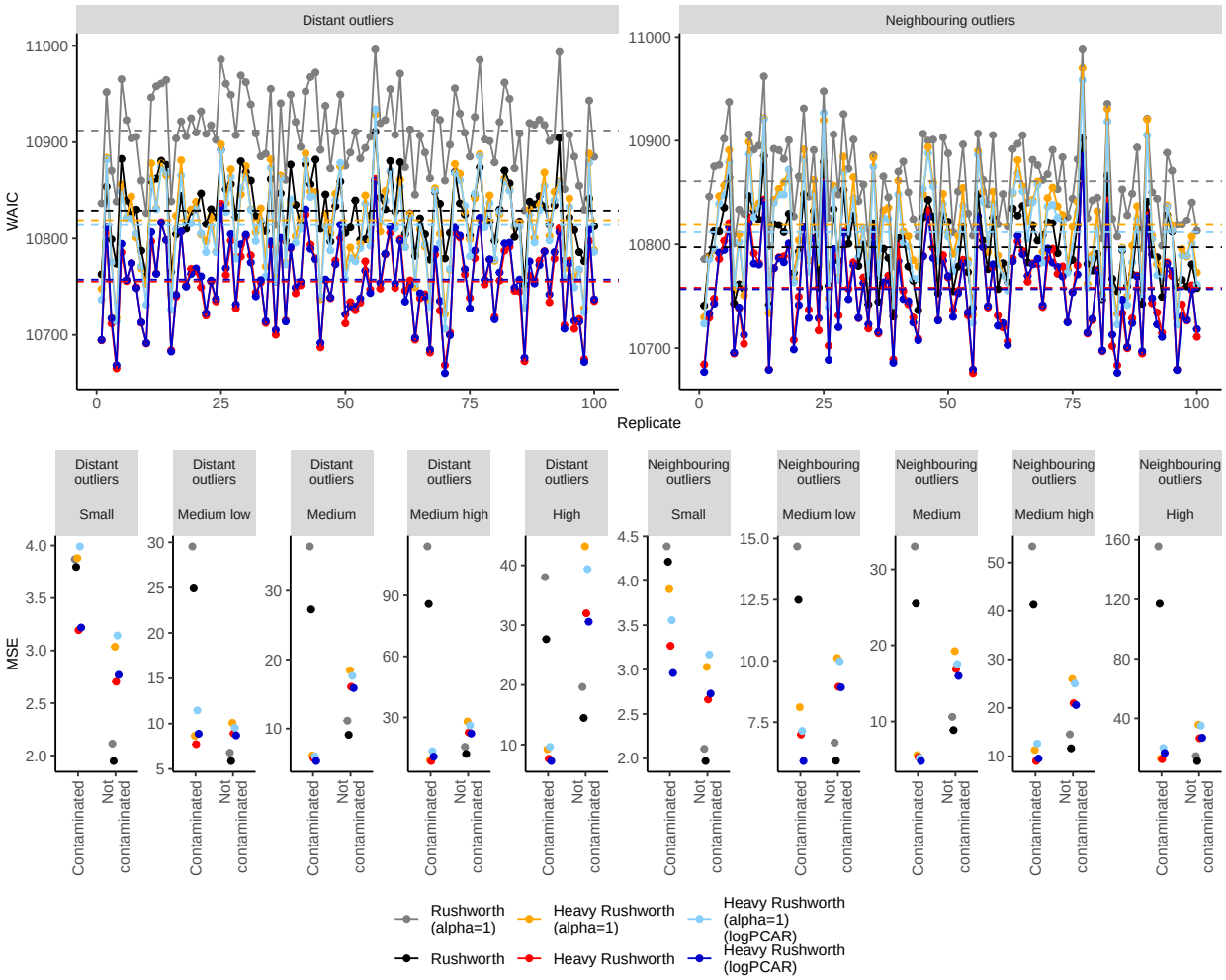


Figure B.5: Top row: WAIC across the 100 replicates for each model and each simulation scenario under the different fitted models. Dashed lines: mean WAIC across the 100 replicates. Bottom row: Average MSE over the 100 simulation replicates for each model, offset category and scenario. The results for the contaminated boroughs are distinguished from the non-contaminated ones.

B.4 French regions

Figure B.6 below displays the French map where the departments are coloured according to their region. To help discuss the results from the analysis of COVID-19 hospitalisations in France during the second wave, the same colours are used here, in Figure B.6, and in Figure 4.5 in Section 4.3.3.

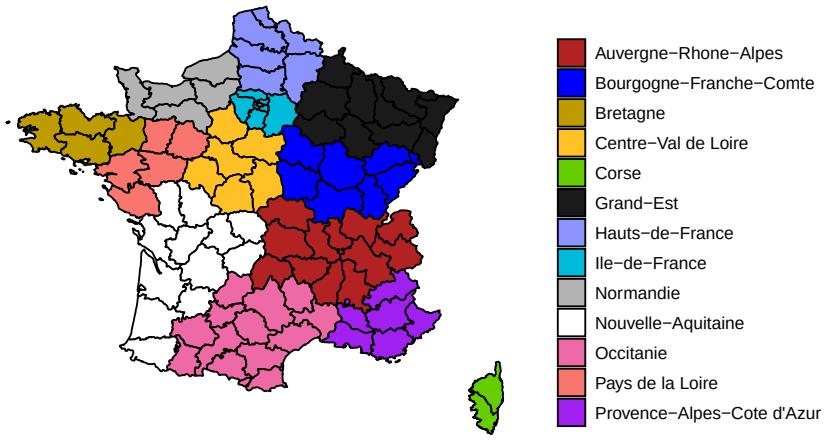


Figure B.6: Map of the French regions.

APPENDIX C

Appendix to Manuscript 3

C.1 Proof of the proposed scaled split conformal prediction interval coverage

In this section, we prove that the proposed scaled split conformal procedure described in Section 5.2.1 yields prediction intervals of the right coverage, when data come from a super population model of the form (5.4). To ease the notation, let $(x_i, y_i) \overset{indep.}{\sim} P_x \times P_{y|x,i},\ i = 1, \ldots, n, \text{new}$, where $P_{y|x,i}$ is such that $\mathbb{E}(y_i \mid x_i) = \mu(x_i)$ and $\mathbb{V}(y_i \mid x_i) = \sigma^2/c_i$, with c_i known for all $i = 1, \ldots, n, \text{new}$. The proposed procedure to compute the prediction interval for y_{new} is as follows:

1. Randomly split $\{(y_i, x_i),\ i = 1, \ldots, n\}$ into two equal sized datasets. Denote by S_1 and S_2 the resulting two sets of indices;
2. Train a model on $\{(y_i, x_i),\ i \in S_1\}$ and predict $\{\widehat{\mu}(x_i),\ i \in S_2\}$;
3. Compute the scaled absolute residuals $R_i = \sqrt{c_i} \times |y_i - \widehat{\mu}(x_i)|,\ i \in S_2$;
4. Find d_α, the k_αth smallest residual R, for $k_\alpha = \lceil (n/2 + 1)(1 - \alpha) \rceil$;

5. Let the prediction interval for y_{new} be $\text{PI}_{(1-\alpha)\%}[y_{\text{new}}] = \widehat{\mu}(x_{\text{new}}) \pm d_\alpha/\sqrt{c_{\text{new}}}$.

The coverage of the proposed prediction interval is computed as follows:

$$\begin{aligned}\text{P}\left(y_{\text{new}} \in [\widehat{\mu}(x_{\text{new}}) \pm d_\alpha/\sqrt{c_{\text{new}}}]\right) &= \text{P}\left(\sqrt{c_{\text{new}}}\,|y_{\text{new}} - \widehat{\mu}(x_{\text{new}})| \leq d_\alpha\right) \\ &= \text{P}\left(R_{\text{new}} \leq R_{\lceil (n/2+1)(1-\alpha)\rceil}\right),\end{aligned}$$

where the R_i's, $i \in S_2$, and R_{new} are i.i.d. (hence, exchangeable), and where $R_{\lceil (n/2+1)(1-\alpha)\rceil}$ denotes the $\lceil (n/2+1)(1-\alpha)\rceil$th smallest R. Additionally, note that for a_{new} and ordered $a_1, \ldots, a_n$ exchangeable, $\text{P}(a_{\text{new}} \leq a_k) = k/(n+1)$ (Angelopoulos and Bates, 2021). Therefore,

$$\begin{aligned}\text{P}\left(y_{\text{new}} \in [\widehat{\mu}(x_{\text{new}}) \pm d_\alpha/\sqrt{c_{\text{new}}}]\right) &= \text{P}\left(R_{\text{new}} \leq R_{\lceil (n/2+1)(1-\alpha)\rceil}\right) \\ &= \lceil (n/2+1)(1-\alpha)\rceil/(n/2+1) \geq 1-\alpha. \qquad \square\end{aligned}$$

C.2 Forward approach

The forward approach can be described in the following steps:

1. Variable selection:

   (a) Fit p simple linear models: $\overline{y}_c^{(s)} \sim \mathcal{N}\left(\eta_0^{(0)} + \overline{x}_{cj}^{(s)}\eta_1^{(0)}, {\sigma^{(0)}}^2/n_c\right),\ c = 1, \ldots, m,\ j = 1, \ldots, p$, and compute the p corresponding AICs;

   (b) Select $\overline{x}_k$ which minimises the AIC;

   (c) Fit $p-1$ linear models: $\overline{y}_c^{(s)} \sim \mathcal{N}\left(\overline{\boldsymbol{z}}_{(1)c}^{(s)\top}\boldsymbol{\eta}^{(1)}, {\sigma^{(1)}}^2/n_c\right),\ \overline{\boldsymbol{z}}_{(1)}^{(s)} = \left[1, \overline{x}_k^{(s)}, \overline{x}_j^{(s)}\right]^\top,\ c = 1, \ldots, m,\ j = 1, \ldots, p,\ j \neq k$, and compute the corresponding AIC;

   (d) Select $\overline{x}_{k'}$ which minimises the AIC;

   (e) Repeat steps (c) and (d) until the AIC is no longer minimised or while the number

of selected variables, $K < m$.

2. Final model and prediction:

   (a) Fit $\overline{y}_c^{(s)} \sim \mathcal{N}\left(\overline{\boldsymbol{z}}_c^{(s)\top}\boldsymbol{\eta}, \sigma^2/n_c\right),\ c = 1, \ldots, m$ to estimate $\widehat{\boldsymbol{\eta}}$, $\widehat{\mathbb{V}}\left(\widehat{\boldsymbol{\eta}}\right)$ and $\widehat{\sigma}$;

   (b) Predict $\widehat{\overline{Y}}_c^{(ns)} = \overline{\boldsymbol{z}}_c^{(ns)\top}\widehat{\boldsymbol{\eta}},\ c = 1, \ldots, M$;

   (c) Estimate the prediction variance, $\widehat{\mathbb{V}}\left(\widehat{\overline{Y}}_c^{(ns)}\right) = \overline{\boldsymbol{z}}_c^{(ns)\top}\widehat{\mathbb{V}}\left(\widehat{\boldsymbol{\eta}}\right)\overline{\boldsymbol{z}}_c^{(ns)} + \widehat{\sigma}^2/(N_c - n_c),\ c = 1, \ldots, M$.

C.3 Random Forest algorithm

A random forest procedure can be described through the following algorithm:

1. Draw a bootstrap dataset $\mathcal{D}^{(b)} = \left\{\left(\overline{y}_c^{(s)(b)}, \overline{\boldsymbol{x}}_c^{(s)(b)}\right),\ c = 1, \ldots, m\right\}$;

2. Train the bth tree $T^{(b)}$ using $\mathcal{D}^{(b)}$ with hyperparameters *mtry* and *nodesize*:

   (a) Let all responses gather in a single node, $\mathcal{A}$;

   (b) Randomly select $mtry \leq p$ covariates. Partition $\mathcal{A}$ into nodes $\mathcal{A}_1$ and $\mathcal{A}_2$ based on $\overline{x}_j \leq c$ and $\overline{x}_j > c$, respectively, for $\overline{x}_j$ one of the *mtry* selected covariates. The covariate and splitting rule are chosen such that

$$\sum_{a=1}^{2} \sum_{c \in \mathcal{A}_a} \left(\overline{y}_c^{(s)(b)} - \left((1/|\mathcal{A}_a|) \sum_{c' \in \mathcal{A}_a} \overline{y}_{c'}^{(s)(b)} \right) \right)^2$$

   is minimised.

   (c) Repeat step (b) until there are a maximum of *nodesize* responses in each final node.

   (d) For a new data point with covariate vector $\overline{\boldsymbol{x}}$, $T^{(b)}$ yields a point estimate equal to the mean responses in the final node that corresponds to $\overline{\boldsymbol{x}}$: $\widehat{\overline{Y}}^{(b)}(\overline{\boldsymbol{x}}) = \sum_{c=1}^{m} w_c^{(b)}(\overline{\boldsymbol{x}})\overline{y}_c^{(s)}$,

where $w_c^{(b)},\ c = 1, \ldots, m$, are weights associated to the outcome sampled means based on the bth bootstrap dataset and tree.

3. Repeat steps 1. and 2. B times.

4. For a new data point with covariate vector $\overline{\boldsymbol{x}}$, the random forest yields a point estimate equal to the average over the B point estimates obtained from the B trees: $\widehat{\overline{Y}} = (1/B)\sum_{b=1}^{B} \widehat{\overline{Y}}^{(b)}(\overline{\boldsymbol{x}}) = \sum_{c=1}^{m} w_c(\overline{\boldsymbol{x}})\overline{y}_c^{(s)}$, where $w_c = (1/B)\sum_{b=1}^{B} w_c^{(b)}(\overline{\boldsymbol{x}}),\ c = 1, \ldots, m$.

C.4 R code: proposed scaled split conformal procedure

Listing C.1: R code to obtain prediction intervals associated with random forest estimates through the proposed scaled split conformal procedure

```
# Sample_data: {(xbar_c^s, ybar_c^s), c=1, ..., m}
# eas: sampled areas
# Pop_data: {(xbar_c^ns, ybar_c^s), c=1, ..., M}, with ybar_c^s=0 if c is not sampled

# Step 1: split sample data and organise dataset
selected_eas_half = sample(eas, length(eas)/2, replace = FALSE)

data_to_train = Sample_data %>% filter(ea %in% selected_eas_half)
data_to_get_residuals = Sample_data %>% filter((!ea %in% selected_eas_half))
data_to_get_final_estimates = Pop_data

Full_Data = bind_rows(data_to_train,
                      data_to_get_residuals,
                      data_to_get_final_estimates)

# Step 2: train a RF on S1 and predict on S2
```

```
  rf = ranger(y ~ ., data = Full_Data[1:nrow(data_to_train),],
              mtry=2, min.node.size = 5, num.trees = 1000, keep.inbag = TRUE)

all_pred = predict(rf, data = Full_Data[(nrow(data_to_train)+1):nrow(Full_Data),])

  # Step 3: Compute the scaled absolute residuals
  residuals = Sample_data %>%
    filter((!ea %in% selected_eas_half)) %>%
    mutate(pred = (all_pred$predictions)[1:nrow(data_to_get_residuals)],
           R_c_scaled = abs(y - pred)*sqrt(n_c))

  # Step 4: Find d_alpha, the relevant quantile for a (1-alpha)% level prediction interval
  d_95 = sort(residuals$R_c_scaled)[ceiling((length(selected_eas_half) + 1)*(1 - 0.05))]
  d_80 = sort(residuals$R_c_scaled)[ceiling((length(selected_eas_half) + 1)*(1 - 0.2))]
  d_50 = sort(residuals$R_c_scaled)[ceiling((length(selected_eas_half) + 1)*(1 - 0.5))]

  # Step 5: Compute the (1-alpha)% level prediction intervals
  predictions = Pop_Data %>%
    mutate(pred =
      (all_pred$predictions)[(nrow(data_to_get_residuals)+1):length(Pop_Data)],
           y_bar_hat = f_c*y_bar_s + (1-f_c)*pred,

           CI_l_95 = f_c*y_bar_s + (1-f_c)*(pred - d_95/sqrt(N_c-n_c)),
           CI_u_95 = f_c*y_bar_s + (1-f_c)*(pred + d_95/sqrt(N_c-n_c)),

           CI_l_80 = f_c*y_bar_s + (1-f_c)*(pred - d_80/sqrt(N_c-n_c)),
           CI_u_80 = f_c*y_bar_s + (1-f_c)*(pred + d_80/sqrt(N_c-n_c)),
```

```
CI_l_50 = f_c*y_bar_s + (1-f_c)*(pred - d_50/sqrt(N_c-n_c)),
CI_u_50 = f_c*y_bar_s + (1-f_c)*(pred + d_50/sqrt(N_c-n_c)))
```

C.5 Design-based simulation study: scaled split conformal procedure

In this section, we present the design-based equivalent of the simulation study from Section 5.3.1. We create a single finite population of $M = 500$ areas of sizes $N_c,\ c = 1, \ldots, M$, with $\min_c(N_c) = 50$ and $\max_c(N_c) = 500$. For $c = 1, \ldots, M$, and $k = 1, \ldots, N_c$, the response variable has distribution

$$y_{ck} \overset{ind.}{\sim} \mathcal{N}(9.5 + x_{1,ck} - x_{2,ck} + 2x_{3,ck} - x_{4,ck} + 2x_{5,ck} + x_{6,ck}, 1),$$

with 6 unit-level covariates, $x_1, \ldots, x_6 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. From the finite population, $R = 500$ samples are drawn according to the same five sampling designs as in Section 5.3.1, which constitute the simulation scenarios:

1. (Stratified) Select all $m = M = 500$ areas and within each area, sample $n_c = 0.5N_c,\ c = 1, \ldots, m$ units;
2. (Stratified) Select all $m = M = 500$ areas and within each area, sample $n_c = 0.7N_c,\ c = 1, \ldots, m$ units;
3. (One-stage) Sample $m = M/2$ areas and within each area, select all $n_c = N_c,\ c = 1, \ldots, m$ units;
4. (Two-stage) Sample $m = M/2$ areas and within each area, sample $n_c = 0.5N_c,\ c = 1, \ldots, m$ units;

5. (Two-stage) Sample $m = M/2$ areas and within each area, sample $n_c = 0.7N_c$, $c = 1, \ldots, m$ units.

Again, for each simulation scenario and in each sample, the estimates described in equation (5.2) are computed using the same four methods as in Section 5.3.1: a linear model that includes the correct six covariates, a linear model that omits x_4, x_5 and x_6, a random forest method that considers all six covariates to grow the trees, and a linear LASSO model. Finally, for each scenario, in each sample and for each modelling method, 50%, 80% and 95% prediction intervals (5.3) are computed following the SC procedure and the proposed scaled SC procedure.

Figure C.1 summarises all the results obtained from this study. Similarly to the model-based simulation study shown in Section 5.3.1, when n_c and $N_c - n_c$ are equal, the data are exchangeable and both SC procedures yield the right coverages. However, when $n_c \neq N_c - n_c$, the data are not exchangeable and the original SC intervals show undercoverage. The proposed scaled SC procedure corrects this undercovage and produces prediction intervals of the right rate.

C.6 Design-based simulation study: prediction methods comparison

In this section, we present the design-based equivalent of the simulation study from Section 5.3.2. We create a single finite population of $M = 1000$ areas of sizes N_c with $\min_c(N_c) = 50$ and $\max_c(N_c) = 500$, according to 3 different models:

A. $y_{ck} \sim \mathcal{N}\left(20 + \boldsymbol{x}_{ck}^\top \boldsymbol{\beta}, 0.5^2\right)$, where the covariates are such that $\boldsymbol{x}_{ck} \sim \mathcal{N}_{100}(\boldsymbol{0}, \boldsymbol{I})$ and with coefficients $\boldsymbol{\beta}^\top = (1, -1, 2, -1, 2, 1, 2, 1, -1, 1, 0, \ldots, 0)$;

B. $y_{ck} \sim \mathcal{N}\left(20 + \boldsymbol{x}_{ck}^\top \boldsymbol{\beta}, 0.5^2\right)$, where the covariates are such that $\boldsymbol{x}_{ck} \sim \mathcal{N}_{100}(\boldsymbol{0}, \Sigma_x)$, with

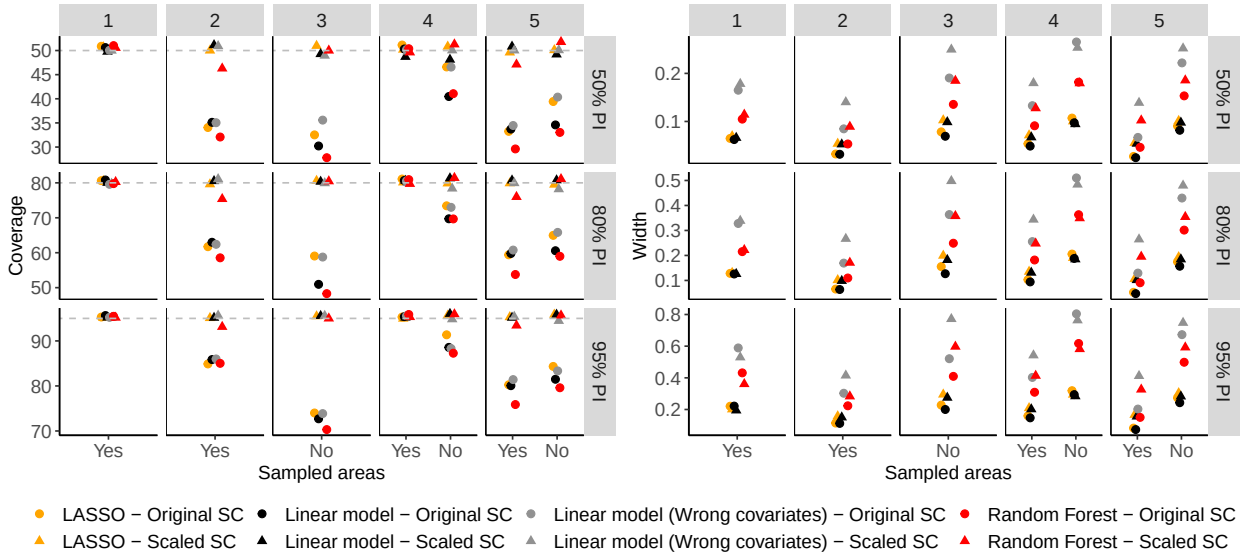


Figure C.1: Coverages and widths of the prediction intervals (PI) obtained from the proposed scaled and original split conformal (SC) procedures for the four modelling methods and across the five scenarios (1-5) in the design-based simulation study. Yes: coverages and widths across the sampled areas; No: coverages and widths across the non-sampled areas.

$$\Sigma_x = \begin{bmatrix} 1 & 0.5 & \dots & 0.5 \\ 0.5 & 1 & \dots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & \dots & 1 \end{bmatrix}, \text{ and } \boldsymbol{\beta}^\top = (1, -1, 2, -1, 2, 1, 2, 1, -1, 1, 0, \dots, 0)/10;$$

C. $y_{ck} \sim \mathcal{N}\left(x_{1,ck}^2 + \exp\left(x_{2,ck}^2\right), 0.3\right)$, with covariates $x_{j,c} \sim \mathcal{U}(-1, 1)$, $j = 1, \dots, 100$.

From each finite population, $R = 100$ samples are drawn following the two sampling schemes:

1. (Stratified) Select all $m = M = 500$ areas and within each area, sample $n_c = 15$, $c = 1, \dots, m$ units;

2. (Two-stage) Sample $m = M/2$ areas and within each area, sample $n_c = 15$, $c = 1, \dots, m$ units.

Like in Section 5.3.2, for each scenario, the estimates and their prediction intervals are computed as described in Section 5.2, assuming known and anonymised EAs..

Figure C.2 shows that the covariate selection pattern is similar to Section 5.3.2. For instance,

the forward and LASSO approach always select the right covariates in Populations A and B, while the random forest always select the correct ones in Population C. Finally, in this design-based setting, the results regarding bias, MSE, coverage and proper interval score of the prediction intervals are similar to the ones shown in Section 5.3.2, in the model-based framework. All methods are virtually unbiased with mean absolute biases between 0 and 0.8, regardless of the population and sampling design. The LASSO, forward and Bayesian approaches yield identical MSEs in all scenarios, while the random forest method outperforms the three in population C. In scenarios A and B, all methods lead to prediction intervals with the right coverage and equivalent proper interval scores. Finally, in population C, all methods yielded under-coverage, the random forest leading to slightly higher rates and smaller proper interval scores.

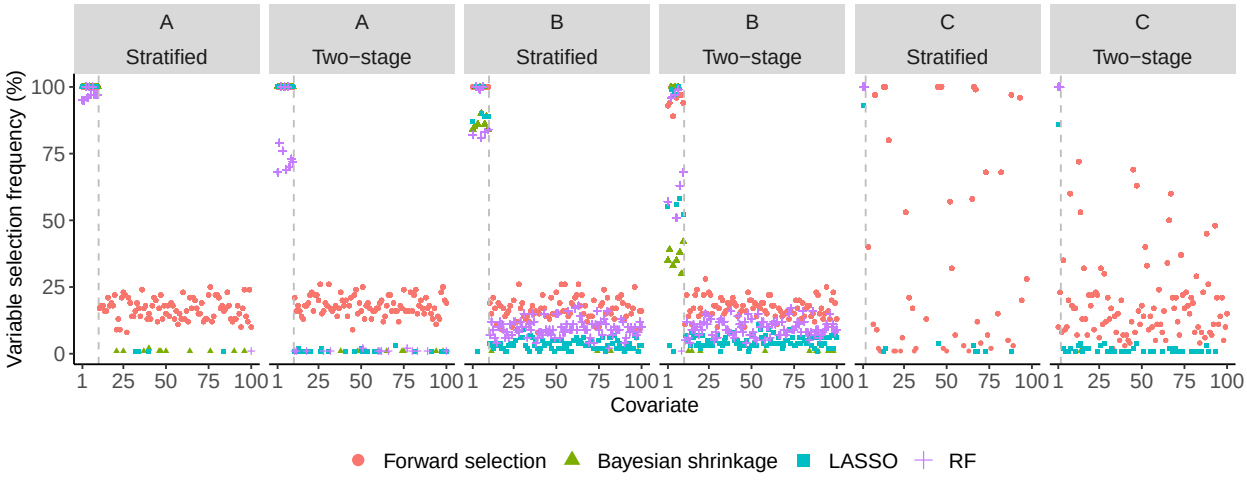


Figure C.2: Covariate selection frequency for each method across the 6 simulation scenarios. Left of the vertical dashed line: true covariates used in the generating models.

C.7 Extra design-based simulation scenarios: prediction methods comparison

This section is a continuation of the design-based simulation study shown in Section C.6. The same finite populations A, B and C are created and, from each, $R = 100$ samples are

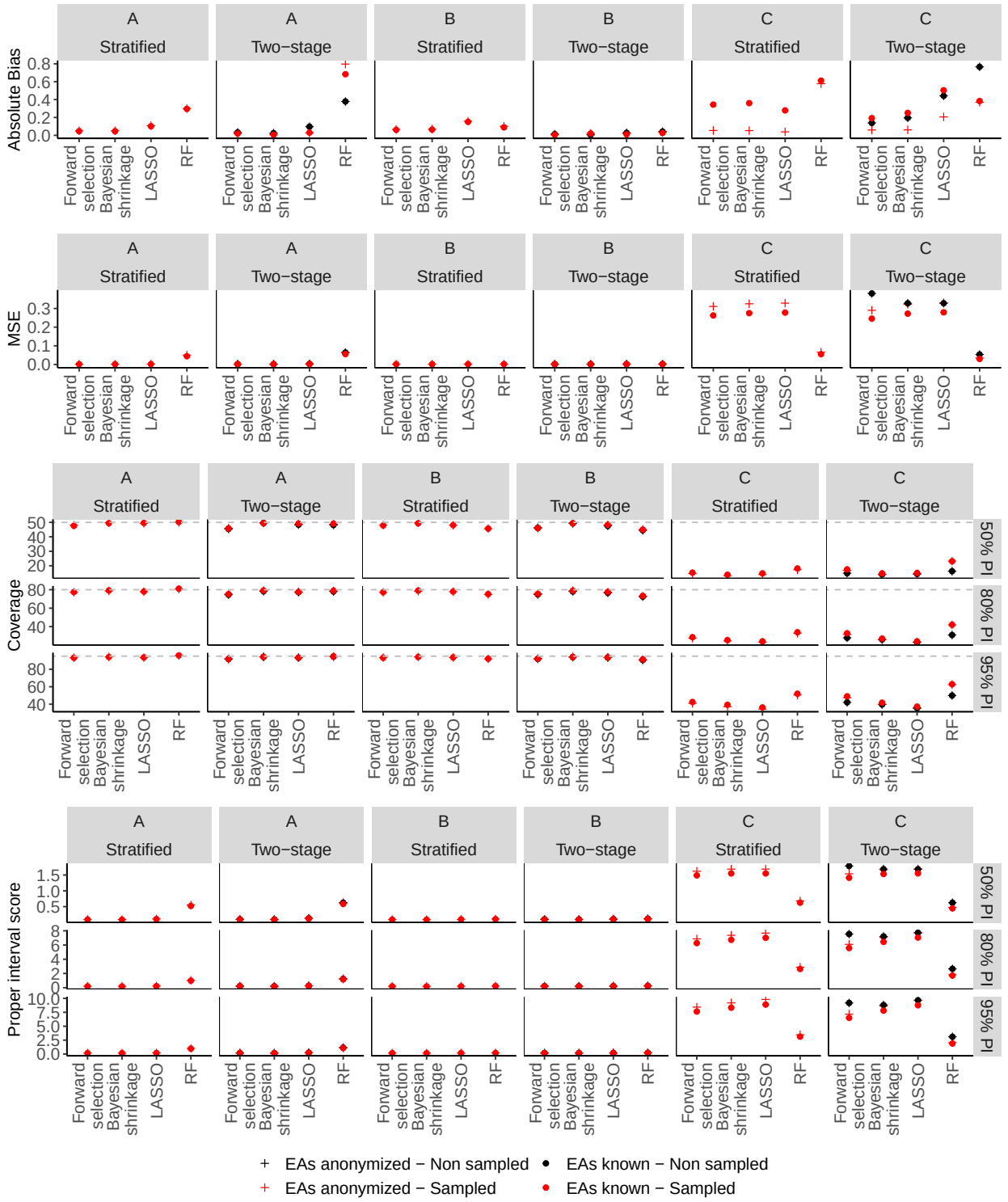


Figure C.3: Mean absolute bias, MSE, coverages and proper scores of the prediction intervals, obtained for each method across the 6 simulation scenarios. RF: Random forest approach.

drawn following the three sampling schemes:

1. (Stratified) Select all $m = M = 500$ areas (EAs) and within each area, sample $n_c = 0.5N_c,\ c = 1, \ldots, m$ units;

2. (One-stage) Sample $m = M/2$ areas and within each area, select all $n_c = N_c,\ c = 1, \ldots, m$ units;

3. (Two-stage) Sample $m = M/2$ areas and within each area, sample $n_c = 0.5N_c,\ c = 1, \ldots, m$ units.

Again, for each scenario, the estimates and their prediction intervals are computed as described in Section 5.2, knowing and anonymising the EAs.

Figure C.4 presents the results (mean absolute bias, MSE, prediction interval coverage and proper score) for each of the 9 simulation scenarios. Like the results in Section 5.3.2 and Appendix C.6, all four modelling methods perform similarly in populations A and B, where the association between the outcome and the covariates is linear. All four methods yield virtually no bias and prediction intervals of the right coverage rate.

In population C, when the sampled EAs are anonymised, all four methods perform slightly worse than when the sampling information is known: the MSE is multiplied by a factor of 3 and the prediction intervals show under-coverage. When the EAs are anonymised, in all three sampling schemes, the random forest approach leads to smaller MSE and proper interval scores. In terms of MSE, regardless of the sampling design, the random forest performs better than the other three modelling methods, knowing or ignoring which EAs have been sampled.

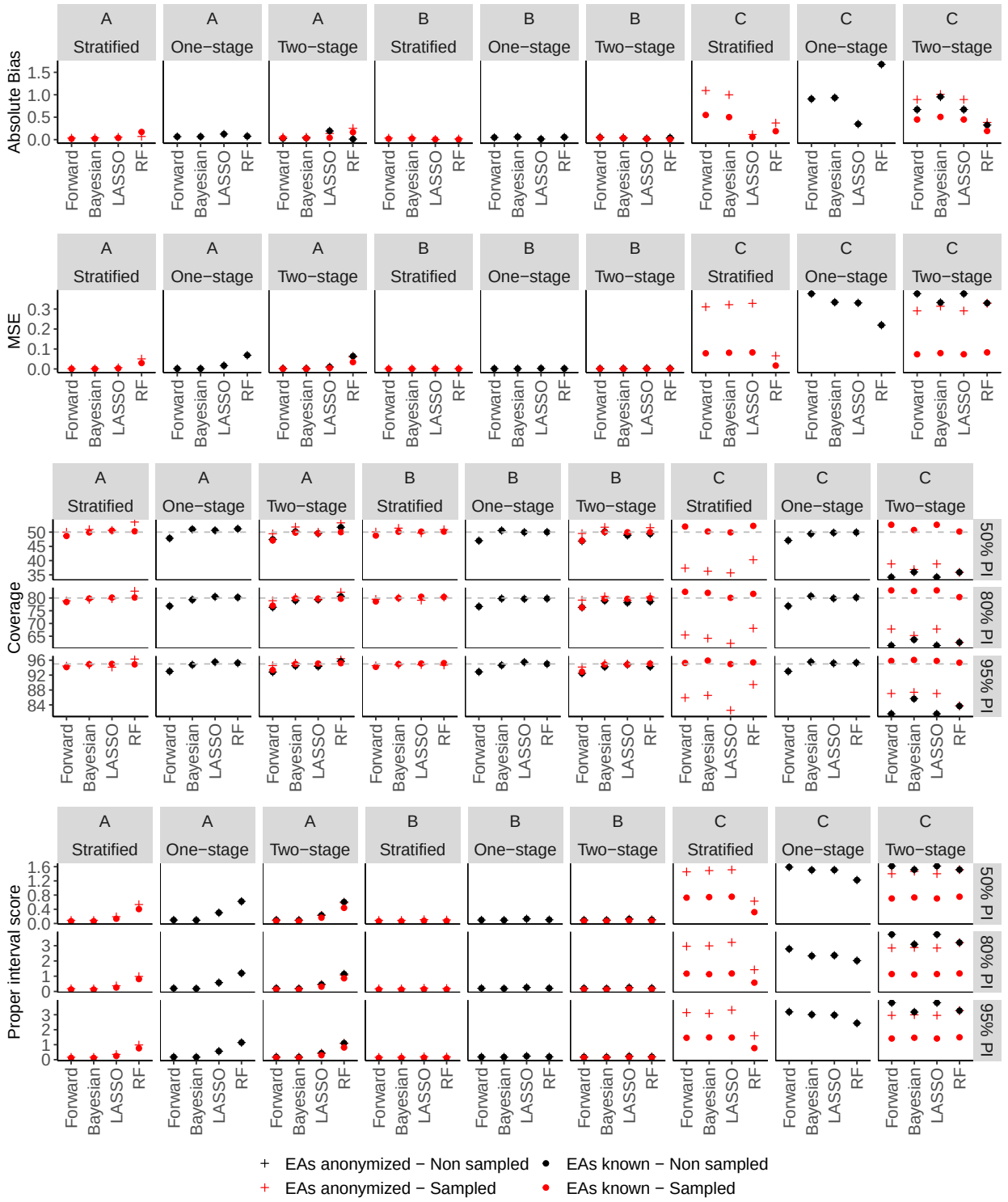


Figure C.4: Mean absolute bias, MSE, coverages and proper scores of the prediction intervals, obtained for each method across the 9 simulation scenarios. Forward: forward selection approach; Bayesian: Bayesian shrinkage approach; RF: Random forest approach

C.8 Model-based simulation study using the Ghanaian data

In this section, we present the results from a model-based simulation study that is similar to the one shown in 5.3.2. The difference is that we make use of the available Ghanaian census data from Section 5.4, in order to consider a realistic set of auxiliary variables. Recall that there are $M = 5019$ EAs in the GAMA and $p = 174$ available variables. In this model-based framework, $R = 100$ finite populations are created following two scenarios, which correspond to a linear and a non-linear relationship between the outcome and covariates. In the linear case, at the unit level (i.e., the household level), the response variable is distributed according to $y_{ck} \overset{i.i.d.}{\sim} \mathcal{N}(9.3+\boldsymbol{x}_{ck}^{\top}\boldsymbol{\beta}, 0.76^2)$, where the intercept and standard deviation values, as well as the 9 variables with non-zero coefficients (summarised in Table C.1 below) were fixed based on the results obtained in Section 5.4. In the non-linear case, we consider two covariates x_{1k} and x_{2k} which are the number of rooms in the kth household dwelling and the number of household members of native nationality, respectively. The response variable is distributed at the unit level as $y_{ck} \overset{i.i.d.}{\sim} \mathcal{N}\left(4 + \mathbb{1}_{[x_{1k}\leq 8]}(x_{1k}/5)^2 + \mathbb{1}_{[x_{2k}\leq 7]}(x_{2k}/5)^2, 0.3^2\right)$. Then, within each finite population, the same set of households are sampled according to the same stratified two-stage sampling design that lead to the GLSS dataset studied in Section 5.4. The strata correspond to the urban and rural indicator and $m = 136$ EAs are sampled with a proportional to size design, wherein 8 EAs are rural ones. Within the sampled EAs, $n_c = 15$ households are systematically sampled.

Covariate	Coefficient
Household head age	-0.02
Nationality: native	0.04
Rooms	0.11
Interweb	0.12
Water: pipe-borne outside dwelling	0.08
Cooking: fuel, gas	0.29
Rubbish: collected	0.15
Floor: cement, concrete	-0.24
Water: sachet, bottled	0.09

Table C.1: Ghanaian covariates and their corresponding coefficient for the model-based simulation study with linear relationship.

Similarly to Section 5.3.2, for each scenario, the estimates and their prediction intervals are computed as described in Section 5.2. Again, the modelling approaches are run knowing and ignoring which EAs have been sampled. Figure C.5 summarises the performance measures obtained for each approach in both simulation scenarios. In terms of bias, similar to Section 5.3.2, all methods were virtually unbiased with absolute mean biases smaller than 1. Interestingly, in terms of MSE, there was no difference between the linear and non-linear populations: the Bayesian shrinkage approach always resulted in smaller MSEs than the other three methods. Finally, regardless of the scenario, the forward selection approach resulted in under-coverage for the prediction intervals, which was not the case for the other three methods. This was also the case in the application shown in Section 5.4. We find it noteworthy that even though they are based on $m/2 = 68$ data points, the prediction intervals computed based on our proposed scaled split conformal approach resulted in correct coverage rates.

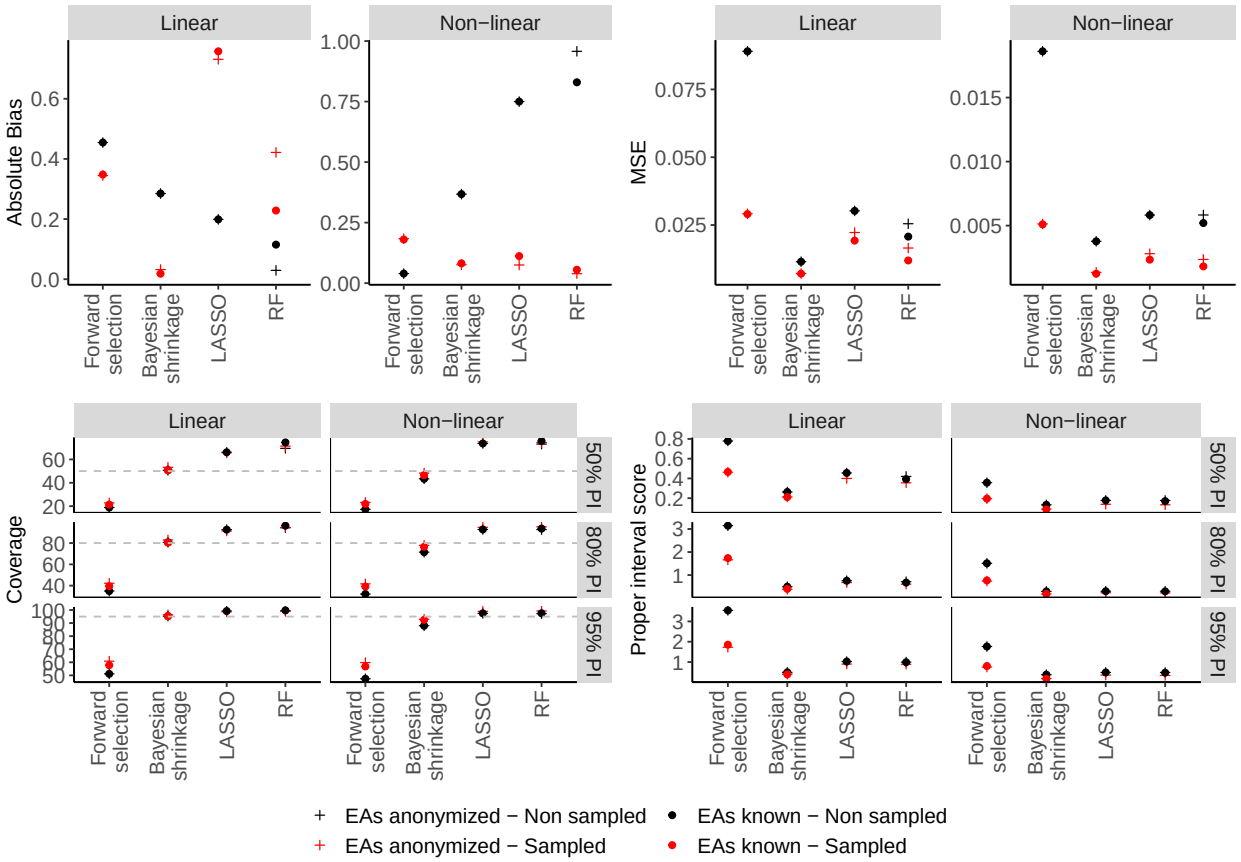


Figure C.5: Mean absolute bias, MSE, coverages and proper scores of the prediction intervals, obtained for each method across the 2 simulation scenarios conducted using the Ghanaian auxiliary information. RF: Random forest approach.

C.9 Detailed results for the model-based simulation study summarised in Section 5.3.2

Scenario		EAs	Forward selection	Bayesian shrinkage	LASSO	RF
A	Stratified	EAs anonymised - Sampled	0.000	0.000	0.001	0.054
A	Stratified	EAs known - Sampled	0.000	0.000	0.001	0.039
A	Two-stage	EAs anonymised - Non sampled	0.000	0.000	0.004	0.064
A	Two-stage	EAs anonymised - Sampled	0.000	0.000	0.000	0.085
A	Two-stage	EAs known - Non sampled	0.000	0.000	0.004	0.061
A	Two-stage	EAs known - Sampled	0.000	0.000	0.000	0.069
B	Stratified	EAs anonymised - Sampled	0.000	0.000	0.000	0.005
B	Stratified	EAs known - Sampled	0.000	0.000	0.000	0.007
B	Two-stage	EAs anonymised - Non sampled	0.000	0.001	0.003	0.007
B	Two-stage	EAs anonymised - Sampled	0.000	0.001	0.004	0.008
B	Two-stage	EAs known - Non sampled	0.000	0.001	0.003	0.005
B	Two-stage	EAs known - Sampled	0.000	0.001	0.004	0.005
C	Stratified	EAs anonymised - Sampled	0.000	0.000	0.003	0.010
C	Stratified	EAs known - Sampled	0.003	0.003	0.005	0.005
C	Two-stage	EAs anonymised - Non sampled	0.020	0.036	0.037	0.019
C	Two-stage	EAs anonymised - Sampled	0.001	0.001	0.001	0.006
C	Two-stage	EAs known - Non sampled	0.020	0.036	0.037	0.015
C	Two-stage	EAs known - Sampled	0.004	0.004	0.003	0.004

Table C.2: Mean absolute bias obtained for each method across the 6 model-based simulation scenarios, knowing and ignoring which EAs have been sampled. RF: Random forest approach.

Scenario		EAs	Forward selection	Bayesian shrinkage	LASSO	RF
A	Stratified	EAs anonymised - Sampled	0.001	0.001	0.002	0.045
A	Stratified	EAs known - Sampled	0.001	0.001	0.002	0.029
A	Two-stage	EAs anonymised - Non sampled	0.001	0.001	0.003	0.061
A	Two-stage	EAs anonymised - Sampled	0.002	0.001	0.003	0.066
A	Two-stage	EAs known - Non sampled	0.001	0.001	0.003	0.041
A	Two-stage	EAs known - Sampled	0.001	0.001	0.002	0.037
B	Stratified	EAs anonymised - Sampled	0.001	0.001	0.002	0.002
B	Stratified	EAs known - Sampled	0.001	0.001	0.001	0.002
B	Two-stage	EAs anonymised - Non sampled	0.001	0.001	0.002	0.002
B	Two-stage	EAs anonymised - Sampled	0.002	0.001	0.002	0.002
B	Two-stage	EAs known - Non sampled	0.001	0.001	0.002	0.002
B	Two-stage	EAs known - Sampled	0.001	0.001	0.002	0.002
C	Stratified	EAs anonymised - Sampled	0.308	0.323	0.326	0.038
C	Stratified	EAs known - Sampled	0.261	0.273	0.276	0.011
C	Two-stage	EAs anonymised - Non sampled	0.363	0.314	0.314	0.087
C	Two-stage	EAs anonymised - Sampled	0.305	0.333	0.340	0.060
C	Two-stage	EAs known - Non sampled	0.363	0.313	0.314	0.057
C	Two-stage	EAs known - Sampled	0.260	0.283	0.289	0.014

Table C.3: MSE obtained for each method across the 6 model-based simulation scenarios, knowing and ignoring which EAs have been sampled. RF: Random forest approach.

Scenario		EAs	Forward selection	Bayesian shrinkage	LASSO	RF
A	Stratified	EAs anonymised - Sampled	48.2	49.9	48.5	45.7
A	Stratified	EAs known - Sampled	47.8	49.7	48.7	54.5
A	Two-stage	EAs anonymised - Non sampled	46.3	49.8	48.1	40.7
A	Two-stage	EAs anonymised - Sampled	46.3	50.0	48.3	40.3
A	Two-stage	EAs known - Non sampled	46.3	49.9	48.1	48.9
A	Two-stage	EAs known - Sampled	46.1	49.7	48.1	49.0
B	Stratified	EAs anonymised - Sampled	48.3	49.9	49.6	45.8
B	Stratified	EAs known - Sampled	48.1	49.8	49.5	47.3
B	Two-stage	EAs anonymised - Non sampled	46.6	50.0	48.7	43.8
B	Two-stage	EAs anonymised - Sampled	46.9	50.1	48.7	43.8
B	Two-stage	EAs known - Non sampled	46.6	50.1	48.7	46.2
B	Two-stage	EAs known - Sampled	46.4	49.6	48.4	46.3
C	Stratified	EAs anonymised - Sampled	14.8	13.3	14.0	22.8
C	Stratified	EAs known - Sampled	15.4	14.0	14.8	33.1
C	Two-stage	EAs anonymised - Non sampled	12.5	12.8	13.1	14.5
C	Two-stage	EAs anonymised - Sampled	17.3	15.2	15.2	23.4
C	Two-stage	EAs known - Non sampled	12.5	12.9	13.1	18.6
C	Two-stage	EAs known - Sampled	17.9	15.9	15.9	36.1

Table C.4: Coverages of the 50% prediction intervals obtained for each method across the 6 model-based simulation scenarios, knowing and ignoring which EAs have been sampled. RF: Random forest approach.

Scenario		EAs	Forward selection	Bayesian shrinkage	LASSO	RF
A	Stratified	EAs anonymised - Sampled	78.3	80.0	78.6	74.1
A	Stratified	EAs known - Sampled	78.0	79.8	78.7	82.2
A	Two-stage	EAs anonymised - Non sampled	75.9	80.1	78.2	68.7
A	Two-stage	EAs anonymised - Sampled	76.3	80.1	77.5	68.1
A	Two-stage	EAs known - Non sampled	75.9	80.0	78.2	79.5
A	Two-stage	EAs known - Sampled	75.6	79.8	77.5	78.9
B	Stratified	EAs anonymised - Sampled	78.5	80.1	79.5	75.4
B	Stratified	EAs known - Sampled	78.3	80.0	79.5	77.4
B	Two-stage	EAs anonymised - Non sampled	76.3	80.1	78.6	72.8
B	Two-stage	EAs anonymised - Sampled	76.4	80.2	78.7	73.0
B	Two-stage	EAs known - Non sampled	76.3	80.1	78.6	75.9
B	Two-stage	EAs known - Sampled	75.6	79.7	78.4	76.1
C	Stratified	EAs anonymised - Sampled	27.5	24.6	22.9	40.5
C	Stratified	EAs known - Sampled	28.6	25.8	24.1	57.0
C	Two-stage	EAs anonymised - Non sampled	25.6	25.0	22.3	27.4
C	Two-stage	EAs anonymised - Sampled	32.6	26.6	23.9	41.2
C	Two-stage	EAs known - Non sampled	25.6	25.0	22.3	33.9
C	Two-stage	EAs known - Sampled	33.7	27.5	24.8	60.6

Table C.5: Coverages of the 80% prediction intervals obtained for each method across the 6 model-based simulation scenarios, knowing and ignoring which EAs have been sampled. RF: Random forest approach.

Scenario		EAs	Forward selection	Bayesian shrinkage	LASSO	RF
A	Stratified	EAs anonymised - Sampled	94.1	95.0	94.2	92.0
A	Stratified	EAs known - Sampled	94.0	94.9	94.3	96.6
A	Two-stage	EAs anonymised - Non sampled	92.9	95.2	94.5	90.8
A	Two-stage	EAs anonymised - Sampled	93.0	95.2	94.3	89.8
A	Two-stage	EAs known - Non sampled	92.9	95.1	94.5	96.8
A	Two-stage	EAs known - Sampled	92.5	94.9	94.1	96.1
B	Stratified	EAs anonymised - Sampled	94.2	95.0	94.8	92.2
B	Stratified	EAs known - Sampled	94.1	95.0	94.8	93.5
B	Two-stage	EAs anonymised - Non sampled	93.0	95.1	94.8	90.9
B	Two-stage	EAs anonymised - Sampled	93.0	95.1	94.7	91.2
B	Two-stage	EAs known - Non sampled	93.0	95.1	94.8	93.0
B	Two-stage	EAs known - Sampled	92.7	94.9	94.6	93.0
C	Stratified	EAs anonymised - Sampled	41.0	37.5	34.4	58.5
C	Stratified	EAs known - Sampled	42.7	39.3	36.1	77.6
C	Two-stage	EAs anonymised - Non sampled	40.7	39.5	36.5	44.4
C	Two-stage	EAs anonymised - Sampled	48.8	38.8	35.8	60.3
C	Two-stage	EAs known - Non sampled	40.7	39.5	36.5	53.4
C	Two-stage	EAs known - Sampled	50.1	40.4	37.6	82.1

Table C.6: Coverages of the 95% prediction intervals obtained for each method across the 6 model-based simulation scenarios, knowing and ignoring which EAs have been sampled. RF: Random forest approach.

Scenario		EAs	Forward selection	Bayesian shrinkage	LASSO	RF
A	Stratified	EAs anonymised - Sampled	0.178	0.170	0.209	0.925
A	Stratified	EAs known - Sampled	0.171	0.162	0.196	0.691
A	Two-stage	EAs anonymised - Non sampled	0.193	0.170	0.255	1.142
A	Two-stage	EAs anonymised - Sampled	0.198	0.174	0.254	1.153
A	Two-stage	EAs known - Non sampled	0.193	0.170	0.255	0.770
A	Two-stage	EAs known - Sampled	0.194	0.169	0.242	0.759
B	Stratified	EAs anonymised - Sampled	0.178	0.170	0.182	0.198
B	Stratified	EAs known - Sampled	0.171	0.162	0.173	0.170
B	Two-stage	EAs anonymised - Non sampled	0.190	0.176	0.205	0.213
B	Two-stage	EAs anonymised - Sampled	0.194	0.179	0.212	0.217
B	Two-stage	EAs known - Non sampled	0.190	0.176	0.205	0.197
B	Two-stage	EAs known - Sampled	0.190	0.174	0.204	0.190
C	Stratified	EAs anonymised - Sampled	6.844	7.347	7.701	2.049
C	Stratified	EAs known - Sampled	6.257	6.717	7.043	0.845
C	Two-stage	EAs anonymised - Non sampled	7.648	7.421	7.879	3.796
C	Two-stage	EAs anonymised - Sampled	6.301	7.345	7.863	2.540
C	Two-stage	EAs known - Non sampled	7.648	7.421	7.879	2.672
C	Two-stage	EAs known - Sampled	5.799	6.745	7.220	0.876

Table C.7: Proper interval scores of the 50% prediction intervals obtained for each method across the 6 model-based simulation scenarios, knowing and ignoring which EAs have been sampled. RF: Random forest approach.

Scenario		EAs	Forward selection	Bayesian shrinkage	LASSO	RF
A	Stratified	EAs anonymised - Sampled	0.178	0.170	0.209	0.925
A	Stratified	EAs known - Sampled	0.171	0.162	0.196	0.691
A	Two-stage	EAs anonymised - Non sampled	0.193	0.170	0.255	1.142
A	Two-stage	EAs anonymised - Sampled	0.198	0.174	0.254	1.153
A	Two-stage	EAs known - Non sampled	0.193	0.170	0.255	0.770
A	Two-stage	EAs known - Sampled	0.194	0.169	0.242	0.759
B	Stratified	EAs anonymised - Sampled	0.178	0.170	0.182	0.198
B	Stratified	EAs known - Sampled	0.171	0.162	0.173	0.170
B	Two-stage	EAs anonymised - Non sampled	0.190	0.176	0.205	0.213
B	Two-stage	EAs anonymised - Sampled	0.194	0.179	0.212	0.217
B	Two-stage	EAs known - Non sampled	0.190	0.176	0.205	0.197
B	Two-stage	EAs known - Sampled	0.190	0.174	0.204	0.190
C	Stratified	EAs anonymised - Sampled	6.844	7.347	7.701	2.049
C	Stratified	EAs known - Sampled	6.257	6.717	7.043	0.845
C	Two-stage	EAs anonymised - Non sampled	7.648	7.421	7.879	3.796
C	Two-stage	EAs anonymised - Sampled	6.301	7.345	7.863	2.540
C	Two-stage	EAs known - Non sampled	7.648	7.421	7.879	2.672
C	Two-stage	EAs known - Sampled	5.799	6.745	7.220	0.876

Table C.8: Proper interval scores of the 80% prediction intervals obtained for each method across the 6 model-based simulation scenarios, knowing and ignoring which EAs have been sampled. RF: Random forest approach.

Scenario		EAs	Forward selection	Bayesian shrinkage	LASSO	RF
A	Stratified	EAs anonymised - Sampled	0.166	0.162	0.198	1.002
A	Stratified	EAs known - Sampled	0.159	0.155	0.188	0.803
A	Two-stage	EAs anonymised - Non sampled	0.172	0.162	0.225	1.162
A	Two-stage	EAs anonymised - Sampled	0.177	0.167	0.234	1.245
A	Two-stage	EAs known - Non sampled	0.172	0.162	0.225	0.919
A	Two-stage	EAs known - Sampled	0.171	0.160	0.223	0.907
B	Stratified	EAs anonymised - Sampled	0.165	0.163	0.173	0.197
B	Stratified	EAs known - Sampled	0.158	0.155	0.165	0.178
B	Two-stage	EAs anonymised - Non sampled	0.172	0.164	0.181	0.212
B	Two-stage	EAs anonymised - Sampled	0.177	0.170	0.187	0.217
B	Two-stage	EAs known - Non sampled	0.172	0.164	0.181	0.195
B	Two-stage	EAs known - Sampled	0.171	0.163	0.179	0.190
C	Stratified	EAs anonymised - Sampled	8.427	9.212	9.887	2.331
C	Stratified	EAs known - Sampled	7.633	8.352	8.965	0.782
C	Two-stage	EAs anonymised - Non sampled	8.990	8.665	9.298	4.322
C	Two-stage	EAs anonymised - Sampled	7.605	9.040	9.842	2.910
C	Two-stage	EAs known - Non sampled	8.990	8.664	9.298	2.890
C	Two-stage	EAs known - Sampled	6.927	8.212	8.934	0.805

Table C.9: Proper interval scores of the 95% prediction intervals obtained for each method across the 6 model-based simulation scenarios, knowing and ignoring which EAs have been sampled. RF: Random forest approach.

References

Adin, A., Congdon, P., Santafé, G., and Ugarte, M. D. (2022). Identifying extreme COVID-19 mortality risks in English small areas: a disease cluster approach. *Stochastic Environmental Research and Risk Assessment*, pages 1–16.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Anderson, C., Lee, D., and Dean, N. (2014). Identifying clusters in Bayesian disease mapping. *Biostatistics*, 15(3):457–469.

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Bedi, T., Coudouel, A., and Simler, K. (2007). *More than a pretty picture: using poverty maps to design better policies and interventions*. World Bank Publications.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995). Bayesian analysis of space—time variation in disease risk. *Statistics in Medicine*, 14(21-22):2433–2443.

Bersson, E. and Hoff, P. D. (2022). Optimal conformal prediction for small areas. *arXiv preprint arXiv:2204.08122*.

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

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Best, N. G., Arnold, R. A., Thomas, A., Waller, L. A., and Conlon, E. M. (1999). Bayesian models for spatially correlated disease and exposure data. *Bayesian Statistics*, 6:131–156.

Breidt, J. F. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brook, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51(3):481–483.

Carabali, M., Harper, S., Neto, A. S. L., de Sousa, G. S., Caprara, A., Restrepo, B. N., and Kaufman, J. S. (2020). Spatiotemporal distribution and socioeconomic disparities of dengue, chikungunya and Zika in two Latin American cities from 2007 to 2017. *Tropical Medicine & International Health*, 26(3):301–315.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Census Bureau (2018). Small area income and poverty estimates (SAIPE) program.

Clayton, D. G. (1996). Generalized linear mixed models. *Markov chain Monte Carlo in practice*, 1:275–302.

Congdon, P. (2017). Representing spatial dependence and spatial discontinuity in ecological epidemiology: a scale mixture approach. *Stochastic Environmental Research and Risk Assessment*, 31(2):291–304.

Cooke, E., Hague, S., and McKay, A. (2016). The Ghana poverty and inequality report: Using the 6th Ghana Living Standards Survey. https://www.unicef.org/ghana/sites/unicef.org.ghana/files/2019-04/Ghana_Poverty_and_Inequality_Analysis_FINAL_3_2016_0_0.pdf, Accessed February 4th 2023.

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(3-4):531–544.

Corral, P., Molina, I., Cojocaru, A., and Segovia, S. (2022). *Guidelines to small area estimation for poverty mapping*. World Bank Publications.

Costemalle, V., Gaini, M., Hazo, J.-B., and Naouri, D. (2021). En quatre vagues, l'épidémie de COVID-19 a causé 116 000 décès et lourdement affecté le système de soins. *INSEE*. https://www.insee.fr/fr/statistiques/5432509?sommaire=5435421#onglet-1.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Dagdoug, M., Goga, C., and Haziza, D. (2023). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542):1234–1251.

Datta, G. S. and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics*, pages 1748–1770.

Datta, J. and Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, 8:111–132.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Temple Lang, D., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26:403–417.

Dean, C., Ugarte, M., and Militino, A. (2001). Detecting interaction between random region and fixed age effects in disease mapping. *Biometrics*, 57(1):197–202.

Dean, N., Dong, G., Piekut, A., and Pryce, G. (2019). Frontiers in residential segregation: Understanding neighbourhood boundaries and their impacts. *Tijdschrift voor economische en sociale geografie*, 110(3):271–288.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and R-software hdi. *Statistical Science*, 30:533–558.

Dong, T. Q. and Wakefield, J. (2021). Modeling and presentation of vaccination coverage estimates using data from household surveys. *Vaccine*, 39(18):2584–2594.

dos Santos, J. P. C., Honório, N. A., and Nobre, A. A. (2019). Definition of persistent areas with increased dengue risk by detecting clusters in populations with differing mobility and immunity in Rio de Janeiro, Brazil. *Cadernos de Saúde Pública*, 35(12).

Dunn, R., Wasserman, L., and Ramdas, A. (2018). Distribution-free prediction sets with random effects. *arXiv preprint arXiv:1809.07441*.

Dupont, E., Wood, S. N., and Augustin, N. H. (2022). Spatial+: a novel approach to spatial confounding. *Biometrics*, 78(4):1279–1290.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2002). *Micro-level estimation of welfare*, volume 2911. World Bank Publications.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1):355–364.

Erciulescu, A. L. and Opsomer, J. D. (2022). A model-based approach to predict employee compensation components. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(5):1503–1520.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277.

Fonseca, T. C., Lobo, V. G., and Schmidt, A. M. (2023). Dynamical non-Gaussian modelling of spatial processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 72(1):76–103.

Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., and Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. *Science of the Total Environment*, 739:140033.

Freitas, L. P., Cruz, O. G., Lowe, R., and Carvalho, M. S. (2019). Space–time dynamics of a triple epidemic: dengue, chikungunya and Zika clusters in the city of Rio de Janeiro. *Proceedings of the Royal Society B*, 286(1912):20191867.

Freitas, L. P., Schmidt, A. M., Cossich, W., Cruz, O. G., and Carvalho, M. S. (2021). Spatio-temporal modelling of the first chikungunya epidemic in an intra-urban setting: The role of socioeconomic status, environment and temperature. *PLOS Neglected Tropical Diseases*, 15(6):e0009537.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2004). *Bayesian data analysis*. CRC press.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7:457–472.

Ghosh, M. (2020). Small area estimation: Its evolution in five decades, with discussion. *Statistics in Transition*, 21(4):1–67.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gómez-Rubio, V., Best, N., Richardson, S., Li, G., and Clarke, P. (2010). Bayesian statistics for small area estimation. *Office for National Statistics, United Kingdom*.

Hájek, J. (1971). Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu. *Foundations of statistical inference*, page 326.

Han, B. (2013). Conditional Akaike information criterion in the Fay-Herriot model. *Statistical Methodology*, 11:53–67.

Hans, C. (2010). Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing*, 20(2):221–229.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143):8.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hogg, J., Cameron, J., Cramb, S., Baade, P., and Mengersen, K. (2023). A two-stage Bayesian small area estimation method for proportions. *arXiv preprint arXiv:2306.11302*.

Honório, N. A., Nogueira, R. M. R., Codeço, C. T., Carvalho, M. S., Cruz, O. G., de Avelar Figueiredo Mafra Magalhães, M., de Araújo, J. M. G., de Araújo, E. S. M., Gomes, M. Q., Pinheiro, L. S., et al. (2009). Spatial evaluation and modeling of dengue seroprevalence and vector density in Rio de Janeiro, Brazil. *PLOS Neglected Tropical Diseases*, 3(11):e545.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Institut national de santé publique du Québec (2024). Ligne du temps COVID-19 au Québec. https://www.inspq.qc.ca/covid-19/donnees/ligne-du-temps, Accessed March 15h, 2024.

Johnson, D. S., Smeeding, T. M., and Torrey, B. B. (2005). Economic inequality through the prisms of income and consumption. *Monthly Labor Review*, 128:11–24.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18):2555–2567.

Krennmair, P. and Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(5):1865–1894.

Lahiri, P. and Suntornchost, J. (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhya B*, 77:312–320.

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

Lawson, A. B. and Clark, A. (2002). Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, 21(3):359–370.

Lee, D. and Lawson, A. (2016). Quantifying the spatial inequality and temporal trends in maternal smoking rates in Glasgow. *The Annals of Applied Statistics*, 10(3):1427.

Lee, D. and Mitchell, R. (2013). Locally adaptive spatial smoothing using conditional auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(4):593–608.

Lee, D., Rushworth, A., and Napier, G. (2018). Spatio-temporal areal unit modeling in R with conditional autoregressive priors using the CARBayesST package. *Journal of Statistical Software*, 84(9):1–39.

Lehot-Couette, M. (2020). COVID-19 : taux d'incidence record, surmortalité... Comment les Ehpad sont frappés par la seconde vague de l'épidémie. https://www.francetvinfo.fr/sante/maladie/coronavirus/confinement/covid-19-taux-dincidence-record-surmortalite-comment-les-ehpad-sont-frappes-par-la-seconde-vague-en-huit-graphiques_4219253.html.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Leroux, B. G., Lei, X., and Breslow, N. (1999). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179–191. Springer.

Lohr, S. L. (2021). *Sampling: design and analysis*. Chapman and Hall/CRC.

Lowe, R., Barcellos, C., Brasil, P., Cruz, O. G., Honório, N. A., Kuper, H., and Carvalho,

M. S. (2018). The Zika virus epidemic in Brazil: from discovery to future implications. *International Journal of Environmental Research and Public Health*, 15(1):96.

Lumley, T. and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1):1–18.

MacNab, Y. C. (2011). On Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods in Medical Research*, 20(1):49–68.

MacNab, Y. C. (2022). Bayesian disease mapping: past, present, and future. *Spatial Statistics*, 50:100593.

MacNab, Y. C. (2023). Revisiting Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods in Medical Research*, 32(1):207–225.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15(1):661–667.

McConville, K. S., Breidt, F. J., Lee, T. C., and Moisen, G. G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5(2):131–158.

McConville, K. S. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2):389–413.

Michal, V., Freitas, L. P., Schmidt, A. M., and Cruz, O. G. (2024). A Bayesian hierarchical model for disease mapping that accounts for scaling and heavy-tailed latent effects. arXiv:2109.10330v2.

Michal, V., Vanciu, L., and Schmidt, A. M. (2022). A joint hierarchical model for the number of cases and deaths due to COVID-19 across the boroughs of Montreal. *Spatial and Spatio-temporal Epidemiology*, 42:100518.

Molina, I., Nandram, B., and Rao, J. (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. *The Annals of Applied Statistics*, 8(2):852–885.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., and DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan. *Spatial and Spatio-temporal Epidemiology*, 31:100301.

Napier, G., Lee, D., Robertson, C., Lawson, A., and Pollock, K. G. (2016). A model to estimate the impact of changes in MMR vaccine uptake on inequalities in measles susceptibility in Scotland. *Statistical Methods in Medical Research*, 25(4):1185–1200.

Newhouse, D. (2023). Small area estimation of poverty and wealth using geospatial data: What have we learned so far? *Calcutta Statistical Association Bulletin*. https://doi.org/10.1177/00080683231198591.

Nguyen, M. C., Corral, P., Azevedo, J. P., and Zhao, Q. (2017). Small area estimation: An extended ELL approach. *Retrieved from World Bank*.

Nobre, A. A., Schmidt, A. M., and Lopes, H. F. (2005). Spatio-temporal models for mapping the incidence of malaria in Pará. *Environmetrics*, 16(3):291–304.

Nogueira, R. M. R., Miagostovich, M. P., Schatzmayr, H. G., dos Santos, F. B., de Araújo, E. S., de Filippis, A. M. B., de Souza, R. V., Zagne, S. M. O., Nicolai, C., Baran, M., et al. (1999). Dengue in the state of Rio de Janeiro, Brazil, 1986–1998. *Memórias do Instituto Oswaldo Cruz*, 94:297–304.

Palacios, M. B. and Steel, M. F. J. (2006). Non-Gaussian Bayesian geostatistical modeling. *Journal of the American Statistical Association*, 101(474):604–618.

Pfeffermann, D. (2002). Small area estimation - new developments and directions. *International Statistical Review*, 70(1):125–143.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1):40–68.

Porwal, A. and Raftery, A. E. (2022). Comparing methods for statistical inference with model uncertainty. *Proceedings of the National Academy of Sciences*, 119(16):e2120737119.

Prefeitura do Rio de Janeiro (2018). Índice de Desenvolvimento Social (IDS) por Áreas de Planejamento (AP), Regiões de Planejamento (RP), Regiões Administrativas (RA), Bairros e Favelas do Município do Rio de Janeiro - 2010. http://www.data.rio/datasets/fa85ddc76a524380ad7fc60e3006ee97.

Rao, J. N. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Raymundo, C. E. and de Andrade Medronho, R. (2021). Association between socio-environmental factors, coverage by family health teams, and rainfall in the spatial distribution of Zika virus infection in the city of Rio de Janeiro, Brazil, in 2015 and 2016. *BMC Public Health*, 21(1):1199.

Reich, B. J. and Ghosh, S. K. (2019). *Bayesian Statistical Methods*. Chapman and Hall/CRC.

Richardson, S., Thomson, A., Best, N., and Elliott, P. (2004). Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives*, 112(9):1016–1025.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent

Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Rushworth, A., Lee, D., and Mitchell, R. (2014). A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatio-temporal Epidemiology*, 10:29–38.

Rushworth, A., Lee, D., and Sarran, C. (2017). An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(1):141–157.

Santafé, G., Adin, A., Lee, D., and Ugarte, M. D. (2021). Dealing with risk discontinuities to estimate cancer mortality risks when the number of small areas is large. *Statistical Methods in Medical Research*, 30(1):6–21.

Santé publique France (2023). Données hospitalières relatives à l'épidémie de COVID-19. https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/, Accessed March 20th, 2023.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, 60:144–162.

Sørbye, S. H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2, http://mc-stan.org/.

Steinberger, L. and Leeb, H. (2016). Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., and Rojas-Perilla, N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4):927–979.

Ugarte, M., Etxeberria, J., Goicoa, T., and Ardanaz, E. (2012). Gender-specific spatio-temporal patterns of colorectal cancer incidence in Navarre, Spain (1990–2005). *Cancer Epidemiology*, 36(3):254–262.

United Nations (2015). Transforming our world: the 2030 Agenda for Sustainable Development. https://sdgs.un.org/2030agenda.

Urdangarin, A., Goicoa, T., and Ugarte, M. D. (2023). Evaluating recent methods to overcome spatial confounding. *Revista Matemática Complutense*, 36(2):333–360.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian analysis*, 16(2):667–718.

Ville de Montréal (2016). Profils sociodémographiques 2016. http://ville.montreal.qc.ca/portal/page?_pageid=6897,68149755&_dad=portal&_schema=PORTAL, Accessed September 15th, 2021.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1):1625–1651.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.

Wakefield, J. (2013). *Bayesian and frequentist regression methods*. Springer.

Wakefield, J., Okonek, T., and Pedersen, J. (2020). Small area estimation for disease prevalence mapping. *International Statistical Review*, 88(2):398–418.

Wang, J. C., Opsomer, J. D., and Wang, H. (2014). Bagging non-differentiable estimators in complex surveys. *Survey Methodology*, 40(2):189–210.

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):3571–3594.

Wieczorek, J. (2023). Design-based conformal prediction. *arXiv preprint arXiv:2303.01422*.

World Bank (2015). Software for poverty mapping. https://www.worldbank.org/en/research/brief/software-for-poverty-mapping.

World Health Organization (2020). Vector-borne diseases. https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases.

Wright, M. N., Wager, S., and Probst, P. (2022). Ranger: A fast implementation of random forests. R package version 0.14.1.

Yan, J. (2007). Spatial stochastic volatility for lattice data. *Journal of Agricultural, Biological, and Environmental Statistics*, 12(1):25–40.

Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. (2019). Random forest prediction intervals. *The American Statistician*, 74:392–406.

Zhou, Z., Mentch, L., and Hooker, G. (2021). V-statistics and variance estimation. *Journal of Machine Learning Research*, 22(287):1–48.