UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

GUILHERME LOPES DE OLIVEIRA

CHALLENGES IN MODELING COUNT DATA: BAYESIAN MODELS FOR CORRECTION OF UNDERREPORTING BIAS AND ESTIMATION OF MORTALITY SCHEDULES

Belo Horizonte
2020

GUILHERME LOPES DE OLIVEIRA

CHALLENGES IN MODELING COUNT DATA: BAYESIAN MODELS FOR CORRECTION OF UNDERREPORTING BIAS AND ESTIMATION OF MORTALITY SCHEDULES
(Desafios na modelagem de dados de contagem: modelos Bayesianos para correção de viés de subnotificação e estimação de curvas de mortalidade)

A dissertation submitted to the *Programa de Pós-graduação em Estatística* of the *Universidade Federal de Minas Gerais* in partial satisfaction of the requirements for the degree of Doctor in Statistics.

Advisor 1: Profa. Dra. Rosangela Helena Loschi
Advisor 2: Prof. Dr. Renato Martins Assunção

Belo Horizonte
2020

Oliveira, Guilherme Lopes de.

O48c Challenges in modeling count data [manuscrito]: Bayesian models for correction of underreporting bias and estimation of mortality schedules / Guilherme Lopes de Oliveira. – 2020.
vii, 121 f. il.

Orientadora: Rosangela Helena Loschi.
Coorientador: Renato Martins Assunção.
Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f.10-13; 42-45; 59-60; 78-81; 109-112; 120-121.

1. Estatística – Teses. 2. Método de Monte Carlo – Teses. 3. Bioestatística – Teses. 4. Teoria bayesiana de decisão estatística – Teses. 5. Poisson, Distribuição de – Teses. 6. Recém-nascidos – Mortalidade – Teses. I. Loschi, Rosangela Helena. II. Assunção, Renato Martins III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. IV.Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

UFMG

ATA DA DEFESA DE TESE DE DOUTORADO DO ALUNO GUILHERME LOPES DE OLIVEIRA, MATRICULADO NO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA, DO INSTITUTO DE CIÊNCIAS EXATAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, REALIZADA NO DIA 03 DE NOVEMBRO DE 2020.

Aos 03 dias do mês de novembro de 2020, às 13h00, em reunião pública virtual (conforme orientações para a atividade de defesa de tese durante a vigência da Portaria PRPG nº 1819), reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologa pelo Colegiado do Programa de Pós-Graduação em Estatística, para julgar a defesa de tese do aluno Guilherme Lopes de Oliveira, intitulada: "*Challenges in modeling count data: Bayesian models for correction of underreporting bias and estimation of mortality schedules*", requisito final para obtenção do Grau de Doutor em Estatística. Abrindo a sessão, a Senhora Presidente da Comissão, Profa. Rosangela Helena Loschi, passou a palavra ao aluno para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do aluno. Após a defesa, os membros da banca examinadora reuniram-se reservadamente sem a presença do aluno e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação:

(x) Aprovada.
( ) Reprovada com resubmissão do texto em ____ dias.
( ) Reprovada com resubmissão do texto e nova defesa em ____ dias.
( ) Reprovada.

RLoschi
Rosangela Helena Loschi -orientadora

Renato Martins Assunção co-orientador

Flávio B. Gonçalves
Flávio Bambirra Gonçalves

Leonardo Soares Bastos

TFonseca
Thaís Cristina Oliveira da Fonseca

Wagner Barreto de Souza

O resultado final foi comunicado publicamente ao aluno pela Senhora Presidenta da Comissão. Nada mais havendo a tratar, a Presidenta encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 03 de novembro de 2020.

Observações:
1. No caso de aprovação da tese, a banca pode solicitar modificações a serem feitas na versão final do texto. Neste caso, o texto final deve ser aprovado pelo orientador da tese. O pedido de expedição do diploma do candidato fica condicionado à submissão e aprovação, pelo orientador, da versão final do texto.
2. No caso de reprovação da tese com resubmissão do texto, o candidato deve submeter o novo texto dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca que então decidirão pela aprovação ou reprovação da tese.
3. No caso de reprovação da tese com resubmissão do texto e nova defesa, o candidato deve submeter o novo texto com a antecedência à nova defesa que o orientador julgar adequada. A nova defesa, mediante todos os membros da banca, deve ser realizada dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca. Baseada no novo texto e na nova defesa, a banca decidirá pela aprovação ou reprovação da tese.

To my family, especially my parents, Afonso and Eva.

Agradecimentos

Esta foi uma jornada gratificante e devo agradecer a todos(as) que de alguma forma contribuíram para essa conquista. Em primeiro lugar, agradeço a Deus por ter colocado a Estatística na minha vida e por ter me dado saúde para chegar até aqui.

Agradeço à minha orientadora, Professora Rosangela Loschi, por sua dedicação, entusiasmo, comprometimento, paciência e por sempre me manter confiante sobre as valiosas contribuições deste trabalho. Esta tese não seria a mesma sem seu apoio constante. A orientação da Rosangela durante os últimos oito anos me ajudou a crescer tanto como profissional quanto como pessoa. Agradeço também ao meu co-orientador, Professor Renato Assunção, por todo o apoio e pelas importantes lições e discussões que contribuíram para meu crescimento como pesquisador.

Agradeço aos meus pais, Afonso e Eva, e aos meus irmãos Bárbara, Felipe, Paula e Wanderson, por sempre acreditarem no meu potencial desde a época da minha graduação, mesmo sem saberem exatamente do que se tratava essa tal ”Estatística”. Em especial, devo agradecer à minha mãe por sempre ter me mostrado o caminho dos estudos como perspectiva de melhora de vida, abrindo todas as portas que podia para facilitar meu caminho.

Eu agradeço imensamente aos meus amigos de longa data Ana Cláudia Silva, Danielle Moreira, Ebert França, Guilherme Veloso, Letícia Nunes, Rafael Aguiar e Vinícius França, irmãos que a vida me deu, por terem sempre me motivado e me feito distrair em momentos de estresse com os estudos e a pesquisa, principalmente. Ao Guilherme Veloso presto meu agradecimento especial pelo companheirismo, amor e paciência ao longo dos últimos anos; além de todo apoio técnico e emocional dado nas frequentes discussões sobre a pesquisa.

Meus colegas da pós-graduação também merecem meu agradecimento por todo o conhecimento compartilhado, pela ajuda constante durante os estudos e por todas as conversas relaxantes, tanto aquelas pelos corredores do DEST-UFMG quando aquelas acompanhadas de uma boa cerveja gelada. Em especial, menciono Cristiano de Carvalho, Douglas Roberto, Erick Amorim, Gabriela Oliveira, Juliana Freitas, Jussiane Gonçalves e Uriel Silva.

Aos Professores Flávio Gonçalves, Leonardo Bastos, Thaís Fonseca e Wagner Souza eu agradeço pela participação na banca examinadora e pelas contribuições dadas para o aprimoramento da versão final desta tese.

Sou extremamente grato aos professores e funcionários do Departamento de Estatística da UFMG que de alguma forma me ajudaram nesta caminhada, em especial o Professor Roberto Quinino e a secretária Rogéria Figueiredo.

Devo agradecer também aos professores e colegas do Departamento de Computação (DECOM) do CEFET-MG, pela compreensão e apoio com a redução parcial dos meus encargos didáticos no DECOM durante o período de elaboração final desta tese.

Agradeço à CAPES pela bolsa de estudos nos dois primeiros anos como aluno do programa de doutorado e durante todo o mestrado; ao CNPq pelo apoio financeiro durante o curso de graduação e à FAPEMIG pelo apoio para participação em congressos ao longo da minha trajetória na UFMG.

Por fim, agradeço à todos(as) os(as) demais que, direta ou indiretamente, passaram pela minha vida, torceram por mim e me deram apoio e incentivo nessa jornada.

Muito obrigado!
Guilherme Oliveira

”It is better to solve the right problem approximately than to solve the wrong problem exactly.” (John Tukey)

Resumo

Em diversas áreas do conhecimento como, por exemplo, Epidemiologia e Demografia, dados de contagem são coletados com o intuito de avaliar ou monitorar os riscos associados aos eventos de interesse. No entanto, muitas vezes esses dados não são completamente registrados. Em vez disso, apenas uma fração do verdadeiro total de eventos é observada, caracterizando o fenômeno conhecido por subnotificação, muito comum em estudos epidemiológicos. Se a subnotificação ocorre e não é levada em consideração, as inferências feitas a partir das contagens observadas serão viesadas e, consequentemente, os riscos relacionados aos eventos de interesse seraão subestimados. Além da questão da subnotificação, dados de contagem podem apresentar alta esparcidade, como geralmente ocorre em estudos demográficos a respeito dos padrões de mortalidade em populações humanas. Nesta tese, nós abordamos estes problemas desafiadores comumente presentes na análise estatística baseada em dados de contagem. Dentre os modelos propostos, tem-se duas abordagens para a correção do viés de subnotificação, as quais foram publicadas em periódicos relevantes em Estatística, além de uma metodologia alternativa para a estimação e suavização de curvas de mortalidade por idade e sexo na presença de dados esparsos, a qual está em estágio de aprimoramento. Um introdução mais aprofundada sobre os problemas práticos abordados é fornecida no capítulo inicial, o qual também traz uma descrição detalhada das contribuições em cada modelo proposto. Os capítulos sequentes são apresentados no formato de coleção de artigos, os quais apresentam metodologias independentes com discussões individuais dos problemas abordados. Em todos os casos, o processo de inferência é feito sob o paradigma Bayesiano. Algumas abordagens disponíveis na literatura são discutidas e, em certos casos, utilizadas para comparação com os modelos propostos. Dados simulados e conjuntos de dados reais são utilizados para explorar e ilustrar as principais caracterÍsticas dos modelos. O capítulo final traz um resumo compacto dos métodos e resultados obtidos nos estudos desenvolvidos ao longo da tese, destacando alguns pontos interessantes para estudos futuros.

Palavras-chave: curvas de mortalidade; identificabilidade; inferência Bayesiana; métodos Monte Carlo via cadeias de Markov, modelo Poisson censurado; modelo Poisson composto; mortalidade neonatal, sub-registro, taxa de incidência de tuberculose; técnica de aumento de dados;

.

Abstract

In several fields, such as epidemiology and demography, count data is collected in order to assess or to monitor the risks associated with the events of interest. However, in many situations only a fraction of the true total of events is observed, characterizing the phenomenon known as underreporting, which is very common in epidemiological studies. If the underreporting occurs and it is not accounted for, the inference made from the observed counts will be biased and, consequently, the risks related to the events of interest will be underestimated. In addition to the issue of underreporting, in some studies the observed counts may be highly sparse, as usually occurs in the analysis of mortality patterns in demographic studies. In this dissertation, we address these challenging problems commonly faced when analyzing count data. Among the proposed models, there are two approaches for the correction of underreporting bias, which have been published in relevant journals in statistics, as well as an alternative methodology for estimating and smoothing mortality curves by age and sex in the presence of sparse data, which is been improved. A broader introduction to the practical problems addressed in the dissertation is provided in the opening chapter, which also provides a detailed description of the contributions related to each proposed model. The subsequent chapters corresponds to a collection of papers, which present independent methodologies with individual discussions of the problems addressed. In all cases, the inference process is made under the Bayesian paradigm. Some approaches available in the statistical literature are discussed and, in some cases, used for comparison with the proposed models. Simulated data as well as real datasets are used to explore and to illustrate the main features of the models. The final chapter summarizes the methods and results obtained throughout the dissertation, highlighting some interesting points for future research.

Keywords: Bayesian inference; censored Poisson model; compound Poisson model; data augmentation; Markov chain Monte Carlo methods, model identifiability; mortality schedules; neonatal mortality, tuberculosis incidence, underreporting.

Contents

Chapter 1

Introduction

The main goal of statistical modeling is to describe the characteristics of a given system and its relationship with possible external factors through probabilistic models. Theoretical knowledge about the phenomenon under study, assumptions about the data collection process, search for parsimony, among other things, may guide the construction of such models.

However, even if the model is adequate to describe the behavior of the data, the inference may be compromised if the process of observation and measurement of the data has been impaired in any way. That is the case, for example, in several studies involving the estimation of rates based on count data in which only a fraction of the total number of events is reported, characterizing what we call *underreporting*.

Although information about unreported events is somewhat *missing*, underreporting differs from the widespread concept of *missing data*. In the common *missing data* problems, information about the lack of part of the observations is available and, therefore, it can be directly incorporated into the analysis. In the case of unreported events, which is the focus in this work, no information is generated regarding the loss of a certain amount of observations. Thus, statistical methods that allow the incorporation of extra information about such a phenomenon throughout the modeling process are of great practical interest.

Another problem that can compromise the inference process when dealing with count data is the low occurrence of cases (e.g., high frequency of null counts) in many sample units. This problem can naturally occur when modeling rare events or when there is a high degree of underreporting. Such a problem commonly arises when one are analyzing the occurrence of events in small populations where there are few sample units at risk. That is the case, for example, when the goal is to estimate mortality rates by sex and age groups (mortality schedules) in small areas. The use of adequate methods and models to deal with such data features is quite relevant to avoid poor and biased inferences.

In this work we approach two problems involving count data: (i) the correction of underreporting bias and (ii) the estimation of mortality schedules in small populations. The text is organized in chapters as a collection of papers in which the methods are presented, applied and discussed. For the appropriate treatment of underreported data we present two Bayesian hierarchical methodologies which can be applied in different situations (Chapters 2 and 4). For the estimation of mortality curves we propose the use of Bayesian dynamic models which have shown to be promising in many scenarios (Chapter 3). Finally, Chapter 5 closes this dissertation with the main concluding remarks, emphasizing the challenge of modeling count data with a defective reporting mechanism as well as the problem of estimating and smoothing mortality patterns in subnational small populations. In Chapter 5 we also suggest some potential topics for future research. In the following Sections 1.1 and 1.2 we briefly present the motivations and main contributions regarding the methods presented in Chapters 2, 3 and 4.

1.1 Modeling Underreported Count Data: Motivation and Contributions

1.1.1 Motivation

Count data is collected in several fields of science, such as criminology, epidemiology and traffic safety. These data are used to assess and monitor the risks inherent to the associated events they represent. The quality of the risk estimates depends on the quality of the available data. Any system for counting and reporting events is prone to errors which may arise from different sources.

In the public health field, for example, the reporting systems for infectious diseases, such as HIV, or chronic diseases, such as leprosy, usually present record failures as a result of diagnosis error or by the fact that some patients avoid diagnosis. The incidence of several other diseases is commonly underreported in several parts of the world, especially in less developed regions [Campbell *et al.*, 2011; Alfonso *et al.*, 2015; Shaweno *et al.*, 2017]. Epidemiologists around the world also points failures in the civil system for reporting deaths, especially for infant deaths in underdeveloped countries and regions with low educational levels [Campos *et al.*, 2007; Szwarcwald *et al.*, 2011; Viswanathan *et al.*, 2010; Xu *et al.*, 2014]. Likewise, Tennekoon [2017] shows that zero inflation in self-reported intentional abortion counts is also related to underreporting. In actuarial sciences, insurance companies usually face an unknown number of total claims, as some claims are usually made late. Such a delay on the number of insurance claims produce underreporting at least for a certain period of time, that is, until the events being properly claimed. As discussed by Bastos *et al.* [2019] and Stoner and Economou [2020] reporting delay is also an obstacle for real-time tracking of epidemics such as dengue, Ebola or COVID-19. An example in industrial production is the total number of products that are broken within a certain period, usually the warranty period. Knowing this number is important for quality management. However, only the number of products returned is known, while the total number also includes products that, although defective, are not returned by customers. A similar problem occurs in notifications of traffic accidents with minor damage [Wood *et al.*, 2016]. In criminology, it is known that crimes associated with factors that cause embarrassment for the victim and those committed by family members are not always reported to the police

[Li *et al.*, 2003]. The same is observed for crimes involving the robbery of goods of low financial value [Tibbetts and Hemmens, 2010]. Another important motivational example in this field is evidenced by the National Victimization Survey (PNV) performed in Brazil in 2013 [PNV, 2013]. The PNV revealed an average underreporting of 80.1% of cases related to twelve types of crimes, including kidnapping, assaults and sexual offense. In the specific case of rape crimes in Brazil, the underreporting is even more several. It is estimated that only 10 % of cases are reported to police agencies [Cerqueira and Coelho, 2014]. Decision-making based on crime rate estimates obtained from these poor quality data will certainly be inappropriate.

Therefore, if the occurrence rates for these crimes are calculated with basis only on the reported cases, the authorities responsible for the public security sector will find difficulties to identify the need for implementation of protective measures or an increase in the number of specialized police agencies, for example.

From the previous discussion, it can be seen that underreporting in count data is a widespread phenomenon. Independently of the source, from the statistical point of view, any incorrectness on the data registration process represents a bias problem. If underreporting occurs and is not accounted for, the inference made from the observed counts will be biased and, consequently, the risks and other amounts associated with the events of interest will be underestimated.

Bias correction in statistical estimators can be made, for example, through techniques such as Bartlett's correction and *bootstrap* [Cordeiro and Cribari, 2014]. However, techniques to correct bias induced by data collection problems, such as underreporting, are less common. Problems related to zero-inflation, delayed data, sample selection and preferential sampling will not be discussed. Some models to address these problems can be find, e.g., in Piancastelli and Barreto-Souza [2019]; Gonçalves and Barreto-Souza [2020]; Bastos *et al.* [2019]; Stoner and Economou [2020]; Bastos and Barreto-Souza [2020]; Dinsdale and Salibian-Barrera [2019] and references there in.

The treatment of the underreporting problem to be approached in this work is not a simple task, as the data itself does not carry any information about the quality (or precariousness) of the event reporting process. Because of that, it is always necessary to introduce extra information in the modeling process so that some correction of the bias induced by this problem can be made. In a general context, the strategy is to specify a joint model for the counting and reporting processes. The few methods currently available in the literature are applicable in restricted situations, being conditioned to the existence of validation datasets or informative prior distributions for specific, and ideally interpretable, parameters. The two approaches that have been frequently considered in the literature are briefly discussed in Sections 1.1.2 and 1.1.3. They serve as the basis for the models we developed in Chapters 2 and 4.

1.1.2 Contributions on Censored Poisson Models

One way to approach the underreporting in count data is to treat the data suspect of suffering from such a phenomenon as censored data. Terza [1985] introduced the so-called censored Poisson regression model by extending the class of the Poisson regression models to the context of censored count data in which the censoring threshold is known and constant for all observations. Caudill and Mixon [1995] extends this model by considering that the censoring thresholds are known and they may vary between observations. This sort of models are approached in details in Cameron and Trivedi [1998]. Bailey *et al.* [2005] considered the class of models proposed by Caudill and Mixon [1995] for handling underreported count data, more specifically, the leprosy counts in the neighborhoods of Olinda, Brazil. In this case, observations suspected of being underreported are treated as censored observations.

In order to assimilate the theoretical construction of the censored Poisson model, consider a region divided into A areas and assume that a total of N individuals are at risk in the region, over a fixed period of time. Denote by T_i the total number of events in the i-th area, $i = 1, \ldots, A$, and by N_i the total number of individuals at risk in that area. Consider that E_i is a known offset quantity representing the expected number of events at area i. The offset E_i allows for the variation in the population size over the areas. In practice one can assume that, for instance, $E_i = N_i$ or $E_i = (N_i \sum_{i=1}^{A} T_i)/N$. Finally, assume that $T_i \mid \theta_i \overset{ind}{\sim} \mathcal{P}oisson(E_i\theta_i)$, where θ_i is the event occurrence risk at area i. The assumption that all T_i variables are fully observed is not realistic in many practical situations. Then, consider that some of them can be right-censored (underreported), that is, for some regions $T_i \geq Y_i$, with Y_i being the reported (observed) count. Let γ_i be a censoring indicator variable whose value is 1 if the count at area i is censored and 0 otherwise. Then, the associated likelihood function is given by

$$f(\boldsymbol{y} \mid \boldsymbol{\gamma}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \left[\frac{e^{E_i\theta_i}(E_i\theta_i)^{y_i}}{y_i!} \right]^{1-\gamma_i} \left[\sum_{y \geq y_i} \frac{e^{E_i\theta_i}(E_i\theta_i)^{y}}{y!} \right]^{\gamma_i} \right\}, \tag{1.1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_A)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_A)$. In this context, the offset representing the expected number of events at area i can be calculated using the naive estimator $E_i = (N_i \sum_{i=1}^{A} Y_i)/N$.

In the approach of Bailey *et al.* [2005], the binary censoring variable $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_A)$ is considered to take into account the possible data reporting failures in some areas. However, in this context, $\boldsymbol{\gamma}$ is not directly observed as in Caudill and Mixon [1995]'s approach. To overcome this problem, Bailey *et al.* [2005] suggest the use of *ad hoc* procedures to define the censored (underreported) areas, based on some socioeconomic indicators and information obtained from specialists. Oliveira [2016] performed a sensitivity analysis on the censored Poisson model using simulated datasets finding that such an approach is quite restrictive because it is very sensitive to small changes in the censoring criterion. Thus, its use is appropriate only for situation in which one precisely know the censored areas.

In order to introduce more flexibility into the censored Poisson model presented in equation

(1.1), in my master's thesis [Oliveira , 2016] I introduced a more general model which considers the censoring indicator variables $\gamma_i,\ i = 1, \ldots, A$, as parameters of the model in order to account for the uncertainty associated to the censored data. A probabilistic structure is assigned to the censoring mechanism instead of requiring its precise prior specification, leading to the called Random-Censoring Poisson Model (RCPM). Under the RCPM, the term γ_i is considered as a latent random variable such that $\gamma_i \mid \pi_i \overset{ind}{\sim} \mathcal{B}ernoulli(\pi_i)\ \forall\ i$, i.e., in each area i the count Y_i is assumed to be underreported with probability π_i. In addition, it is assumed that, given $\boldsymbol{\theta}$, the observed counts and the censoring variables are independent. Then, the joint model for $\boldsymbol{Y}$ and $\boldsymbol{\gamma}$ is hierarchically obtained through the likelihood function given in equation (1.1) and $f(\boldsymbol{\gamma} \mid \boldsymbol{\pi}) = \prod_{i=1}^{n} \pi_i^{\gamma_i}\ (1 - \pi_i)^{1-\gamma_i}$. Under the RCPM, it is possible to estimate both the risks associated with the event of interest and the probability of each observation being censored (underreported). The key point is to model the uncertainty about the censoring probabilities $\boldsymbol{\pi} = (\pi_1, ..., \pi_A)$, for which is considered a logistic regression. Extra information coming, for instance, from expert's opinion is required to model the uncertainty about the regression parameters. Inference is made under the Bayesian paradigm and a data augmentation technique [Tanner and Wong, 1987] is proposed for sampling from the full conditional posterior distributions.

Although the theoretical structure which defines the RCPM was introduced in Oliveira [2016], such an approach was substantially improved during the first three semesters of my Ph.D. The MCMC scheme introduced in Oliveira [2016] for sampling from the posterior distribution was modified, improving adequacy and efficiency of the estimation process. More specifically, the Metropolis-Hastings algorithm and the blocking strategy used to sample from the model parameters was completely reformulated. The competing model introduced by Moreno and Girón [1998] was implemented. A broader simulation study was performed also including a comparison with the approach of Moreno and Girón [1998]. A sensitivity analysis considering different prior distributions for parameters related to the reporting process was accomplished under both the RCPM and the Moreno and Girón [1998]'s approaches. The application to infant mortality data in Minas Gerais, Brazil, was also improved by including a sensitivity analysis to the prior specifications for the censoring probabilities $\boldsymbol{\pi}$ and also by considering datasets from different periods of time. The RCPM methodology and the improvements mentioned above generated a paper published in Statistics in Medicine [Oliveira *et al.* , 2017]. The paper with the complete model specification, the proposed MCMC scheme, simulation studies and applications, as well as its supplementary material, are presented in Chapter 4. Chapters 2 and 3 present the most important contributions of this dissertation, which are described in the following.

1.1.3 Contributions on Compound Poisson Models

Another usual approach in the context of underreported data is to treat the model for the observed counts as a composition of the model associated with the total (unobserved) counts. The resulting model belongs to the class of the so-called compound Poisson models (CPM), which is based on a Binomial thinning scheme.

As in the previous section, assume that $T_i \mid \theta_i \overset{ind}{\sim} \mathcal{P}oisson(E_i\theta_i)$, $i = 1, \ldots, A$, where θ_i is the event occurrence risk and E_i is a known offset quantity representing the expected number of events at area i. For the number of reported events, Y_i, assume that $Y_i \mid T_i, \epsilon_i \overset{ind}{\sim} \mathcal{B}inomial(T_i, \epsilon_i)$, where ϵ_i represents the reporting probability or, equivalently, the proportion of the total count T_i that is effectively recorded. It can be shown that the marginal distribution of Y_i with respect to (w.r.t.) the total unobserved count T_i is

$$Y_i \mid \theta_i, \epsilon_i \overset{ind}{\sim} \mathcal{P}oisson(E_i\theta_i\epsilon_i), \; \forall \; i = 1, \ldots, A. \tag{1.2}$$

Although it is a quite attractive modeling strategy, if no additional information is available, the model given in expression (1.2) is not identifiable. It is straightforward seeing that in each area only the product $\eta_i = \theta_i\epsilon_i$ is estimated from the observed count Y_i, since, any other combination of parameters, say $\tilde{\theta}_i$ and $\tilde{\epsilon}_i$, which provide the same product $\eta_i = \tilde{\theta}_i\tilde{\epsilon}_i$ generates the same likelihood function.

The lack of identifiability inherent to the CPM can be overcome by imposing restrictions in the parametric space or introducing extra information into the analysis. A detailed discussion on this issue is presented in Chapter 2. In the following, we summarize the strategies currently proposed in the literature in order to introduce the reader to the modeling difficulties under the CPM.

Firstly, we discuss about the approach proposed by Schmertmann and Gonzaga [2018], which make use of additional information to directly elicit an informative prior distribution for each $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, A$. In their application, such extra information comes from additional samples obtained by an active search procedure available from previous related studies. A similar strategy is used by Moreno and Girón [1998]. However, this approach is feasible only for practical situations where there is available strong prior information about the reporting probability within each sample unit i. Also, especially in areas experiencing the worst quality in the registration process, the uncertainty surrounding the available information can be so high that it becomes useless to guarantee the model identifiability.

Another approach for the CPM, which has been considered in several works in the literature (e.g., Whittemore and Gong [1991]; Powers *et. al* [2010]; Papadopoulos and Silva [2008]; Stoner *et al.* [2019]), assumes that, in each area, both the risk θ_i and the reporting probability ϵ_i depend on a set of covariates, say $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$, such that

$$\log(\theta_i) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} \qquad \text{and} \qquad \text{logit}(\epsilon_i) = \alpha_0 + \boldsymbol{w}_i'\boldsymbol{\alpha}, \tag{1.3}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ denote the fixed effects and β_0 and α_0 denote the intercept terms, $i = 1, \ldots, A$. The model defined by equation (1.3) is usually called by Pogit (Poisson-Logistic) model. Eventually, random effects are included in the model in order to take into account spatial and regional information complementary to that contained in the available covariates. From the modeling point of view, $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ can be the same or one can be a subset of the other. A convenient and less restrictive modeling strategy would be to include all available variables as regressors in both parts of the model. However, Papadopoulos and Silva [2008] show that, even with the inclusion of covariates, it is not possible to identify the model without using extra information. These authors suggest considering the inclusion of signal restrictions or exclusion restrictions in components of the fixed effects $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ and intercepts terms β_0 and α_0. Such identifiability constraints may not be feasible in general practical situations.

Still under the Pogit model, another way to get around the intrinsic lack of identification is to include an additional validation dataset. This is the case in the problem addressed by Whittemore and Gong [1991], Stamey *et al.* [2006], Powers *et. al* [2010] and Dvorzak and Wagner [2015]. The main dataset of interest to these authors concerns to the number of deaths caused by cervical cancer in four European countries for women in four age groups. The diagnosis of the cause of death is subject to error leading to under-registration of the total amount of deaths. In this particular application, a sample of 50 physicians from each country filled out a death certificate for a specific patient who had died by cervical cancer and the number of correct death certificates in each country was recorded. With this, a validation (gold-standard) dataset is available as a proxy for quality of the cervical cancer deaths registration process in each country. The incorporation of these data into the likelihood function makes it possible to identify the model and estimate the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ separately. However, in general applications, validation datasets are rarely available and they can be very expensive to obtain. Stoner *et al.* [2019] proposed a different Bayesian modeling framework when applying the Pogit model to Brazilian tuberculosis data. Nevertheless, similar to the previously mentioned approaches, their model can only be applied when strong prior information is available for some specific model parameters.

From the previous discussion it becomes evident the difficulty related to the wide application of the CPM as it has been approached in the literature. Its identifiability depends on specific additional information about the reporting process that is not always available. With focus on practical situations in which trustful prior information is only available for areas known to experience the highest data quality levels, we propose in Chapter 2 a novel Bayesian approach for the compound Poisson model. We model the probability of underreporting in a different way by assuming a dependence among such probabilities provided that the areas of interest are grouped according to their data quality. We use such a clustering structure to relate the reporting probabilities such that they decrease as we move from the best group to the worst ones. A deep discussion on the model identifiability is provided. We obtain constraints for model identifiability and we prove that only prior information about the reporting probability in areas experiencing the best data quality is required to identify the model. We assess some model characteristics using simulation experiments and illustrate its use by analyzing Brazilian infant mortality datasets. Our approach for the CPM also requires the availability of specific prior information to guarantee the model identifiability, but it emerges as an alternative modeling strategy for the adequate treatment of underreported count data. The proposed methodology has been accepted for publication in Bayesian Analysis [Oliveira *et al.*, 2020]. In addition, a new application of the model in a real epidemiological problem is provided in the appendix of Chapter 2. We reanalyze the Brazilian tuberculosis data considered by Stoner *et al.* [2019]. This analysis is used to exemplify the potential application of the proposed model and the necessary paths to adapt the available data to the model's structure in general contexts. We use a conventional procedure to define the clusters. Our solution is compared with the results obtained under the approach given in Stoner *et al.* [2019].

The research developed in Chapter 2 originated from discussions in the scope of the project "Sensitivity analysis and Bayesian robustness in partition model with application in neo-natal mortality mapping". This is an international cooperation project Brazil-Italy CNPq-CNR, grant 19/2011, 2012-14 coordinated by researchers Márcia D'élia Branco (USP, Brazil) and Fabrizio Ruggeri (CNR-IMATI, Italy) and with the participation of Rosangela Helena Loschi (UFMG, Brazil) and Raffaele Argiento (Università Cattolica del Sacro Cuore, Italy).

1.2 Estimation and Smoothing of Mortality Schedules: Motivation and Contributions

The challenge of estimating human mortality rates is faced in different fields, specially in demography. The death rates are used along with other indicators, such as the probability of dying and the number of survivors for each age or specific age groups, to create a specific tabulation called life table. Life expectancy at birth, an output of a life table, is a widely used indicator to compare levels of mortality and health status among populations. A complete set of age-specific mortality rates is the starting point for deriving the value of life expectancy at birth and other summary indicators of mortality or longevity. Often, life tables are also calculated separately for males and females.

Some important goals of mortality modeling include describing the shape of mortality curves by age and sex (referenced in this work as *mortality schedules*), estimating and projecting mortality patterns over time and investigating differences in mortality patterns across different populations. The data is available in the form of counts of deaths and individuals exposure to risk in each specific age and sex stratum. The representation of mortality patterns is a special issue in small populations where observed data tends to be sparse and erratic. Also, it is a difficult task when dealing with populations from less developed regions because data coming from such areas is usually unreliable or incomplete.

Traditionally, mortality rates from large populations that present good data quality are used to observe empirical and mathematical regularities in the mortality schedules, which could be applied to support the estimation in sparse (and small) populations. Mortality models defined under such an approach are commonly known as relational models. Human mortality schedules tend to present an specific pattern when plotted on the logarithm scale (see an example for males' mortality in Panel (A) of Figure 1.1), which justifies the use of well-established curves in the relational models. To exemplify the problem with the sparsity of data observed in small areas, which impairs a simple estimation of the complete underlying mortality schedule, we present in Panels (B) and (C) of Figure 1.1 the observed mortality rates in two selected Brazilian municipalities with quite different population sizes. It can be noticed that, for the largest selected municipality, given in Panel (B), the common males' mortality pattern is easily identified whereas for the smallest one, given in Panel (C), it is quite difficult to recognize the common shape for the mortality curve due to the large amount of ages with a null count, that is, with no observed death at the specific age.

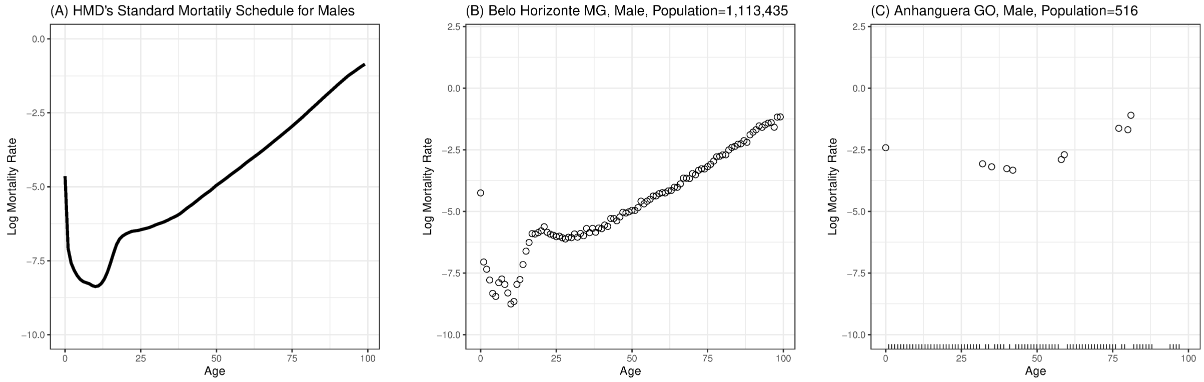


Figure 1.1: Open circles in Panels (B) and (C) show the observed mortality rates (in the logarithm scale) for males in two Brazilian municipalities for each single-year of age $0, 1, ..., 99$. Tick marks on the horizontal axis represent ages with no observed deaths. The black curve in Panel (A) displays an standard mortality schedule for males obtained from all life tables available in the Human Mortality Database [Wilmoth *et al.*, 2020].

As a wide variety of types and sources of data have become available, mortality modeling has focused on the development of flexible relational models that perform well in several contexts [Wilmoth *et al.*, 2012; Alkema and New, 2014; Alexander *et al.*, 2017; Gonzaga and Schmertmann, 2016; Clark, 2019]. Chapter 3 brings our proposal of using Bayesian dynamic models for estimating and smoothing mortality curves. The log-mortality rates are related to a standard mortality schedule through regression models whose parameters are dynamically related across the age intervals. In our approach, the standard mortality schedule is obtained from the Human Mortality Database (HMD), a database that contains detailed population and mortality data (life tables) for 41 countries judged to have good data quality [Wilmoth *et al.*, 2020]. Preliminary results using simulated and real datasets are presented and discussed, including comparison with the approach of Gonzaga and Schmertmann [2016]. Some points for future investigation are discussed. The goal is to improve the proposed methodology to a posterior submission in a peer-reviewed journal.

References

Alexander, M., Zagheni, E. and Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography*, **54**(6), 2025–2041.

Alfonso, J., Lovseth, E., Samant, Y., and Jolm, J. (2015). Work-related skin diseases in Norway may be underreported: data from 2000 to 2013. *Contact Dermatitis*, **72**(6), 409–412.

Alkema, L. and New, J.R. (2014). Global estimation of child mortality using a Bayesian B-spline bias-reduction model. *The Annals of Applied Statistics*, 8(4), 2122–2149.

Bailey, T.C., Carvalho, M.S., Lapa, T.M., Souza, W.V., and Brewer, M.J. (2005). Modeling of under-detection of cases in disease surveillance. *Annals of Epidemiology*, **15**(5), 335–343.

Bastos, L.S., Economou, T., Gomes, M.F.C., Villela, D.A.M., Coelho, F.C., Cruz, O.G., Stoner, O., Bailey, T., and Codeço, C.T. (2019). A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in Medicine*, **38**(22), 4363–4377.

Bastos, F.B., and Barreto-Souza, W. (2020). Birnbaum-Saunders sample selection model. *Journal of Applied Statistics*, DOI: 10.1080/02664763.2020.1780570.

Cameron, A.C., and Trivedi, P.K. 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK.

Campbell, L., Hills, S., Fischer, M., Jacobson, J., Hoke, C., Hombach, J., Marfin, A., Solomon, T., Tsai, T., Tsu, V., and Ginsburg, A. (2011). Estimated global incidence of Japanese encephalitis: a systematic review. *Bulletin of the World Health Organization*, **89**(10), 766–774.

Campos, D., Loschi, R.H., and França, E. (2007). Mortalidade neonatal precoce hospitalar em Minas Gerais: Associação com variáveis assistenciais e a questão da subnotificação (in Portuguese). *Revista Brasileira de Epidemiologia*, **10**(2), 223–238.

Caudill, S.B., and Mixon Jr., F.G. (1995). Modeling household fertility decisions: Estimation and testing censored regression models for count data. *Empirical Economics*, **20**, 183–196.

Cerqueira, D., and Coelho, D.S.C. (2014) Estupro no Brasil: uma radiografia segundo os dados da Saúde (in Portuguese). Nota técnica do Instituto de Pesquisa Econômica Aplicada (IPEA), N^o 11. Available at www.ipea.gov.br/portal/images/stories/PDFs/nota_tecnica/140327_notatecnicadiest11.pdf.

Clark, S.J. (2019). A General Age-Specific Mortality Model With an Example Indexed by Child Mortality or Both Child and Adult Mortality. *Demography* **56**, 1131–1159.

Cordeiro, G.M., and Cribari-Neto, F. (2014). *An Introduction to Bartlett Correction and Bias Reduction*. New York: Springer.

Dinsdale, D. and Salibian-Barrera, M. (2019). Methods for preferential sampling in geostatistics. *Journal of the Royal Statistical Society, Applied Statistics, Series C*, **68**(Part 1), 181–198.

Dvorzak, M., and Wagner, H. (2016). Sparse Bayesian modelling of underreported count data. *Statistical Modelling*, 16(1), 24–46.

Gonçalves, J.N., and Barreto-Souza, W. (2020). Flexible regression models for counts with high-inflation of zeros. *METRON*, **78**, 71–95.

Gonzaga, M.R. and Schmertmann, C.P. (2016). Estimating age- and sex-specific mortality rates for small areas with TOPALS regression: an application to Brazil in 2010. *Revista Brasileira de Estudos de População*, **33**(3), 629–652.

Li, T., Trivedi, P,K. and Guo, J. (2003) Modeling Response bias in count: A structural approach with an application to the national crime victimization survey data. *Sociological Methods and Research*, **31**, 514–544.

Moreno, E., and Girón J. (1998). Estimating with incomplete count data: A Bayesian approach. *Journal of Statistical Planning and Inference*, **66**(1), 147–159.

Oliveira, G.L. (2016). Modeling Underreported Infant Mortality Data with a Random Censoring Poisson Model. Master's thesis, Statistics Department, Universidade Federal de Minas Gerais, Belo Horizonte. Available at http://hdl.handle.net/1843/BUBD-A89PEH.

Oliveira, G.L., Loschi, R.H. and Assunção, R.M. (2017). A random-censoring Poisson model for underreported data. *Statistics in Medicine*, **36**(30), 4873–4892.

Oliveira, G.L., Argiento, R., Loschi, R.H., Assunção, R.M., Ruggeri, F. and Branco, M.D. (2020). Bias correction in clustered underreported data. To appear at *Bayesian Analysis.*

Papadopoulos, G., and Silva, J.M.C.S. (2008). Identification issues in models for underreported counts. Discussion Paper Series No 657. Depart. of Economics, University of Essex.

Piancastelli, L.S.C., and Barreto-Souza, W. (2019). Inferential aspects of the zero-inflated Poisson INAR(1) process. *Applied Mathematical Modelling*, **74**, 457–468.

PNV - Pesquisa Nacional de Vitimização. (2013) Sumário Executivo SENASP (in Portuguese). Available at www.crisp.ufmg.br/wp-content/uploads/2013/10/Sumario_SENASP_final.pdf.

Powers, S., Gerlach, R. and Stamey, J. (2010). Bayesian variable selection for Poisson regression with underreported responses. *Computational Statistics and Data Analysis*, **54**, 3289–3299.

Schmertmann, C., and Gonzaga, M. R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*, **55**(4), 1363–1388.

Shaweno, D., Trauer, J.M., Denholm. J.T., and McBryde, E.S. (2017). A novel Bayesian geospatial method for estimating tuberculosis incidence reveals many missed TB cases in Ethiopia. *BMC Infectious Diseases*, **17**(662).

Stamey, J.D., Young, D.M., and Boese, D.(2006). A Bayesian hierarchical model for Poisson rate and reporting-probability inference using double sampling. *Australian & New Zealand Journal of Statistics*, **48**(2), 201–212.

Stoner, O; Economou, T; Drummond, G. (2019). A Hierarchical Framework for Correcting Under-Reporting in Count Data. *Journal of the American Statistical Association*, **114**(528), 1481–1492.

Stoner, O; and Economou, T. (2020). Multivariate hierarchical frameworks for modeling delayed reporting in count data. *Biometrics*, **76**(3), 1481–1492.

Szwarcwald, C.L., Morais Neto, O.L., Escalante, J.J.C., Souza Jr., P.R.B., Frias, P.G., Lima, R.B., and Viola, R.C. (2011). Busca ativa de óbitos e nascimentos no Nordeste e na Amazônia Legal: estimação das coberturas do SIM e do SINASC nos municípios brasileiros (in Portuguese). *Saúde Brasil 2010: uma análise da situaç ao de saúde e de evidências selecionadas de impacto de ações de vigilância em saúde.* Ministério da Saúde, Brasília, 79–98.

Tanner, M.A. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation . *Journal of the American Statistical Association*, **82**, 528–540.

Tennekoon, V.S. (2017). Counting unreported abortions: A binomial-thinned zero-inflated Poisson model. *Demographic Research*, **36**(2), 41–72.

Terza, J.P. (1985). A Tobit-type Estimator for the Censored Poisson Regression Model. *Economics Letters*, **18**, 361–365.

Tibbetts, S.G. and Hemmens, C. (2010). Criminological Theory, A Text/Reader. *Sage Publication*, London, UK.

Viswanathan, K., Becker, S., Hansen, P., Kumar, D., Kumar, B., Niayesh, H., Peters, D., and Burnham, G. (2010). Infant and under-five mortality in Afghanistan: current estimates and limitations. *Bulletin of the World Health Organization*, **88**(8), 576–583.

Whittemore, A.S., and Gong, G. (1991). Poisson Regression with Misclassified Counts: Application to Cervical Cancer Mortality Rates. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **40**(1), 81–93.

Wilmoth, J., Zureick, S., Canudas-Romo, V., Inoue, M. and Sawyer. C. (2012) A flexible two-dimensional mortality model for use in indirect estimation *Popul Stud (Camb)*, **66**(2), 1–28.

Wilmoth, J.R., Andreev, K., Jdanov, D., Glei, D.A. and Riffe, T. with the assistance of Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., Vachon, P., Winant, C. and Barbieri, M. (2020). Methods Protocol for the Human Mortality Database. University of California at Berkeley (United States) and the Max Planck Institute for Demographic Research (Rostock, Germany). Last Revised: August 8, 2020 (Version 6). Available at https://www.mortality.org/Public/Docs/MethodsProtocol.pdf.

Wood, J.S., Donnell, E.T., and Fariss, C.J.(2016) A method to account for and estimate underreporting in crash frequency research. *Accident Analysis & Prevention*, **95**(Part A), 57–66.

World Human Organization. (2006) *Neonatal and perinatal mortality: Country, regional and global estimates.* WHO Library. Cataloguing-in-Publication Data.

Xu, Y., Zhang, W., Yang, R. Zou, B., and Zhao, Z. (2014). Infant mortality and life expectancy in China. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, **20**, 379–385.

Chapter 2

Bias correction in clustered underreported data

Abstract[1]

Data quality from poor and socially deprived regions have given rise to many statistical challenges. One of them is the underreporting of vital events leading to biased estimates for the associated risks. To deal with underreported count data, models based on compound Poisson distributions have been commonly assumed. To be identifiable, such models usually require extra and strong information about the probability of reporting the event in all areas of interest, which is not always available. We introduce a novel approach for the compound Poisson model assuming that the areas are clustered according to their data quality. We leverage these clusters to create a hierarchical structure in which the reporting probabilities decrease as we move from the best group to the worst ones. We obtain constraints for model identifiability and prove that only prior information about the reporting probability in areas experiencing the best data quality is required. Several approaches to model the uncertainty about the reporting probabilities are presented, including reference priors. Different features regarding the proposed methodology are studied through simulation. We apply our model to map the early neonatal mortality risks in Minas Gerais, a Brazilian state that presents heterogeneous characteristics and a relevant socio-economical inequality.

Keywords: compound Poisson; generalized Beta distribution; Jeffreys prior; model identifiability; neonatal mortality; underreporting.

[1] The content of this chapter has been accepted for publication in *Bayesian Analysis* [Oliveira *et al.*, 2020].

2.1 Introduction

The estimation of economic, health and social indicators in underdeveloped and developing countries has been a challenging task due to the low quality of the available data. In such areas, even with the recent advances, data coming from official collection systems usually experience considerable underreporting of events. To cite an example, it is common to miss the report of infants who die shortly after birth. If not accounted for, such a phenomenon typically lead to the underestimation of vital statistics, compromising the definition of appropriate government intervention policies and distribution of financial resources.

In the statistical literature, the bias problem induced by a defective data reporting process is commonly handled by considering hierarchical models that accommodate truncated or censored observations. For mapping the risks associated to count events subjected to underreporting, Bailey *et al.* [2005] consider the censored Poisson regression model proposed by Caudill and Mixon Jr. [1995] assuming that, for suspected areas, the observed count represents a right-censoring threshold for the true non-observed total number of events. This approach relies on the fact that, *a priori*, all areas experiencing underreporting are precisely known. Bailey *et al.* [2005] consider *ad-hoc* procedures to determine the censored (underreported) areas. Later, Oliveira *et al.* [2017] define a random-censoring Poisson model (RCPM) introducing more flexibility in the analysis of underreported count data. The RCPM allows for the estimation of both the associated occurrence rates and the probability of each area to experience censoring. The authors showed that quality of posterior estimates is related to the availability of informative prior distributions for the censoring probabilities.

The compound Poisson model (CPM) is an alternative approach to deal with potentially underreported counts. It allows for the joint modeling of the event occurrence rates and the associated reporting probabilities. The main difference between RCPM and CPM is that the former models the underreporting status of each area: Is area i suffering from underreporting or not? In turn, the latter models the area-specific probability of each particular event being reported, then all areas are, in principle, subject to underreporting.

To guarantee the CPM identifiability, it is necessary to introduce prior information on the reporting process. This has been carried out in different ways in the literature depending on the context and the type of information available. For example, Whittemore and Gong [1991], Stamey, Young and Boese [2006] and Dvorzak and Wagner [2015] resort to a validation dataset on the reporting process. This refers to another independent data source, free of underreporting, that can be used to calibrate the bias induced by the underreporting in the main dataset under analysis. Such additional gold standard dataset does not necessarily have to be on the same scale as the primary data but it has to be available for each sample unit. Thus, validation datasets are rarely available and they can be very expensive to obtain. All three previous papers use the same illustrative example which is based on a single validation dataset of severely restrictive extent. Specifically, their validation dataset is based on a 1987 study that selected a sample of 203 physicians divided in four groups according to their nationality (England, Belgium, France and Italy). In each group, the sample of physicians was asked to complete a specimen death certificate for the case history of a single 51-year-old woman with an ulcerating tumor of the cervix. The certificate had enough information to induce the correct classification of the patient as a victim of cervical cancer. However, the groups reached different proportions of death certificates correctly coded as cervical cancer. The result is then used as a gold-standard estimation of the correct diagnosis and completion of death certificates for this specific cancer as the underlying cause of death. Hence, this validation dataset is outdated and should be looked cautiously if used for recent death data. Furthermore, it is useful only for one single cancer (cervical cancer) in four specific countries, being hardly generalizable for other sorts of cancers or other regions.

Moreno and Girón [1998] resort to a different strategy as they did not have a validation dataset in their studies of reported assaults in Málaga and Stockholm. They provide a detailed investigation under the CPM whenever conjugate families are considered to independently model the prior uncertainty for the reporting probabilities and the occurrence rates. The authors emphasize that prior information on the reporting probabilities is expected to be included to make feasible the posterior estimation. Such information can be obtained through specific surveys or from experts' opinion and then it must be conveniently used to set the hyperparameters of the conjugate prior distributions. Following Moreno and Girón [1998]'s approach, Schmertmann and Gonzaga [2018] consider the CPM to estimate the age-specific mortality and life expectancy for small areas with defective vital records in Brazil. Probabilistic prior information on the death registration coverage in each area is considered to elicit an informative Beta prior distribution for the death reporting probability in three age groups. The authors derived such a prior information from standard demographic estimation techniques, such as the Death Distribution Methods, and also from intensive field audits conducted by Brazilian public health researchers.

As an alternative to these previous models, Stoner, Economou, and Drummond [2019] present a Bayesian hierarchical CPM to account for the underreporting in tuberculosis counts in Brazil. To complement the partial information in the data, their model only requires an informative prior distribution for the mean reporting rate. To elicit such an informative prior across all Brazilian microregions, the authors consider external estimates of the overall tuberculosis detection rate derived by the World Health Organization through an inventory study.

Trustful prior information about the overall mean reporting process is not always available. Sometimes, one can only count with pieces of prior information on the reporting process for some subsets of areas, obtained through local inventory studies (local active search for cases) or experts' opinion. In many epidemiological studies, for example, one may only know *a priori* that the severity levels of underreporting are likely associated with some socioeconomic indicators or, merely, that less socially deprived areas properly record a greater percentage of their events, producing more reliable information [see Campos, Loschi and França, 2007; Bailey *et al.*, 2005; Silva *et al.*, 2017, for instance]. That is the case, for example, when mapping the infant mortality rates in underdeveloped regions, such as Africa and Latin America, based on data coming from defective death registration systems [World Health Organization, 2006; Alkema and New, 2014; Alexander and Alkema, 2018].

Inspired by situations in which validation datasets are unaccessible and reliable prior information about the reporting process is only available for areas experiencing the best data quality, we propose a new hierarchical Bayesian approach for the CPM (Section 2.2). It considers that the areas composing the region of interest are ordered according to data quality categories. If it is reasonable to additionally cluster the areas into homogeneous groups, then the model becomes more parsimonious. The clusters may be defined with basis on experts' opinion or applying some clustering technique to data quality indicators provided by previous studies and surveys. In our model, the data quality clustering of the areas is a tool used to model their reporting probabilities. We leverage the clusters to create a hierarchical structure in which the reporting probabilities decrease as we move from the best data quality areas to the worst ones. The novelty in our approach is that only an informative prior distribution about the reporting probability at areas experiencing the best data quality is required to ensure identifiability (Section 2.2.1). To model the event occurrence rates, we consider a set of potential covariates through a regression structure. Bayesian variable selection is incorporated into the model to identify regressors with a non-zero effect.

Extensive simulation studies are presented to evaluate the performance of the proposed model in different scenarios (Section 2.3). We apply the developed Bayesian methodology to estimate the early neonatal mortality rates in Minas Gerais State, Brazil, for the periods 1999–2001 and 2009–2011 (Section 2.4), where the death counts are known to be underreported [Campos, Loschi and França, 2007]. In this context, the proposed approach is attractive because neither validation datasets nor prior knowledge about the overall mean reporting probability are available. Section 2.5 closes the paper with our main conclusions.

2.2 Model specification

Consider a region divided into A areas and denote by T_i the total number of events at area i, for $i = 1, \ldots, A$. Assume that $T_i \mid \lambda_i \overset{ind}{\sim} \mathcal{P}oisson(\lambda_i)$, where λ_i is the mean expected counts in the ith area. The relative risk for the event at area i is given by $\theta_i = \lambda_i / E_i$, where E_i is a known offset quantity representing the expected number of events in such area. The offset E_i allows for a variation in the population size over the areas. In the context of underreported data, T_i is not fully observed for, at least, part of the areas, so that the reported number of events Y_i may corresponds to only a fraction of T_i. To consider this data feature, each event occurring at area i is associated to a binary random variable $Z_{i,t} \sim \mathcal{B}ernoulli(\epsilon_i)$ that determines whether the tth event will be reported or not, where $\epsilon_i \in [0, 1]$ represents the associated reporting probability. The random variables in the sequence $Z_{i,1}, Z_{i,2}, , Z_{i,3}, \ldots$ are assumed as being identically distributed, mutually independent and also independent of T_i. Consequently, $Y_i = \sum_{t=1}^{T_i} Z_{i,t}$ has a compound Poisson distribution in which $Y_i \mid T_i, \epsilon_i \overset{ind}{\sim} \mathcal{B}inomial(T_i, \epsilon_i)$ and

$T_i \mid \theta_i \overset{ind}{\sim} \mathcal{P}oisson(E_i\theta_i)$. By marginalizing the joint distribution of (Y_i, T_i) over T_i, it follows that the observed count Y_i has the conditional distribution

$$Y_i \mid \theta_i, \epsilon_i \overset{ind}{\sim} \mathcal{P}oisson(E_i\theta_i\epsilon_i). \tag{2.1}$$

The model in expression (2.1) will be refereed as compound Poisson model (CPM) throughout this work. To model the relative risks $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_A)$, we assume that they are related to a set of p potential covariates such that $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$, $i = 1, \ldots, A$. Random effects may be included in the log-linear predictor to capture any residual spatial or local variation in the relative risks. The greatest challenge under the CPM is the modeling of the reporting probabilities $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_A)$. If no further information is available, only the parameter $\eta_i = \theta_i\epsilon_i$ is identified from the observed data Y_i since any parameter combination such that $\tilde{\eta_i} = \tilde{\theta_i}\tilde{\epsilon_i} = \theta_i\epsilon_i$ yields the same likelihood function.

Different approaches to model $\boldsymbol{\epsilon}$ have been discussed in the literature. Moreno and Girón [1998] and Schmertmann and Gonzaga [2018] directly model the uncertainty about ϵ_i using informative beta prior distributions. A more general approach assumes that $\epsilon_i = g(H_{1i}, \ldots, H_{mi})$, where $H_{1i}, \ldots, H_{mi}$ is a set of m covariates related to the reporting process and g is any non-negative function such that $0 < g(H_{1i}, \ldots, H_{mi}) < 1$ for all i. There are many possible choices for g. The most popular one is to assume that g is a logistic function, as in Whittemore and Gong [1991], Dvorzak and Wagner [2015] and Stoner, Economou, and Drummond [2019]. As discussed in Section 2.1, all these approaches require either strong prior information about $\boldsymbol{\epsilon}$ or validation datasets to ensure model identifiability.

One of the main goals in this work is to model $\boldsymbol{\epsilon}$ in situations where no validation dataset is available to guarantee model identifiability and whenever reliable prior information about the percentage of underreporting is only available for areas where data are known to be better reported. In this context, we assume that $\epsilon_i = g(H_{1i}, \ldots, H_{mi}) = 1 - \gamma - f(H_{1i}, \ldots, H_{mi})$, where $\gamma \in [0, 1)$ is the basal underreporting probability in the area with the best data quality and f is any non-negative function such that $f(H_{1i}, \ldots, H_{mi}) < 1 - \gamma$ for all i. The function f captures the increase in the basal underreporting probability explained by the covariates. If f equals to zero in the best area, then $f(H_{1l}, \ldots, H_{ml})$ denotes the increase in the underreporting probability for area l when compared to the best one. As in the model proposed by Stoner, Economou, and Drummond [2019], covariates $H_{1i}, \ldots, H_{mi}$ are assumed to be different from $X_{1i}, \ldots, X_{pi}$ to guarantee model identifiability. This model limitation may be avoided only if validation datasets are accessible as in Dvorzak and Wagner [2015]. A further discussion on this issue is given in Section 2.2.1.

The definition of a general f which satisfies all these constraints is not a simple task. To facilitate its construction, we assume that it is possible to sort the areas according to their data quality. Additionally, we assume that the reporting probabilities are equal for areas where the covariates related to the reporting process experience similar values. For this purpose, we assume that the A areas are grouped into K known data quality clusters, where $K \leq A$. We allow for

$K = A$ to account for situations in which there is no prior information for clustering the areas. However, if such information is available and then $K < A$, we obtain a more parsimonious model and more data information to estimate the reporting probabilities throughout the areas.

In practice, there are many ways to define the clusters. We may consider some grouping proposals available from previous works or to be guided by experts' information. Another possibility is to perform usual clustering techniques based on available covariates that are related to data quality in the region of interest.

Based on such grouping structure, we use a convenient coding scheme to represent the clustering indicator variable, which is different from variables in $X_{1i}, \ldots, X_{pi}$. Let $\mathbf{h}_i = (h_{1i}, \ldots, h_{Ki})^T$ be the clustering variable composed by binary quantities $h_{1i}, \ldots, h_{Ki}$ and defined according to the following split-coding scheme: if area i belongs to cluster j then $h_{li} = 1$ for all $l \leq j$ and $h_{li} = 0$ otherwise. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$, where $\gamma_j \in [0, 1)$ for all $j = 1, \ldots, K$, such that $\sum_{j=1}^{K} \gamma_j < 1$. We assume that the reporting probability at area i is given by

$$\epsilon_i = 1 - \mathbf{h}_i^T \boldsymbol{\gamma}. \tag{2.2}$$

The constraint imposed on $\boldsymbol{\gamma}$ is necessary to guarantee that $\epsilon_i \neq 0 \ \forall \ i$, and, consequently, to ensure non-null mean for the associated Poisson distribution.

The proposed model has some interesting features. Firstly, to be identifiable, it only requires information about the reporting probabilities in the best areas (see discussion in Section 2.2.1). Besides that, ϵ_i is represented in terms of interpretable parameters, which facilitates prior elicitation. For a given area i, $\mathbf{h}_i = (1, 0, \ldots, 0)^T$ and $\mathbf{h}_i = (1, 1, \ldots, 1)^T$ represent the two most extreme situations. If $\mathbf{h}_i = (1, 0, \ldots, 0)^T$ then the ith area has the highest level of data quality. We will assume that data in such area are recorded with a higher probability ($\epsilon_i = 1 - \gamma_1$) if compared to the areas in the remaining data quality groups. At the other extreme situation, if $\mathbf{h}_i = (1, 1, \ldots, 1)^T$ then the ith area lies in the worst data quality category. Data in this region are recorded with a lower probability ($\epsilon_i = 1 - \gamma_1 - \cdots - \gamma_K$) if compared to those areas belonging to clusters with better data quality. Thus, the parameter γ_1 represents the probability of not recording an event in areas classified in the highest level of data quality. The parameter γ_2 is the increment on such probability for areas experiencing the second highest data quality level, and so on. Another attractive feature of the proposed model is that, although the clustering indicator variable cannot be used to also model the relative risks $\boldsymbol{\theta}$, the covariates used for clustering are indirectly taken into consideration when estimating $\boldsymbol{\theta}$, since the areas belonging to the same cluster are homogeneous w.r.t. such clustering covariates.

2.2.1 On model identifiability

The lack of identifiability of the compound Poisson model presented in expression (2.1) has been discussed by several authors [Whittemore and Gong, 1991; Moreno and Girón, 1998; Stamey, Young and Boese, 2006; Papadopoulos and Silva, 2012; Dvorzak and Wagner, 2015;

Schmertmann and Gonzaga, 2018; Stoner, Economou, and Drummond, 2019]. All these previous works impose some constraints on $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ to attain model identifiability.

A well-known way to overcome non-identifiability problems requires extra information about the reporting probabilities $\boldsymbol{\epsilon} = (\epsilon_1, \ldots \epsilon_A)$. In the most extreme cases, all components of vector $\boldsymbol{\epsilon}$ should be fixed at a known quantity. Moreno and Girón [1998] and Schmertmann and Gonzaga [2018] show that this extreme constraint may be relaxed when the target of the statistical analysis is to estimate the relative risks. This is done by incorporating external estimates of registration coverage through very informative prior distributions about each component of $\boldsymbol{\epsilon}$.

To the best of our knowledge, there are two approaches to obtain an identifiable model when sets of covariates, say $\boldsymbol{X}$ and $\boldsymbol{H}$, are used to model the relative risk $\boldsymbol{\theta}$ and the reporting probability $\boldsymbol{\epsilon}$, respectively. The first one requires extra information from independent validation datasets [Whittemore and Gong, 1991; Stamey, Young and Boese, 2006; Dvorzak and Wagner, 2015]. This is a rare situation in practice that, however, does not require the intersection of $\boldsymbol{X}$ and $\boldsymbol{H}$ to be empty. The second one, adopted by Papadopoulos and Silva [2012] and Stoner, Economou, and Drummond [2019], creates some kind of linear separability of the covariates sets $\boldsymbol{X}$ and $\boldsymbol{H}$. Stoner, Economou, and Drummond [2019] build $\boldsymbol{X}$ and $\boldsymbol{H}$ by splitting the set of all available covariates into two disjoint sets based on experts' opinion. Hence, there is an empty intersection between the covariates in the sets $\boldsymbol{X}$ and $\boldsymbol{H}$ but this is not enough to guarantee identifiability. In their modeling framework, they also had available an informative prior distribution for the overall mean reporting rate which was sufficient to complete the identifiability conditions. Papadopoulos and Silva [2012] allow intersection between the two sets of covariates but impose prior information to establish appropriate constraints on the parametric space, such as restrictions on the signs or exclusion of some coefficients. This avoids the need for validation datasets.

Our approach also assumes, as in Stoner, Economou, and Drummond [2019], that the clustering covariates associated with $\boldsymbol{\epsilon}$ are not considered in the log-linear predictor of the relative risks $\boldsymbol{\theta}$. In principle, this constraint seems to be quite restrictive. Nevertheless, for model identifiability, what is required is the lack of strict mathematical collinearity between $\boldsymbol{X}$ and $\boldsymbol{H}$, but not their statistical independence. Thus, the two disjoint sets $\boldsymbol{X}$ and $\boldsymbol{H}$ may be correlated. In many practical situations, we can and probably we will have the two sets composed by covariates carrying similar information, measuring related aspects of the areas. For instance, to estimate infant mortality rates, one expects that poor social-economic conditions will affect both the relative risks and the reporting probabilities. It is true that to avoid the identifiability issues we must not use the same covariates when modeling θ and ϵ. However, we are allowed to use correlated variables, since our identifiability assumption requires just the strict empty intersection between the two sets, not the orthogonality of the information they carry. This makes our model much more attractive for practical implementation with respect to some of the previously proposed alternatives.

If the number of clusters K is smaller than the initial number of areas A, the clustering strategy proposed in expression (2.2) imposes a reduction in the parametric space related to the CPM in expression (2.1). Even under such a reduction and assuming that $\boldsymbol{X}$ and $\boldsymbol{H}$ are disjoint sets, the proposed CPM remains unidentifiable. Its identification will depend on the only trustful prior information we have available: the percentage of data reported in areas with the best data quality. Nevertheless, if such a piece of information is not available, other constraints for model identification are possible as discussed in the following.

Assume $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$, $i = 1, \ldots, A$, and denote by A_j the subset of areas belonging to the j-th data quality cluster, for $j = 1, \ldots, K$. Under these assumptions, the log-likelihood function associated with the proposed model is

$$\begin{aligned} l(\boldsymbol{\Psi};\boldsymbol{y}) &= \sum_{j=1}^{K}\sum_{i\in A_j}\left\{-E_i \exp\left\{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}\right\}\left(1-\sum_{l=1}^{j}\gamma_l\right)\right. \qquad (2.3) \\ &+ \left. y_i\left(\log E_i + \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \log\left(1-\sum_{l=1}^{j}\gamma_l\right)\right) - \log y_i!\right\}, \end{aligned}$$

where $\boldsymbol{\Psi} = (\beta_0, \beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_K)$. As the proposed model belongs to the exponential family, we obtain that $T(\boldsymbol{y}) = \left(\sum_{i=1}^{A} y_i, \sum_{i=1}^{A} y_i X_{1i}, \ldots, \sum_{i=1}^{A} y_i X_{pi}, \sum_{i\in A_1} y_i, \sum_{i\in A_2} y_i, \ldots, \sum_{i\in A_K} y_i\right)$ is the $(p+K+1)$-dimensional sufficient statistic for the parameter vector $\boldsymbol{\Psi}$. Note that, the first coordinate of vector $T(\boldsymbol{y})$ is a linear combination of the last K coordinates. Thus, the number of unknown parameters exceeds by one the number of linearly independent pieces of sample information (sufficient statistics). This implies that only $p + K$ parameters can be estimated without additional information [McHugh, 1956; Picci, 1977; Huang, 2005].

Proposition 2.2.1. *The proposed model under the specification in expression (2.3) is identifiable if β_0 or one of the coordinates of vector $\boldsymbol{\gamma}$ is fixed at a known value.*

Proof. Firstly, fix β_0 at a known value. In this case, model identifiability follows by noticing that the vector of sufficient statistics associated to the parameter vector $\boldsymbol{\Psi}^* = (\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_K)$ is given by $T^*(\boldsymbol{y}) = \left(\sum_{i=1}^{A} y_i X_{1i}, \ldots, \sum_{i=1}^{A} y_i X_{pi}, \sum_{i\in A_1} y_i, \sum_{i\in A_2} y_i, \ldots, \sum_{i\in A_K} y_i\right)$, which is composed by independent pieces of information. Similarly, without losing generality, let γ_1 to be known. Under this assumption the sufficient statistics related to the parameter vector $\boldsymbol{\Psi}^{**} = (\beta_0, \beta_1, \ldots, \beta_p, \gamma_2, \ldots, \gamma_K)$ are given in $T^{**}(\boldsymbol{y}) = \left(\sum_{i=1}^{A} y_i, \sum_{i=1}^{A} y_i X_{1i}, \ldots, \sum_{i=1}^{A} y_i X_{pi}, \sum_{i\in A_2} y_i, \ldots, \sum_{i\in A_K} y_i\right)$. In this case, the proof follows straightforwardly by noticing that the first coordinate of $T^{**}(\boldsymbol{y})$ can not be recovered as a linear combination of the last $p + K - 1$ coordinates as it depends on $\sum_{i\in A_1} y_i$. □

Our proposal is to approach situations in which trustful prior information is only available about γ_1. This parameter is easily interpretable as the underreporting probability in those areas having the best data quality. Thus, only prior information about the proportion of unrecorded data in such areas is required to identify the proposed model. Despite its appealing interpretation, the precise choice of the value for γ_1 may not be a simple task in practical situations. However, it is possible to obtain from experts some pieces of information about the most likely values for such parameter. This information may be suitably expressed by means of a non-degenerated informative prior distribution for γ_1 thus relaxing the requirement of exactly knowing its value (for further discussion on the use of prior information to attain model identification see Gustafson *et al.* [2005]).

Another way to investigate model identifiability is to consider the associated Fisher information. The Fisher information plays an important role in the asymptotic theory of maximum likelihood estimation as well as in Bayesian reference analysis. Besides that, Rothenberg [1971] showed that a model that belongs to the exponential family is globally identifiable if the Fisher information matrix is nonsingular. Let $\Lambda(j) = \left(1 - \sum_{l=1}^{j} \gamma_l\right)$ and $\mu_{ij} = E_i \exp\{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}\} \Lambda(j)$. The Fisher information matrix $\mathcal{I}(\boldsymbol{\Psi})$ resulting from expression (2.3) is given by

$$\mathcal{I}(\boldsymbol{\Psi}) = \begin{bmatrix} \sum\limits_{j=1}^{K}\sum\limits_{i\in A_j} \mu_{ij} & \cdots & \sum\limits_{j=1}^{K}\sum\limits_{i\in A_j} \mu_{ij}X_{pi} & \sum\limits_{j=1}^{K}\sum\limits_{i\in A_j} \frac{-\mu_{ij}}{\Lambda(j)} & \sum\limits_{j=2}^{K}\sum\limits_{i\in A_j} \frac{-\mu_{ij}}{\Lambda(j)} & \cdots & \sum\limits_{i\in A_K} \frac{-\mu_{ij}}{\Lambda(K)} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum\limits_{j=1}^{K}\sum\limits_{i\in A_j} \mu_{ij}X_{pi} & \cdots & \sum\limits_{j=1}^{K}\sum\limits_{i\in A_j} \mu_{ij}X_{pi}^2 & \sum\limits_{j=1}^{K}\sum\limits_{i\in A_j} \frac{-\mu_{ij}X_{pi}}{\Lambda(j)} & \sum\limits_{j=2}^{K}\sum\limits_{i\in A_j} \frac{-\mu_{ij}X_{pi}}{\Lambda(j)} & \cdots & \sum\limits_{i\in A_K} \frac{-\mu_{ij}X_{pi}}{\Lambda(K)} \\ \sum\limits_{j=1}^{K}\sum\limits_{i\in A_j} \frac{-\mu_{ij}}{\Lambda(j)} & \cdots & \sum\limits_{j=1}^{K}\sum\limits_{i\in A_j} \frac{-\mu_{ij}X_{pi}}{\Lambda(j)} & \boldsymbol{\sum\limits_{j=1}^{K}\sum\limits_{i\in A_j} \frac{\mu_{ij}}{\Lambda(j)^2}} & \boldsymbol{\sum\limits_{j=2}^{K}\sum\limits_{i\in A_j} \frac{\mu_{ij}}{\Lambda(j)^2}} & \cdots & \boldsymbol{\sum\limits_{i\in A_K} \frac{\mu_{ij}}{\Lambda(K)^2}} \\ \sum\limits_{j=2}^{K}\sum\limits_{i\in A_j} \frac{-\mu_{ij}}{\Lambda(j)} & \cdots & \sum\limits_{j=2}^{K}\sum\limits_{i\in A_j} \frac{-\mu_{ij}X_{pi}}{\Lambda(j)} & \boldsymbol{\sum\limits_{j=2}^{K}\sum\limits_{i\in A_j} \frac{\mu_{ij}}{\Lambda(j)^2}} & \boldsymbol{\sum\limits_{j=2}^{K}\sum\limits_{i\in A_j} \frac{\mu_{ij}}{\Lambda(j)^2}} & \cdots & \boldsymbol{\sum\limits_{i\in A_K} \frac{\mu_{ij}}{\Lambda(K)^2}} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum\limits_{i\in A_K} \frac{-\mu_{ij}}{\Lambda(K)} & \cdots & \sum\limits_{i\in A_K} \frac{-\mu_{ij}X_{pi}}{\Lambda(K)} & \boldsymbol{\sum\limits_{i\in A_K} \frac{\mu_{ij}}{\Lambda(K)^2}} & \boldsymbol{\sum\limits_{i\in A_K} \frac{\mu_{ij}}{\Lambda(K)^2}} & \cdots & \boldsymbol{\sum\limits_{i\in A_K} \frac{\mu_{ij}}{\Lambda(K)^2}} \end{bmatrix},$$

Which is a matrix of order $(1+p+K)\times(1+p+K)$. The $K \times K$ sub-matrix highlighted in bold will () be considered in Section 2.2.2 to build the Jeffreys prior for $\boldsymbol{\gamma}$ given $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$.

Proposition 2.2.2. *The Fisher matrix information $\mathcal{I}(\boldsymbol{\Psi})$ associated with the model given in expressions (2.1) and (2.2) is singular.*

Proof. Denote by $\mathcal{C}_\kappa$ the column vector of $\mathcal{I}(\boldsymbol{\Psi})$ associated to the parameter $\kappa \in \boldsymbol{\Psi}$ such that $\mathcal{I}(\boldsymbol{\Psi}) = \left[\mathcal{C}_{\beta_0} \ldots \mathcal{C}_{\beta_p}\ \mathcal{C}_{\gamma_1}\ \mathcal{C}_{\gamma_2} \ldots \mathcal{C}_{\gamma_K}\right]$. Let $\xi_0 = 1$, $\xi_1 = (1 - \gamma_1)$ and $\xi_j = -\gamma_j$ for $j = 2, \ldots, K$. Assuming these non-null constants, it follows that $\xi_0 \mathcal{C}_{\beta_0} + \xi_1 \mathcal{C}_{\gamma_1} + \sum_{j=2}^{K} \xi_j \mathcal{C}_{\gamma_j} = 0$. Thus, $\mathcal{I}(\boldsymbol{\Psi})$ is a singular matrix. From the results in Rothenberg [1971] it follows that the associated statistical model is not locally identifiable for at least a subset of the parametric space, thus characterizing the model lack of identifiability since local identification is a necessary condition to global identification. □

As previously shown in Proposition 2.2.1, model identifiability is achieved provided that the parameter γ_1 is fixed at a known value. In the general case, it is difficult to prove directly that the Fisher information matrix $\mathcal{I}(\boldsymbol{\Psi})$ is nonsingular when we fix γ_1. However, some special cases are amenable to analytic treatment and they are illuminating for this identifiability discussion as we show in Proposition 2.2.3.

Proposition 2.2.3. *Assume that the A areas experience a common relative risk such that $\log(\theta_i) = \beta_0$, for $i = 1, \ldots, A$. If γ_1 is fixed at a known value $\gamma_1^0 \in [0, 1]$ then the Fisher information matrix associated with the model given in expressions (2.1) and (2.2) is nonsingular.*

Proof. The Fisher information matrix $\mathcal{I}(\boldsymbol{\Psi}^*)$ under this model specification is obtained from $\mathcal{I}(\boldsymbol{\Psi})$ by removing the columns and rows related to parameters $\beta_1, \ldots, \beta_p$ and γ_1 and setting $\gamma_1 = \gamma_1^0$. After some calculation, we obtain that the determinant of $\mathcal{I}(\boldsymbol{\Psi}^*)$ is

$$\det \mathcal{I}(\boldsymbol{\Psi}^*) = \left(\sum_{i \in A_1} \mu_{i1} \right) \left[\prod_{j=2}^{K} \sum_{i \in A_j} \mu_{ij} \left(1 - \gamma_1^0 - \sum_{l=2}^{j} \gamma_l \right)^{-2} \right].$$

All sum terms in $\det \mathcal{I}(\boldsymbol{\Psi}^*)$ are positive. Consequently, we have $\det \mathcal{I}(\boldsymbol{\Psi}^*) > 0$ implying that $\mathcal{I}(\boldsymbol{\Psi})^*$ is a nonsingular matrix. From Theorem 3 in [Rothenberg, 1971], it follows that the associated statistical model is globally identifiable. □

The previous propositions provide some mathematical constraints for model identifiability, which are necessary to guarantee that all parameters can be estimated from the observed data. However, it is important noting that such constraints do not guarantee that all parameters will be well estimated, that is, having theoretical identifiability may not guarantee the practical identifiability. Even for an identifiable model, large sample sizes might be required to obtain good parameter estimates in some situations. On the other hand, for a non-identifiable model, some parameters might not be estimated even with large datasets if the identifiability constraints are not considered.

Remark 2.2.1. *As suggested by an anonymous referee, an equivalent representation of our model is obtained by considering the parameterization*

$$\epsilon_i = \exp\left\{ -\mathbf{h}_i^T \boldsymbol{\delta} \right\}, \tag{2.4}$$

where $\delta_1 = -\log(1 - \gamma_1)$, $\delta_j = -\log\left(1 - \sum_{l=1}^{j} \gamma_l\right) + \log\left(1 - \sum_{l=1}^{j-1} \gamma_l\right)$ and $\mathbf{h}_i$ is as defined in equation (2.2). Under this parametrization, the likelihood function is given by

$$\begin{aligned} l(\boldsymbol{\Psi}; \boldsymbol{y}) &= \sum_{i=1}^{A} \{ -E_i \exp\{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} - \delta_1 - \delta_2 h_{2,i} - \ldots - \delta_K h_{K,i}\} \\ &+ y_i \left(\log E_i + \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} - \delta_1 - \delta_2 h_{2,i} - \ldots - \delta_K h_{K,i} \right) \}. \end{aligned}$$

Concerning the model identification, the parametrization in (2.4) is quite attractive as it leads to a regular Poisson generalized linear model (GLM). By framing the model as a GLM, the conditions for model identification are easily found, especially the requirement that $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ are associated with disjoint sets of covariates. Also, as the first component of $\mathbf{h}_i$ is equal to 1 for all i, such parameterization makes it clear that δ_1 works like a second intercept for which an informative prior must be elicited. However, such a parametrization brings some additional *challenges to model the uncertainty about $\boldsymbol{\epsilon}$. While γ_j has a clear and meaningful interpretation for practitioners, δ_j is interpreted as the ratio between the reporting probability in cluster $j-1$ and cluster j in the log scale, for $j = 2, ..., K$. As for δ_1, it is the log of the proportion of recorded data in a perfect scenario where $\epsilon = 1$ in relation to the proportion of data recorded in the best cluster. Besides those interpretation issues, we also have a challenge regarding the appropriate prior specification for $\boldsymbol{\delta}$. To ensure a valid Poisson model we must have $\delta_j > 0$ for all j. As, a priori, we only have trustful information about ϵ_1 and we know that $0 < \epsilon_K \leq \epsilon_{K-1} \leq \cdots \leq \epsilon_2 \leq \epsilon_1 \leq 1$, we can not simply assume independent positive distributions for each parameter δ. Notice that $\delta_1 = \log(1) - \log(\epsilon_1)$ and $\delta_l = \log(\epsilon_{l-1}) - \log(\epsilon_l)$, for $l = 2, \ldots, K$. Then, we must transform the prior information of ϵ_1 to the log-scale and use it to build a distribution with positive support for δ_1. Then, the prior distribution of δ_2 should be such that the distribution of $\delta_2 + \delta_1 = -\log(\epsilon_2)$ is a truncated distribution putting all probability mass in values higher than δ_1. Similar constraints should be imposed to the prior distributions of the remaining δs.*

2.2.2 Prior distributions

In this section, we detail the prior distributions for the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_A)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$ which are required to complete our model specification.

2.2.2.1 Modeling the prior uncertainty about $\boldsymbol{\gamma}$

As a starting point, we could consider independent informative Beta distributions by eliciting $\gamma_j \overset{ind}{\sim} Beta(\alpha_j, \nu_j)$, $j = 1, \ldots, K$, where the hyperparameters $\alpha_j > 0$ and $\nu_j > 0$ should be elicited by experts. This strategy was considered by Schmertmann and Gonzaga [2018] in their particular application to estimate age-mortality rates in Brazil. This is a cumbersome approach as it might lead to some difficulties in the computational implementation of our model. First of all, the constraint $\sum_{j=1}^{K} \gamma_j < 1$ should be satisfied since some events are recorded even in areas belonging to the worst data quality cluster and also because, to have a valid Poisson model, ϵ_i must be non-null for all i. Furthermore, some dependence among the γ_j's is desirable. To deal with the first problem, we may consider a Dirichlet distribution on the augmented vector $(\gamma_1, \ldots, \gamma_K, \gamma_{K+1})$, where $\gamma_{K+1} = 1 - \sum_{j=1}^{K} \gamma_j$ is the percentage of data recorded in the worst cluster. More interestingly, both issues may be jointly addressed as described below. We propose considering a joint prior for $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$ as follows:

$$\left.\begin{array}{rcl} \gamma_1 & \sim & GBeta(\alpha_1, \nu_1; a_1, a_1^*), \\ \gamma_k \mid \gamma_{1:k-1} & \sim & GBeta\left(\alpha_k, \nu_k; a_k[1 - \sum_{j=1}^{k-1} \gamma_j], a_k^*[1 - \sum_{j=1}^{k-1} \gamma_j]\right), \ k = 2, \ldots, K, \end{array}\right\} \tag{2.5}$$

where $GBeta(\alpha, \nu; a, b)$ denotes the generalized Beta distribution with probability density function (p.d.f.) given by $f(x \mid \alpha, \nu; a, b) = \frac{\Gamma(\alpha+\nu)}{\Gamma(\alpha)\Gamma(\nu)(b-a)} \left(\frac{x-a}{b-a}\right)^{\alpha-1} \left(1 - \frac{x-a}{b-a}\right)^{\nu-1}$, $x \in (a, b)$, $\alpha > 0$, $\nu > 0$, $a \in \mathbb{R}$, $b \in \mathbb{R}$ with $a < b$. Such generalized Beta distribution can be obtained as the linear transformation $X = a + (b-a)B$, where $B \sim Beta(\alpha, \nu)$. In our case, $0 \leq a_j < a_j^* \leq 1$,

$j=1,\ldots,K$. By letting $a_j=0$ and $a_j^*=1$ for all $j=1,\ldots,K$, the prior distribution in expression (2.5) is the well-known stick-breaking representation of the Dirichlet process, in which we consider independent random variables $Z_j \sim Beta(\alpha_j,\nu_j)$, $j=1,\ldots,K$, and we let $\gamma_1=Z_1$ and $\gamma_j=Z_j\prod_{l=1}^{j-1}(1-Z_l)$ for $j=2,\ldots,K$. This is an advantageous feature we consider to facilitate the computational implementation of the generalized Beta prior distribution given in expression (2.5).

If we set $\alpha_j=\nu_j=1$, for $j=1,\ldots,K$, the conditional prior distributions given in expression (2.5) corresponds to a simpler model which is based on conditional uniform distributions so that

$$\left.\begin{array}{rcl}\gamma_1 & \sim & \mathcal{U}(a_1,a_1^*),\\ \gamma_k\mid\gamma_{1:k-1} & \sim & \mathcal{U}\left(a_k[1-\sum_{j=1}^{k-1}\gamma_j],a_k^*[1-\sum_{j=1}^{k-1}\gamma_j]\right),\ k=2,\ldots,K,\end{array}\right\} \tag{2.6}$$

where $0\leq a_j<a_j^*\leq 1$, $j=1,\ldots,K$. The uniform prior distribution in expression (2.6) is more parsimonious and easier to be elicited. In turn, the generalized Beta prior distribution in expression (2.5) is more flexible and provides different shapes for the marginal prior distribution of each γ_j. Thus, the choice between the prior distributions given by expressions (2.5) and (2.6) will depend on the information that the practitioner has available. In practice, the choice of all prior hyperparameters might be driven by experts' opinion or guided by results of previous studies. Special attention, however, should be given to the prior distribution of γ_1 as it plays an important role in the model identification. As discussed in Section 2.2.1, prior distribution of γ_1 has to be informative, putting a significant probability mass in the subset of the parametric space indicated by the experts as containing the most likely values for such parameter.

Independently of the prior that is chosen for $\boldsymbol{\gamma}$, the generalized Beta or the particular case of the conditional uniform, by assuming the structure in expression (2.2), the increment in the underreporting probability associated with each cluster j amounts just to a fraction of what is left after considering the probabilities of the previous (better) groups. Thus, the prior distribution for ϵ_i outside the best cluster inherits the prior information for the reporting probability in the best areas.

The unconditional prior expectation and variance of ϵ_i are useful whenever an informative prior distribution for γ_1 or any other component of parameter vector $\boldsymbol{\gamma}$ is to be elicited. Assuming the distribution in (2.5), respectively, the prior unconditional expectation and variance of ϵ_i, for all $i\in A_j$, i.e., all areas classified in the jth data quality cluster, for $j=1,\ldots,K$, are

$$\mathrm{E}(\epsilon_i)=\prod_{l=1}^{j}\{1-c_l\}\quad\text{and}\quad \mathrm{V}(\epsilon_i)=\mathrm{V}\left(\sum_{l=1}^{j-1}\gamma_l\right)\left[d_l+(1-c_l)^2\right]+d_l\left[1-\mathrm{E}\left(\sum_{l=1}^{j-1}\gamma_l\right)\right]^2,$$

where $c_l=a_l+(a_l^*-a_l)\alpha_l[\alpha_l+\nu_l]^{-1}$ and $d_l=[(a_l^*-a_l)^2\alpha_l\nu_l]\left[(\alpha_l+\nu_l)^2(\alpha_l+\nu_l+1)\right]^{-1}$. For the particular case in which $a_j = 0$ and $a_j^* = 1$ for all j, it follows that

$$\mathrm{E}(\gamma_j) = \frac{\alpha_j}{\alpha_j + \nu_j} \prod_{l=1}^{j-1} \frac{\nu_l}{\alpha_l + \nu_l}, \quad \text{and}$$
$$\mathrm{V}(\gamma_j) = \mathrm{E}(\gamma_j) \left(\frac{\alpha_j + 1}{\alpha_j + \nu_j + 1} \prod_{l=1}^{j-1} \frac{\nu_l + 1}{\alpha_l + \nu_l + 1} - \mathrm{E}(\gamma_j) \right).$$

Similar results under the conditional uniform prior distribution in expression (2.6) are provided in the Supplementary Material [Oliveira *et al.* Supplement, 2020].

In the Bayesian modeling framework, another way to model the prior uncertainty about the model parameters is to consider the Jeffreys' approach [Jeffreys, 1946]. Let $Y_i \mid \theta_i, \epsilon_i \overset{ind}{\sim} \mathcal{P}oisson(E_i\theta_i\epsilon_i)$ in which $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$. We assume that, *a priori*, $\boldsymbol{\gamma}$ is independent of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ and we only focus on the Jeffreys prior for $\boldsymbol{\gamma}$. The Fisher information matrix for the vector $\boldsymbol{\gamma}$, given $\boldsymbol{\beta}$, is the bottom right $K \times K$ submatrix highlighted in bold in $\mathcal{I}(\boldsymbol{\Psi})$ which is given in Section 2.2.1. Consequently, the Jeffreys prior distribution for $\boldsymbol{\gamma}$ becomes

$$\pi_J(\boldsymbol{\gamma} \mid \boldsymbol{\beta}) \propto \sqrt{\prod_{j=1}^{K} \left(1 - \sum_{l=1}^{j} \gamma_l\right)^{-1}}. \tag{2.7}$$

Our goal is to prove that the prior in expression (2.7) is a proper distribution and, more importantly, we aim to investigate the level of prior information about γ_1 that is induced by the Jeffreys prior.

Proposition 2.2.4. *The Jeffreys prior distribution for $\boldsymbol{\gamma}$ given in expression (2.7) is proper.*

Proof. The proof of Proposition 2.2.4 follows straightforwardly by noticing that the Jeffreys prior given in expression (2.7) may be represented as $\pi_J(\boldsymbol{\gamma} \mid \boldsymbol{\beta}) \propto \psi(\gamma_1)\psi(\gamma_2 \mid \gamma_1) \cdots \psi(\gamma_K \mid \gamma_1, \ldots, \gamma_{K-1})$, where $\psi(\gamma_1)$ is the kernel of the generalized Beta distribution $GBeta(1, 1/2; 0, 1)$ and $\psi(\gamma_j \mid \gamma_1, \ldots, \gamma_{j-1})$ is the kernel of a $GBeta\left(1, 1/2; 0, 1 - \sum_{l=1}^{j-1} \gamma_l\right)$, for $j = 2, \ldots K$. Consequently, $\pi_J(\boldsymbol{\gamma} \mid \boldsymbol{\beta})$ is proper as it belongs to the generalized Beta family of distributions given in expression (2.5). □

Assuming the Jeffreys prior in expression (2.7), the prior expected value of γ_1 is 0.6667 and its marginal prior distribution concentrates most probability mass around large values (see Figure 2.1). It is expected that such prior does not provide good posterior estimates for the model parameters whenever the true percentage of underreported events in areas with the best data quality is small and far from that prior expected value. Particularly, it is not an appropriate prior to model the uncertainty about γ_1 in the case study addressed in the paper where the probability of underreporting in the best areas is expected to be close to zero. To illustrate the effect of the marginal Jeffreys prior distribution of γ_1 on the joint Jeffreys prior for $\boldsymbol{\gamma}$, we also present in Figure 2.1 the joint Jeffreys prior distribution for parameters γ_1 and γ_2. As the prior associated to γ_1 is centered around large values, the most probable prior values for the vector (γ_1, γ_2) are associated to large values for γ_1 and small values for γ_2.

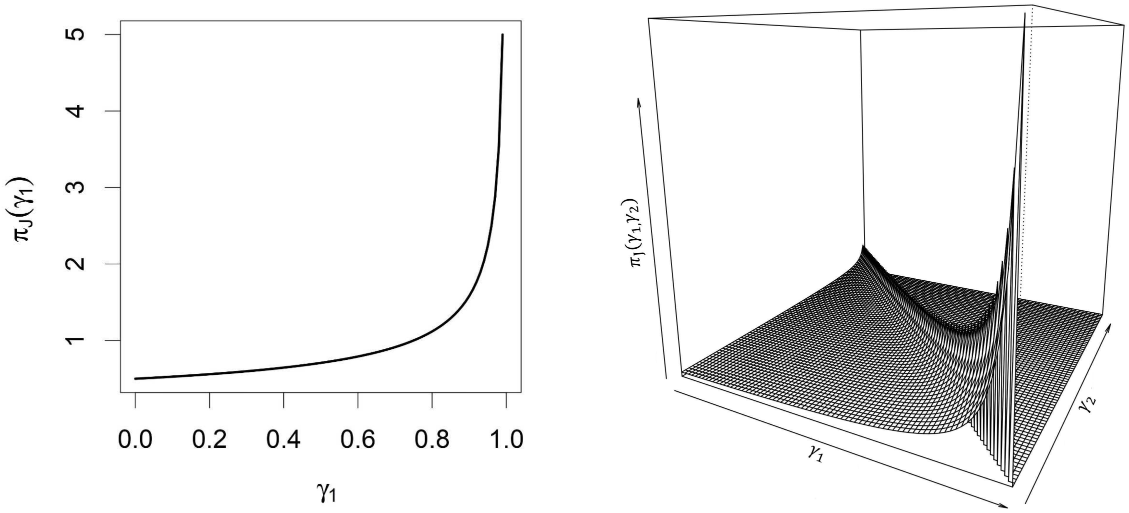


Figure 2.1: Marginal Jeffreys prior for γ_1 (left) and the joint Jeffreys prior for γ_1 and γ_2 (right).

2.2.2.2 Modeling the prior uncertainty about $\boldsymbol{\theta}$

To model the uncertainty about the relative risk $\boldsymbol{\theta}$, assume that p covariates are available such that $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$, $i = 1, \ldots, A$. The intercept β_0 represents a common term affecting the risk of all areas with prior $N(0, \sigma^2_{\beta_0})$. To model the prior uncertainty about the fixed effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, we assume that $\boldsymbol{\beta} \mid \boldsymbol{\Sigma}_\beta \sim \mathbf{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$, where $\mathbf{N}_p$ denotes the p-variate Gaussian distribution and $\boldsymbol{\Sigma}_\beta = \mathrm{diag}\{\sigma^2_1, \ldots, \sigma^2_p\}$. It is also appealing to consider some technique to perform Bayesian variable selection. The goal is to identify covariates that are statistically significant (non-zero effect) to explain the relative risks. The stochastic search variable selection (SSVS) method, proposed by George and McCulloch [1993], assigns a spike-slab mixture of Gaussian distributions to the fixed effects $\boldsymbol{\beta}$. The spike element concentrates closely around zero, reflecting whether the covariate should be included in the model. The slab component has a sufficiently large variance to allow the covariate effect to spread over larger values. Thus, to complete the SSVS prior specification we, additionally, assume that

$$\begin{aligned} \sigma^2_m \mid \omega_m, \sigma^2_{slab}, \sigma^2_{spike} &\overset{ind}{\sim} (1-\omega_m)\delta_{\sigma^2_{spike}}(\sigma^2_m) + \omega_m \delta_{\sigma^2_{slab}}(\sigma^2_m) \\ \omega_m \mid \rho_m &\overset{ind}{\sim} Bernoulli(\rho_m), \end{aligned} \tag{2.8}$$

where $\delta_x(\cdot)$ denotes the Kronecker delta concentrated at point x and the hyperparameters σ^2_{slab}, σ^2_{spike} and ρ_m, for $m = 1, \ldots, p$, should be specified (see example in Section 2.3).

To allow for local differences among the risks, apart from the covariates pattern, a more complete model with regional effects $\boldsymbol{u} = (u_1, \ldots, u_A)$ can be considered in the log-linear regression by assuming that $u_i \overset{iid}{\sim} \mathbf{N}(0, \sigma^2_u)$, $i = 1, \ldots, A$. Spatial effects $\boldsymbol{s} = (s_1, \ldots, s_A)$ that quantify the influence of neighboring areas can also be added into the regression structure such that $\log(\theta_i) = \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + s_i + u_i$. The usual joint prior distribution for $\boldsymbol{s}$ is the intrinsic conditional autoregressive distribution (ICAR) with variance parameter σ_s^2 (see Besag, York, and Mollié [1991] for details on the ICAR prior definition). We further assume that the model variance parameters are such that $\sigma_s^2 \sim \text{IG}(a_s, d_s)$ and $\sigma_u^2 \sim \text{IG}(a_u, d_u)$, where IG denotes the Inverse-Gamma distribution. The parameters β_0, $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{s}$ are considered as being independent.

Assuming the prior distributions discussed in this section, the joint posterior distribution for all model parameters is not known in closed form. Posterior inference can be carried out through a Markov chain Monte Carlo (MCMC) scheme. The posterior full conditional distributions that can be considered for sampling from the joint posterior distribution are given in the Supplementary Material [Oliveira *et al.* Supplement, 2020].

2.3 Simulated data studies

In this section, we investigate the performance of the proposed model through Monte Carlo simulations. To mimic our case study presented in Section 2.4, we consider the map of Minas Gerais State that is composed of $A = 75$ areas. A total of 100 datasets are generated from Poisson distributions such that $Y_i \overset{ind}{\sim} \mathcal{P}oisson(E_i\theta_i\epsilon_i)$, for $i = 1, \ldots, 75$, where $\epsilon_i = 1 - \mathbf{h}_i^T\boldsymbol{\gamma}$ and the expected number of cases E_i is known and equal to the one available for the case study. We also consider the same clustering indicator variable used in the case study, which has $K = 4$ data quality categories (clusters), partitioning the map in groups with a total of 28, 16, 14 and 17 areas, respectively, from the best to the worst category. This clustering variable is based on the adequacy index (AI) introduced by França *et al.* [2006]. We provide a detailed explanation of the clustering construction in Section 2.4. We set $\boldsymbol{\gamma} = (0.05, 0.10, 0.15, 0.20)$ imposing that 5% of events are not reported in those areas classified at the highest level of data quality whereas only 50% of events are reported in those areas belonging to the worst data quality cluster. To generate the relative risks, we consider independent observations from five covariates such that $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_5 X_{5i}$, where $\beta_0 = 0.50$ and $\boldsymbol{\beta} = (-0.25, -0.25, 0, 0, 0.25)$. These covariates are different from the clustering covariate. They were selected from our real dataset such that part of them are correlated with the clustering covariate. All covariates considered here are provided in the Supplementary Material [Oliveira *et al.* Supplement, 2020].

When fitting the simulated datasets, three different structures are considered for the relative risk $\boldsymbol{\theta}$. In Model 1, we let $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_5 X_{5i}$, where $\beta_m \overset{iid}{\sim} N(0, 100)$ for $m = 0, \ldots, 5$. Model 2 differs from Model 1 by considering a variable selection scheme on the set of covariates through the SSVS prior distribution given in expression (2.8) with $\sigma^2_{spike} = 0.001$, $\sigma^2_{slab} = 100$ and $\rho_m = 0.5$ for $m = 1, \ldots, 5$. Model 3 differs from Model 2 by the inclusion of both local and spatially structured random effects in the log-linear regression such that $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_5 X_{5i} + u_i + s_i$, where $u_i \overset{iid}{\sim} \mathbf{N}(0, \sigma_u^2)$ is the local effect of area i and $\boldsymbol{s} = (s_1, \ldots, s_A)$ denotes the spatial effects having the ICAR prior distribution [Besag, York, and Mollié, 1991] with precision parameter $\tau_s = \sigma_s^{-2}$. The neighboring structure inherent to the map of case study in Section 2.4 is adopted to model the spatial effects $\boldsymbol{s}$ and we further assume that the model precision parameters are modeled as $1/\sigma_s^2 \sim \text{Gamma}(0.5, 0.0005)$ and $1/\sigma_u^2 \sim \text{Gamma}(2, 0.01)$.

The prior specification for $\boldsymbol{\gamma}$ differs throughout the simulation studies and it will be properly described in each case. Basically, the joint prior distributions given in expressions (2.5) and (2.6) are elicited with different levels of information, specially focusing on the prior distribution for the parameter γ_1 which is associated with the model identifiability.

Posterior estimates (posterior means) for the relative risks, $\boldsymbol{\theta}$, are compared in terms of bias (Bias), relative mean squared error (RMSE) and the nominal coverage of the 95% highest posterior density intervals (Cov.) averaged over the $R = 100$ Monte Carlo replications. Specifically, the $bias = \left[\sum_{r=1}^{R}\sum_{i=1}^{A}\left(\hat{\theta}_{i,r} - \theta\right)\right]/(R \times A)$ and $RMSE = \left[\sum_{r=1}^{R}\sum_{i=1}^{A}\left(\frac{\hat{\theta}_{i,r}-\theta}{\theta}\right)^2\right]/(R \times A)$. All simulations were performed in OpenBUGS (available at http://www.openbugs.net/w/FrontPage) through the *rbugs* package from software R [R Core Team , 2015]. A sample of the BUGS code is provided in the Supplementary Material [Oliveira *et al.* Supplement, 2020]. For each generated dataset, the MCMC scheme considered a total of 100,000 iterations, being the first 50,000 discarded as a burn-in period and a lag of 25 iterations was selected to avoid autocorrelated posterior samples.

2.3.1 Simulation Study I: comparing the generalized Beta and the conditional uniform priors for $\boldsymbol{\gamma}$

In this study, we mainly evaluate the sensitivity of the posterior estimates of $\boldsymbol{\theta}$ when different degrees of information are assumed in the prior distributions for $\boldsymbol{\gamma}$ defined in expressions (2.5) and (2.6). In both cases, two different levels of prior information, named partially informative and fully informative, are considered. The partially informative case assumes an informative prior only for the parameter γ_1. Here, that is attained by choosing hyperparameters such that the prior $\pi(\gamma_1)$ is centered and highly concentrated around the true value of γ_1. We elicited $\gamma_1 \sim GB(2.9, 55.1; 0, 1)$ under the generalized Beta prior and $\gamma_1 \sim U(0, 0.10)$ under the conditional uniform prior. For all remaining γ_j, $j = 2, \ldots, 4$, the associated prior distribution assumes $a_j = 0$, $a_j^* = 1$ and, additionally for the generalized Beta case, it also considers $\alpha_j = \nu_j = 1$. By doing so, we impose a strong constraint on the reporting probability associated to areas belonging to the best data quality cluster but, for all the remaining areas, the only prior information is the one inherited from the prior of γ_1. Finally, in the case of fully informative prior distributions, all hyperparameters a_j and a_j^* and, additionally α_j and ν_j^* in the generalized Beta case, are chosen such that $\pi(\gamma_j)$ is centered and highly concentrated around the true value of γ_j, for $j = 1, \ldots, 4$. For comparison purposes, we also consider the standard Poisson model which does not take underreporting into account.

Table 2.1 summarizes the results. By eliciting an informative prior distribution only for pa-

Table 2.1: Bias, relative mean squared error (RMSE) and nominal coverage of 95% credible intervals (Cov.) for the estimated relative risks $\boldsymbol{\theta}$; Simulation Study I.

	RMSE	Bias	Cov.	RMSE	Bias	Cov.
	proposed model with generalized Beta prior on $\boldsymbol{\gamma}$					
	partially informative			fully informative		
Model 1	0.001	0.004	0.989	0.001	−0.004	0.991
Model 2	0.001	0.004	0.993	0.001	−0.003	0.993
Model 3	0.002	0.002	0.997	0.002	−0.003	0.997
	proposed model with conditional uniform prior on $\boldsymbol{\gamma}$					
	partially informative			fully informative		
Model 1	0.001	−0.001	0.988	0.001	−0.001	0.989
Model 2	0.001	−0.001	0.993	0.001	−0.002	0.992
Model 3	0.002	−0.003	0.997	0.002	−0.003	0.996
	standard Poisson model (underreporting ignored)					
Model 1	0.069	−0.622	0.069	–	–	–
Model 2	0.069	−0.621	0.106	–	–	–
Model 3	0.076	−0.626	0.424	–	–	–

rameter γ_1 (partially informative case), the proposed model provides good posterior estimates for the risks with bias and RMSE close of zero. The results are quite close to those obtained under informative prior for all components of parameter vector $\boldsymbol{\gamma}$ (fully informative case). In general, we observe a slight difference between results obtained under the generalized Beta prior and the conditional uniform distributions for $\boldsymbol{\gamma}$, where the former has a greater number of hyperparameters to be chosen. Results under Models 1–3 are quite similar showing that spatial and local effects do not significantly influence the posterior inferences. This is an interesting result as the data are generated from a model that does not include any spatial or local correlation. It should be also mentioned that the non-significant (null) effect of covariates X_3 and X_4 (results not shown) is well identified even under Model 1 which does not consider variable selection.

Table 2.1 also shows that, as expected, the standard Poisson model fails in estimating the relative risks, $\boldsymbol{\theta}$, whenever applied to analyze underreported data. It produces very poor estimates, always underestimating the relative risks no matter the structure imposed to model them. The RMSE under such a model is reasonably small but the 95% credible intervals do not contain the true value of the relative risk for the great majority of the Monte Carlo replications, which means that the posterior distribution for $\boldsymbol{\theta}$ tends to put negligible probability mass around its true value.

2.3.2 Simulation Study II: effect of the prior uncertainty about γ_1

The prior distribution for parameter γ_1 plays an important role in model identification and, consequently, in the quality of the posterior estimates. In this section, we reexamine the datasets considered in Section 2.3.1 fitting the proposed model with different partially informative prior distributions for $\boldsymbol{\gamma}$, that is, an informative prior distribution is considered only for the component γ_1. A sensitivity analysis is performed in order to evaluate the influence of such prior distribution on the posterior inference.

The evaluation metrics for the posterior estimates of $\boldsymbol{\theta}$ under six different conditional uniform priors for γ_1 (Table 2.2) show that the relative risks tend to be underestimated if, *a priori*, we elicited $\gamma_1 \sim U(0.0, 0.01)$ and $\gamma_1 \sim U(0.0, 0.05)$. Such prior distributions put all probability mass below 0.05 which is the true value of γ_1. On the other hand, the risks tend to be overestimated whenever the prior expectation exceeds the true value of γ_1. The highest the difference between the prior expectation $E(\gamma_1)$ and the true value of γ_1, the highest are the bias and RMSE of the posterior estimates of $\boldsymbol{\theta}$. This is not a surprising result and it evidences the importance of searching for reliable prior information about parameter γ_1 in practical situations.

Table 2.2 also shows that, if we assume $\gamma_1 \sim U(0, 0.05)$ or $\gamma_1 \sim U(0, 0.15)$, the prior means differ from the true value of γ_1 by the same amount. Although the latter prior imposes much higher prior variance than the former, the posterior estimates present similar absolute values for the bias and the RMSE in both cases. This suggests that quality of posterior estimates under the proposed model are more strongly related to the prior expectation of γ_1 than to its prior variance. Such an idea is supported by the results in Table 2.3 which exhibits some evaluation metrics related to posterior inference for $\boldsymbol{\theta}$ assuming different partially informative generalized Beta prior distributions for $\boldsymbol{\gamma}$. In all cases, $\gamma_1 \sim GB(\alpha_1, \nu_1; 0, 1)$ where hyperparameters α_1 and ν_1 are chosen such that this prior is centered around the true value of γ_1, that is, $E(\gamma_1) = 0.05$, but the prior uncertainty about γ_1 varies from 0.00002 to 0.00950.

Table 2.2: Bias and relative mean squared error (RMSE) for the estimated relative risks $\boldsymbol{\theta}$ assuming partially informative conditional uniform priors to $\boldsymbol{\gamma}$ with six levels of prior information on γ_1 ($E(\gamma_1)$ and $V(\gamma_1)$ are different in all cases); Simulation Study II.

	RMSE	Bias	RMSE	Bias	RMSE	Bias
	$\gamma_1 \sim$ **U(0.0,0.01)**		$\gamma_1 \sim$ **U(0.0,0.05)**		$\gamma_1 \sim$ **U(0.0,0.15)**	
Model 1	0.004	−0.091	0.002	−0.050	0.002	0.055
Model 2	0.003	−0.090	0.002	−0.050	0.002	0.047
Model 3	0.004	−0.093	0.002	−0.053	0.002	0.049
	$\gamma_1 \sim$ **U(0.0,0.30)**		$\gamma_1 \sim$ **U(0.0,0.50)**		$\gamma_1 \sim$ **U(0.0,0.70)**	
Model 1	0.014	0.225	0.062	0.440	0.137	0.570
Model 2	0.015	0.227	0.065	0.467	0.112	0.527
Model 3	0.014	0.215	0.078	0.494	0.240	0.766

Table 2.3 shows that the RMSE approaches zero in all cases. As expected, the bias tends to increase as the prior uncertainty about γ_1 increases. If the generalized Beta prior distributions with $V(\gamma_1) = 0.00024$ and $V(\gamma_1) = 0.00226$ are assumed, the biases are much smaller than those observed in Table 2.2 under priors $\gamma_1 \sim U(0.0, 0.05)$ and $\gamma_1 \sim U(0.0, 0.15)$ whose variances are similar (respectively, $V(\gamma_1) = 0.00021$ and $V(\gamma_1) = 0.00188$). Moreover, the bias and RMSE under the prior $U(0.0, 0.30)$, which has variance equal to 0.0075, are much higher than those obtained when assuming a generalized Beta prior with a variance equal to 0.0095. In summary,

these findings provide more evidence that the posterior inference is more influenced by the prior expectation of γ_1 than by its prior variance.

Table 2.3: Bias and relative mean squared error (RMSE) for the estimated relative risks $\boldsymbol{\theta}$ assuming partially informative generalized Beta priors to $\boldsymbol{\gamma}$ with six distinct levels of information on γ_1 ($E(\gamma_1) = 0.05$ (true γ_1) and a different prior variance in each case); Simulation Study II.

	RMSE	Bias	RMSE	Bias	RMSE	Bias
	$\mathbb{V}(\gamma_1) = 0.00002$		$\mathbb{V}(\gamma_1) = 0.00024$		$\mathbb{V}(\gamma_1) = 0.00083$	
Model 1	0.001	0.002	0.002	0.003	0.001	0.005
Model 2	0.001	0.000	0.001	0.002	0.001	0.005
Model 3	0.002	0.001	0.002	0.000	0.002	0.002
	$\mathbb{V}(\gamma_1) = 0.00144$		$\mathbb{V}(\gamma_1) = 0.00226$		$\mathbb{V}(\gamma_1) = 0.00950$	
Model 1	0.001	0.006	0.001	0.007	0.002	0.017
Model 2	0.001	0.006	0.001	0.009	0.002	0.028
Model 3	0.002	0.004	0.002	0.007	0.002	0.026

Table 2.4 exhibits the averaged posterior means for parameters β_0, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ under three out of the different partially informative conditional uniform prior distributions for γ_1 considered in previous studies. Results for Models 1–3 are quite similar, thus we only present the results obtained under Model 3. The vector of fixed effects $\boldsymbol{\beta}$ and variable selection parameter $\boldsymbol{\omega}$ are well estimated regardless of the prior distribution elicited for γ_1 but very little is learned about γ_1 from the data. The posterior mean of γ_1 tends to be close to its prior expectation, reinforcing the importance of obtaining reliable prior information about this parameter. Posterior estimates for the remaining components of $\boldsymbol{\gamma}$ become worse as the prior expectation of γ_1 gets far from its true value and the prior variance of γ_1 increases.

Table 2.4: Averaged posterior means of β_0, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ under three different prior specifications on parameter γ_1; Simulation Study II.

Parameter	True Value	$U(0, 0.01)$		$U(0, 0.10)$		$U(0, 0.70)$	
		Mean	$\hat{\omega}$	Mean	$\hat{\omega}$	Mean	$\hat{\omega}$
β_0	**0.500**	**0.450**	–	**0.500**	–	**0.780**	–
β_1	−0.250	−0.250	1.000	−0.250	1.000	−0.250	1.000
β_2	−0.250	−0.260	1.000	−0.260	1.000	−0.260	1.000
β_3	0.000	0.000	0.020	0.000	0.020	0.000	0.020
β_4	0.000	0.000	0.030	0.000	0.030	0.000	0.030
β_5	0.250	0.240	0.990	0.240	0.990	0.240	0.990
γ_1	**0.05**	**0.005**	–	**0.048**	–	**0.261**	–
γ_2	0.100	0.103	–	0.099	–	0.077	–
γ_3	0.150	0.155	–	0.148	–	0.114	–
γ_4	0.200	0.211	–	0.202	–	0.156	–

Goodness of posterior estimation for parameters β_0 and γ_1 are closely related, which is not a surprising result given the identifiability issues discussed in Section 2.2.1. The intercept β_0 is overestimated (resp., underestimated) if γ_1 is also overestimated (resp., underestimated). Since

β_0 directly affects the estimation of the relative risks $\boldsymbol{\theta}$, by overestimating (resp., underestimating) β_0, the relative risks $\boldsymbol{\theta}$ is overestimated (resp., underestimated) inducing the larger positive (resp., negative) bias shown in Table 2.2.

2.3.3 Simulation Study III: breaking the identification constraints

Our goal here is to show the effect of using the same source of information to model both the relative risk $\boldsymbol{\theta}$ and the reporting probability $\boldsymbol{\epsilon}$. We consider two different scenarios. In the first one, the same covariate is present in both sets $\boldsymbol{X}$ and $\boldsymbol{H}$. Consequently, as the constraints for the model identification are not fulfilled, we should have problems to estimate the model parameters. In the second scenario, we will use the same variable but coded in two different ways: In $\boldsymbol{X}$ it is continuous while for $\boldsymbol{H}$ it is considered in a discretized scale obtained by breaking its continuous range into four intervals and coding them with dummy variables. In this case, despite the very strong correlation between $\boldsymbol{X}$ and $\boldsymbol{H}$, we should obtain good posterior estimates for all model parameters.

We consider the same four clusters used in the previous simulation studies, which are based on a variable called adequacy index (AI) available in our case study (Section 2.4). In the first scenario, named Categorical AI, the variable AI is considered in its discretized version with four categories indicating the clusters and the variable AI enter in this discretized form in both $\boldsymbol{X}$ and $\boldsymbol{H}$. In the second scenario, named Continuous AI, its discretized version is maintained in $\boldsymbol{H}$ but, for $\boldsymbol{X}$, we consider the original continuous AI re-scaled to have a zero mean and a unit standard deviation. To generate the datasets, we set $\boldsymbol{\gamma} = (0.05, 0.10, 0.15, 0.20)$ and assume the covariates $X_{1i}, \ldots, X_{4i}$ as in the previous studies. In the Continuous AI scenario, we let $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_5 X_{5i}$, where X_{5i} is the AI in its continuous scale, $\beta_0 = 0.15$ and $\boldsymbol{\beta} = (-0.25, -0, 25, 0, 0, -0.25)$. In the Categorical AI scenario, we assume $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_4 X_{4i} + \beta_{5,1} D_{1i} + \cdots + \beta_{5,3} D_{3i}$, where $\beta_0 = 0.15$ and $\boldsymbol{\beta} = (-0.25, -0.25, 0, 0, 0.25, 0.50, 0.75)$. The dummy variable D_{li} represents the lth level of the discretized AI for $l = 1, 2, 3$. To analyze the data, we consider the partially informative conditional uniform prior for $\boldsymbol{\gamma}$ in which $\gamma_1 \sim U(0, 0.10)$.

As expected, Table 2.5 shows that the posterior inferences for the relative risks are much worse if we break the identifiability constraints (Categorical AI case). However, such estimates do not lose quality if we consider strongly correlated variables to model $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ (Continuous AI case). In the Categorical AI case, Table 2.6 shows confounding between the parameters $\boldsymbol{\gamma}$ and the effects of the dummy variables, being all these parameters poorly estimated. This problem is not experienced by the parameters in the Continuous AI case. These findings are in perfect agreement with the theoretical identifiability results discussed in Section 2.2.1.

Table 2.5: Bias, relative mean squared error (RMSE) and nominal coverage of 95% credible intervals (Cov.) for the estimated relative risks $\boldsymbol{\theta}$; Simulation Study III.

	RMSE	Bias	Cov.	RMSE	Bias	Cov.
	Continuous AI			Categorical AI		
Model 1	0.002	−0.009	0.982	5.789	3.785	0.880
Model 2	0.002	−0.009	0.986	12.654	4.207	0.885
Model 3	0.002	−0.010	0.995	11.489	4.670	0.823

Table 2.6: Averaged posterior means of β_0, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ under Model 2; Simulation Study III.

	Continuous AI			Categorical AI		
Parameter	True Value	Posterior Mean	$\hat{\omega}$	True Value	Posterior Mean	$\hat{\omega}$
β_0	0.150	0.141	–	0.150	0.146	–
β_1	−0.250	−0.248	1.000	−0.250	−0.248	1.000
β_2	−0.250	−0.253	1.000	−0.250	−0.251	1.000
β_3	0.000	−0.001	0.018	0.000	0.002	0.016
β_4	0.000	0.002	0.018	0.000	0.003	0.020
β_5	−0.250	−0.255	0.999	**0.250**	**0.507**	0.938
β_6	–	–	–	**0.500**	**1.144**	0.997
β_7	–	–	–	**0.750**	**1.811**	0.996
γ_1	0.050	0.048	–	0.050	0.048	–
γ_2	0.100	0.093	–	**0.100**	**0.256**	–
γ_3	0.150	0.152	–	**0.150**	**0.272**	–
γ_4	0.200	0.196	–	**0.200**	**0.182**	–

2.3.4 Comments on further simulation studies

Section S.3 of the Supplementary Material [Oliveira *et al.* Supplement, 2020] presents additional simulation studies exploring other features of the proposed model. In the following, we present the main results obtained from such studies. A discussion about the misspecification of the number of data quality clusters, K, is provided in Section S.3.1 of the Supplementary Material. In summary, for the simulated datasets, we note that the misspecification of K introduces more bias as well as higher variability in the posterior estimates of $\boldsymbol{\theta}$. Both bias and RMSE are much higher if the number of clusters assumed when fitting the proposed model is smaller than the true value of K if compared with the case of assuming a value for K that is greater than the true one.

We also evaluate whether the number of areas within the best and worst data quality clusters significantly affects the posterior inference for the relative risks $\boldsymbol{\theta}$ (Section S.3.2 of the Supplementary Material). In summary, we observed that having a greater number of areas within the best data quality cluster decreases the bias in the posterior estimates of $\boldsymbol{\theta}$. This is an expected behavior since, whenever the number of areas within the best group is larger, the model induces an informative prior for a greater number of areas.

Finally, from the study presented in Section S.3.3 of the Supplementary Material, we note that, if the data are correctly recorded (that is, assuming $\epsilon_i = 1 \; \forall \; i$), the relative risks $\boldsymbol{\theta}$ are overestimated under our approach and the bias magnitude depends on the prior knowledge about $\boldsymbol{\gamma}$ (see Table 2.10). In this context, as expected, the standard Poisson model performs very well presenting both bias and RMSE close to zero. However, the standard Poisson model always underestimates the relative risks if counts are partially recorded (see Table 2.1), and the bias magnitude depends on the amount of underreporting in the data. Therefore, it is important mentioning that the proposed model shows better results whenever fitted to analyze perfectly recorded data (in terms of bias and RMSE) than the standard Poisson model does whenever fitted to analyze underreported data. In practical situations, the relative risk estimates may guide the definition of government policies for control and intervention. Thus, the underestimation of such quantities leads to undesirable consequences, for instance, if we are mapping disease or mortality risks.

2.4 Early neonatal mortality data in Minas Gerais, Brazil

Our goal here is to map the relative risk of early neonatal mortality (ENM) in Minas Gerais State (MG), Brazil, and also to identify factors that are possibly associated to the event occurrence. The ENM refers to the deaths occurring in the first seven days of life. Quality of infant mortality information produced in MG is usually underreported [Campos, Loschi and França, 2007], mainly in the socio-economically more deprived areas located in northern and northeastern regions of the state. In order to define efficient policies to diminish the number of early neonatal deaths and properly distribute the financial resources, it is important to correctly estimate the associated risks.

The counts were obtained from the *Sistema de Informações sobre Mortalidade* (SIM) and *Sistema de Informações sobre Nascidos Vivos* (SINASC) from the National Health System of the Brazilian Ministry of Health (BMH). The 853 municipalities of MG were grouped in $A = 75$ microregions (areas) following the official division suggested by the BMH. Two periods of time comprising the two most recent Brazilian Demographic Censuses are considered, namely, 1999–2001 and 2009–2011.

To analyze the datasets, we fit the proposed model assuming that Y_i and E_i are, respectively, the observed and the expected counts of ENM at area $i = 1, \ldots, 75$. We assume $Y_i \mid \theta_i, \epsilon_i \overset{ind}{\sim} \mathcal{P}oisson(E_i \theta_i \epsilon_i)$ for all i. We consider the usual naive estimator for the offset E_i given by $E_i = n_i \left(\sum_{i=1}^{A} y_i / \sum_{i=1}^{A} n_i \right)$, where n_i represents the total number of newborn children at risk in the ith area and y_i is the observed count of early neonatal deaths in such area. For comparison purposes, we also fit the standard Poisson model which ignores the underreporting in its structure by assuming $\epsilon_i = 1$ for all areas.

The ENM relative risk assumes a log-linear regression structure which includes local and spatial random effects, that is, $\log(\theta_i) = \beta_0 + \mathbf{X}_i \boldsymbol{\beta} + u_i + s_i$, $i = 1, \ldots, 75$. Five covariates are introduced in this regression model: the Municipal Human Development Index (MHDI), the proportion of mothers with more than twelve years of formal education (MomEduc), the proportion of children with weight at birth smaller than 2.5 Kg (LowWeight), the proportion of children who were born with some congenital anomaly (Anomaly) and the proportion of mothers who made seven or more prenatal visits during the pregnancy (Prenatal). The MHDI was collected from the Atlas of Human Development in Brazil (2010) and the other four covariates were obtained from the DATASUS repository, maintained by the BMH.

To define the clustering structure, we consider the adequacy index (AI) introduced by França *et al.* [2006] as a measure of the quality of infant mortality data collected in Minas Gerais. Based on the adequacy index, França *et al.* [2006] proposed a partition of the $A = 75$ microregions of MG into $K = 4$ groups: MG1 (most adequate, $AI > 70.0$, 28 microregions), MG2 (group intermediate A, $50.1 < AI < 70.0$, 16 microregions), MG3 (group intermediate B, $20.0 < AI < 50.0$, 14 microregions) and MG4 (less adequate group, $AI < 20.0$, 17 microregions). We consider these four groups to analyze the ENM data in both periods, 1999–2001 and 2009–2011. Since there is an expectation of improved data reporting quality in recent years, the $K = 4$ clusters induced by this partition may be more heterogeneous in the period 1999–2001. In order to provide a sensitivity analysis and also attempting to reduce the effect of within cluster heterogeneity, we divide each of the previous groups in two new groups obtaining another clustering structure with $K = 8$ categories of data quality. The median of the AI within each of the four initial groups is considered for defining the new partition into eight groups. Panels (b) and (d) of Figure 2.2 display the groups defined in both cases (each color corresponds to a different group).

2.4.1 About prior elicitation

To complete the model specification a prior distribution must be elicited for each parameter, with special attention to the informative prior needed for parameter γ_1. According to experts' opinion, the reporting probability in areas experiencing the best data quality likely approaches one. Based on the information obtained from the specialists (local epidemiologists and health researchers) for both periods of interest, we adopt the conditional uniform prior distribution given in expression (2.6) eliciting an informative prior distribution only for parameter γ_1 (partially informative prior distribution). When considering the clustering structure with $K = 4$ data quality groups, we set $\gamma_1 \sim U(0, 0.10)$ for period 1999–2001 and, as an improvement on data reporting quality is expected in more recent years, for period 2009–2011 it is assumed $\gamma_1 \sim U(0, 0.05)$. When fitting the data with $K = 8$ clusters, we set the prior $\gamma_1 \sim U(0, 0.05)$ for both periods.

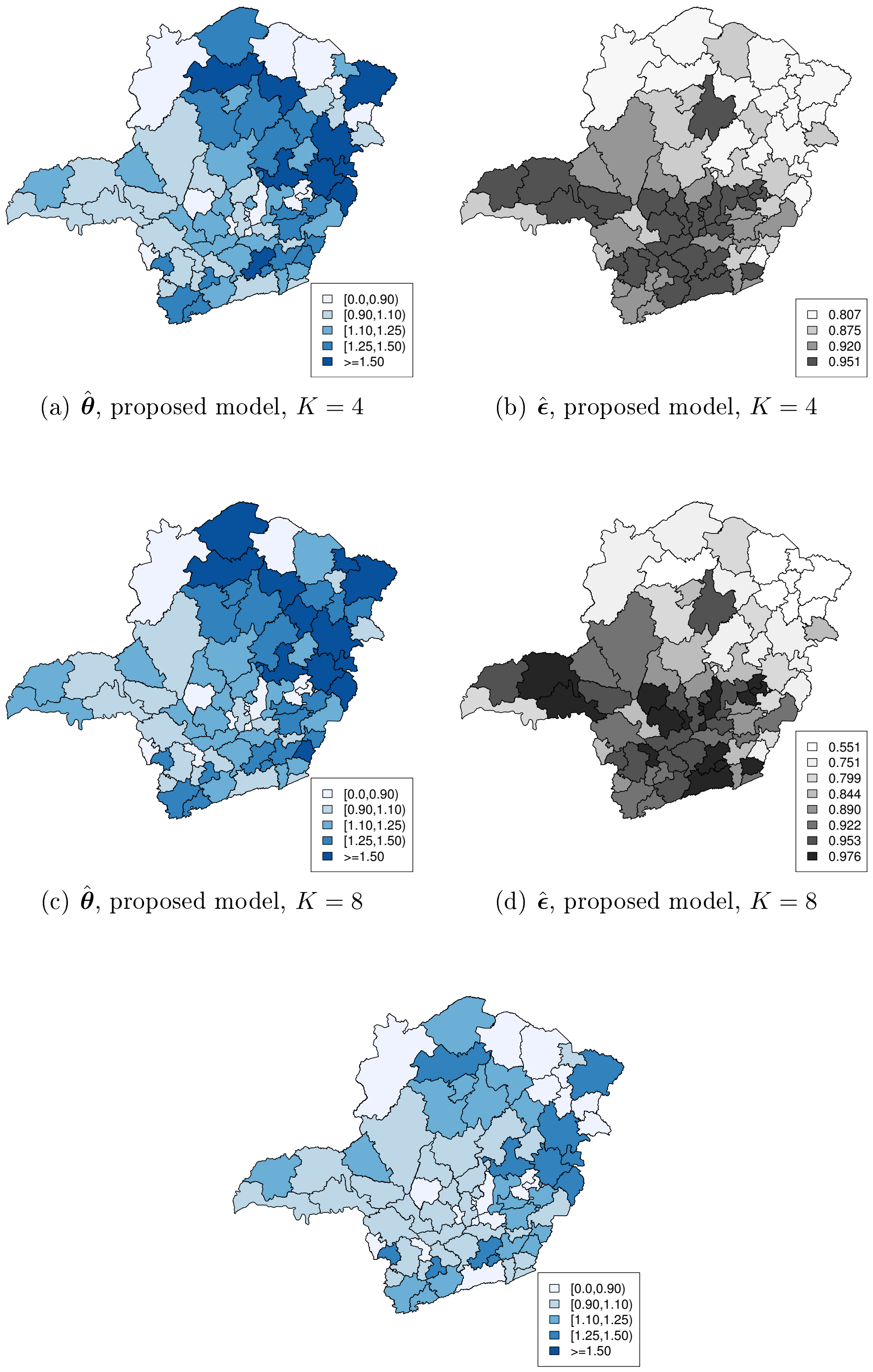
To model the prior uncertainty about the relative risks, $\boldsymbol{\theta}$, we assume the structure of Model 3 described in the simulation studies (Section 2.3). We set $\beta_0 \sim \mathrm{N}(0, 100)$ and perform a variable selection by eliciting the SSVS prior given in expression (2.8) for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)$ with $\sigma^2_{slab} = 100$, $\sigma^2_{spike} = 0.001$ and $\rho_m = 0.5$, $m = 1, \ldots, 5$. For parameters $\boldsymbol{s}$, $\boldsymbol{u}$, σ^2_s and σ^2_u we assume the prior distributions elicited in the simulated studies (Section 2.3). Also, for the MCMC performed in OpenBUGS, we consider the same specifications as in the simulated studies. The complete dataset and the BUGS code considered in this case study are provided in the Supplementary Material [Oliveira *et al.* Supplement, 2020].

2.4.2 Posterior results

Figures 2.2 and 2.3 show the posterior estimates of the ENM risks in MG for periods 1999–2001 and 2009–2011, respectively. By fitting the proposed model, we estimate the probability of recording the events in each area, see Panels (b) and (d) of Figures 2.2 and 2.3. Panel (d) of Figure 2.2 show that, for the period 1999–2001, the posterior mean for the probability of recording an early neonatal death at areas with the worst data quality is 0.551. Such estimate increases to 0.806 in the period 2009–2011 (Panel (d), Figure 2.3) indicating an improvement in the data reporting process in North and Northeast areas. The same occurred for the other areas showing that an improvement in data reporting process spread out over the state. For those areas classified in the best data quality cluster, the estimated reporting probability tends to be close in both periods, which is expected as the posterior estimate for parameter ϵ_i in the best group is quite influenced by its prior mean (see the discussion in Sections 2.2.1 and 2.3.2).

Posterior estimates for the relative risks under the standard Poisson model are displayed in Panel (e) of Figures 2.2 and 2.3. For the period 1999–2001 (Figure 2.2), such estimates shows that areas in the North and Northeast regions of Minas Gerais experienced the lowest ENM risks, being smaller than the risk obtained for Belo Horizonte city, the capital of the Minas Gerais State. This finding goes against the results obtained in some epidemiological studies that relate the quality of data to socioeconomic and access to health care indicators (e.g., Campos, Loschi and França [2007]). Because the North and Northeast regions are the poorest and present the lowest socio-educational indicators in the state, experts believe that the ENM risks in such areas are much higher than those estimated through the standard Poisson model, evidencing the incapacity of this model to account for a high underregistration level. In relation to the most recent period 2009–2011 (Figure 2.3), the spatial distribution of the posterior estimates provided by the standard Poisson model are more compatible to what is expected by the specialists. The posterior estimates for the ENM risks in the poorest areas (North and Northeast) are higher than the ones obtained for more developed regions of Minas Gerais. It points to an improvement in the quality of the data reporting process as indicated by the estimates for the reporting probabilities obtained under the proposed model in both periods. Moreover, compared to the estimates for period 1999–2001 (Figure 2.2), the ENM risks for most regions in South and Southwest of Minas Gerais decrease by 2009–2011. These results are possibly indicating the advance in the socio-economic conditions and the access to health care in Minas Gerais.

Panels (a) and (c) of Figures 2.2 and 2.3 show that the proposed model provides estimates for the ENM risks in Minas Gerais that are more compatible with the findings in Campos, Loschi and França [2007], especially in northeastern areas for both periods. Its performance is



(a) $\hat{\boldsymbol{\theta}}$, proposed model, $K = 4$

(b) $\hat{\boldsymbol{\epsilon}}$, proposed model, $K = 4$

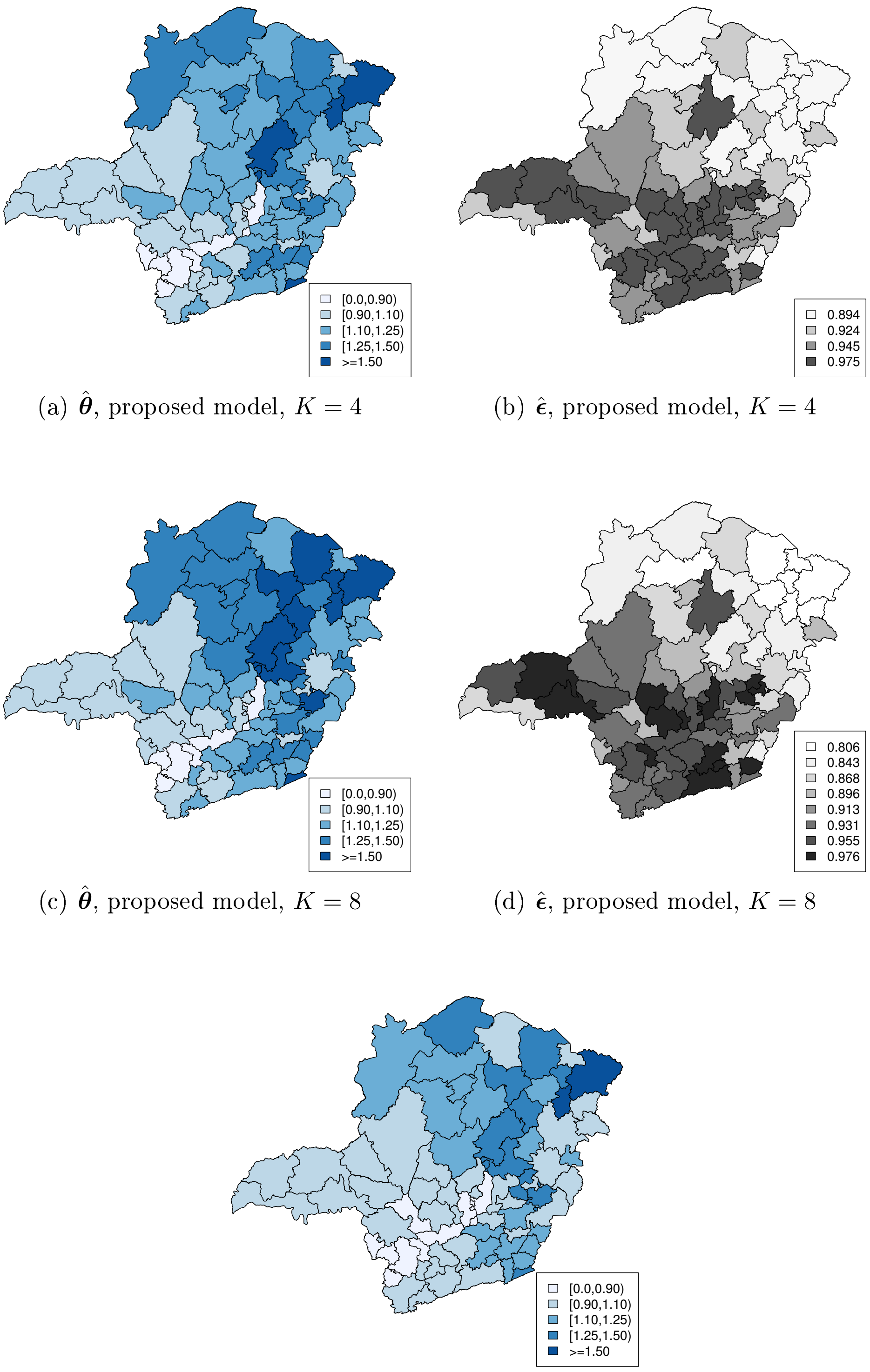
(c) $\hat{\boldsymbol{\theta}}$, proposed model, $K = 8$

(d) $\hat{\boldsymbol{\epsilon}}$, proposed model, $K = 8$

(e) $\hat{\boldsymbol{\theta}}$, standard Poisson model

Figure 2.2: Posterior mean for the relative risks, $\boldsymbol{\theta}$, of early neonatal mortality (Panels (a) and (c)) and the reporting probabilities, $\boldsymbol{\epsilon}$, (Panels (b) and (d)) under the proposed model with $K = 4$ (Panels (a) and (b)) and $K = 8$ (Panels (c) and (d)) and the standard Poisson model (Panel (e)); Minas Gerais data, period 1999-2001.

specially good when estimating the ENM risks in the period 1999–2001, in which data quality is more questionable. By accounting for underreporting, the proposed model corrects at least part of the underestimation experienced by the poorest microregions of the state providing



(a) $\hat{\boldsymbol{\theta}}$, proposed model, $K = 4$

(b) $\hat{\boldsymbol{\epsilon}}$, proposed model, $K = 4$

(c) $\hat{\boldsymbol{\theta}}$, proposed model, $K = 8$

(d) $\hat{\boldsymbol{\epsilon}}$, proposed model, $K = 8$

(e) $\hat{\boldsymbol{\theta}}$, standard Poisson model

Figure 2.3: Posterior mean for the relative risks, $\boldsymbol{\theta}$, of early neonatal mortality (Panels (a) and (c)) and the reporting probabilities, $\boldsymbol{\epsilon}$, (Panels (b) and (d)) under the proposed model with $K = 4$ (Panels (a) and (b)) and $K = 8$ (Panels (c) and (d)) and the standard Poisson model (Panel (e)); Minas Gerais data, period 2009-2011.

more realistic estimates for the ENM risks in such areas. As expected, for areas experiencing a good data quality, estimation under both the proposed and the standard Poisson models are similar. As observed for the standard Poisson model, the maps for the ENM relative risks estimated under the proposed model in period 2009–2011 disclose a decrease in the risk for most microregions in South and Southwest of Minas Gerais if compared to period 1999–2001.

Table 2.7 summarizes the results under the fitted models. The log pseudo-marginal likelihood (LPML) criterion [Ibrahim, Chen, and Sinha, 2001] points that data from 1999–2001 are better fitted by the proposed model with $K = 8$ data quality clusters whereas for period 2009–2011 the proposed model with $K = 4$ provides the best data fit. The expected improvement in the quality of the data reporting process in the most recent period, 2009–2011, makes the microregions more homogeneous in relation to such data feature. Therefore, a smaller number of data quality categories is actually expected. For each period, only results related to the best fitted models are considered in the following analysis.

Table 2.7: Posterior summaries for the regression effects β_0 and $\boldsymbol{\beta}$ under proposed and standard Poisson models; Minas Gerais data in both periods 1999–2001 and 2009–2011. We provide the posterior mean (Mean), the standard deviation (St.Dev.), the posterior probability of being positive ($\mathbb{P}(\beta > 0)$) and the posterior inclusion probability ($\hat{\omega}$).

Covariate	Mean	St.Dev.	$\mathcal{P}(\beta > 0)$	$\hat{\omega}$	.Mean	St.Dev.	$\mathcal{P}(\beta > 0)$	$\hat{\omega}$
			proposed model with $K = 4$					
	1999–2001 (LPML = −334.107)				2009–2011 (**LPML = −281.833**)			
Intercept	0.834	0.410	0.989	–	1.402	0.632	0.998	–
MHDI	−1.592	0.647	0.000	1.000	−0.860	0.725	0.150	**0.670**
MomEduc	−0.398	1.144	0.428	0.250	−0.218	0.558	0.399	0.190
LowWeight	1.694	2.462	0.706	0.706	−0.688	1.435	0.380	0.274
Anomaly	2.653	7.731	0.604	0.534	3.685	6.588	0.687	**0.553**
Prenatal	0.080	0.216	0.630	0.159	−0.949	0.552	0.084	**0.791**
			proposed model with $K = 8$					
	1999–2001 (**LPML = −325.948**)				2009–2011 (LPML = −283.863)			
Intercept	1.986	0.181	1.000	–	1.946	0.300	1.000	–
MHDI	−3.369	0.311	0.000	**1.000**	−1.400	0.465	0.005	0.989
MomEduc	−0.033	0.615	0.491	0.128	−0.120	0.357	0.425	0.140
LowWeight	−0.095	0.592	0.483	0.168	−0.843	1.944	0.393	0.253
Anomaly	2.450	7.211	0.579	0.476	3.586	6.446	0.644	0.515
Prenatal	0.104	0.212	0.678	0.222	−1.170	0.306	0.000	1.000
			standard Poisson model					
	1999–2001 (LPML = −338.997)				2009–2011 (LPML = −286.665)			
Intercept	2.007	0.238	1.000	–	2.006	0.507	1.00	–
MHDI	−3.797	0.486	0.000	1.000	−1.686	0.837	0.044	0.894
MomEduc	−0.086	0.785	0.470	0.171	−0.058	0.279	0.444	0.091
LowWeight	0.545	1.374	0.587	0.294	−2.260	2.932	0.255	0.507
Anomaly	1.499	8.679	0.542	0.548	3.028	6.292	0.643	0.513
Prenatal	0.097	0.240	0.625	0.228	−0.934	0.507	0.050	0.865

Assuming that a covariate X_m, $m = 1, \ldots, 5$, should be included into the model whenever $\hat{\omega}_m \geq 0.5$, where $\hat{\omega}_m$ denotes the posterior estimate for the associated inclusion probability, then Table 2.7 shows that different sets of covariates are significant to explain the ENM risks in the two analyzed periods. Under the best models, only the covariate MHDI shows to be significant (likely non-zero effect) to explain the ENM risk for the period 1999–2001 while, for the period 2009–2011, MHDI, Anomaly and Prenatal were significant. As expected in practice, the effect of the covariate MDHI is negative in both periods, indicating that the highest the MHDI, the smallest the ENM risk. The effect of MHDI is smaller in the period 2009–2011. Also for this most recent period, we observe that the ENM risk is smaller in areas with a high proportion of mothers who have made seven or more prenatal visits during the pregnancy. Furthermore, the positive effect associated to the proportion of children who were born with some congenital anomaly (Anomaly) indicates that such characteristic has been an important factor to the occurrence of early neonatal deaths in recent years. Covariates MomEduc and LowWeight, usually pointed out as important factors to explain the infant mortality rate, are not significant in the best model for both periods considered in our study.

In closing, it is important to mention that the relative risk estimates provided by the proposed and the standard Poisson models are closer in the period 2009–2011 than their estimates obtained for the period 1999–2001. This is an evidence of improvement in the quality of the ENM data recorded in the civil registration systems SIM and SINASC in Minas Gerais State.

2.5 Discussion

We presented a novel Bayesian modeling framework to analyze potentially underreported count data. We propose a clustering scheme that relates the reporting probabilities among the areas according to a previous data quality partitioning. Auxiliary variables and experts' opinion can be considered to assess data quality throughout the areas. One interesting feature of the proposed model is that, to ensure its identifiability, only an informative prior for the underreporting probability in areas experiencing the best data quality is required. That is attractive because in the best areas information about the reporting probability tends to be easily accessed.

Naturally, some care should be taken as the posterior inference tends to be highly influenced by our prior specification for parameter γ_1, the underreporting probability in the best areas. In the simulation experiments, a sensitivity study involving different levels of prior information for γ_1 was performed. The results indicated that if the specified prior mean for γ_1 turns out to be widely different from the truth, then the bias correction is likely to be inaccurate. Therefore, in practical situations, it is truly relevant searching for reliable information about this particular prior distribution, especially the associated prior mean.

Our model was applied to correct the underreporting bias in a Brazilian neonatal mortality dataset. In this case, previous works guided the partitioning of the region according to the data quality experienced by its microregions. It is worth mentioning that in other case studies in which the clustering structure may not be previously available, one can apply usual clustering techniques to define the groups with basis on covariates related to the quality of the reporting system. In our application, some local epidemiologists and health researchers provided information about the reporting process in areas where data are known to be better recorded. This information is used to elicit the required informative prior distribution for γ_1. It is likely that a different prior specification in the neonatal mortality application might result in different inference on the reporting probabilities. Consequently, it also affects the bias correction on the mortality relative risks. However, the subjective nature of the solution for completely underreported data is not unique. In Bailey *et al.* [2005], for example, a different choice for the threshold used to define the censored areas can lead to different predictions. That may be also observed in the model introduced by Oliveira *et al.* [2017] if a different informative prior is elicited for the censoring probabilities. Also, in the approach proposed by Stoner, Economou, and Drummond [2019], a distinct prior specification to the mean reporting rate could lead to quite different posterior inference. The usage of a complete validation dataset (as, e.g., Whittemore and Gong [1991]; Stamey, Young and Boese [2006]; Dvorzak and Wagner [2015]) might be a less subjective approach depending on the quality, quantity and experimental design of collecting such data. In many cases, as the one analyzed here, the elicitation of an informative prior distribution for one parameter is a feasible and reasonable solution.

The precise mapping of risks related to vital statistics is an important tool to guide health policies that may lead to a reduction of events such as infant mortality. Estimates for the event reporting probabilities, which provide a measure of severity of underreporting, help to decide about where additional resources for surveillance programs would be most necessary and effective. The model introduced in this work is another attractive tool to account for underreporting bias in this context.

It is an interesting topic for future research to introduce partitioning models, such as Dirichlet process or product partition models, for underreported data. Such kind of models will allow us to also infer about the clusters throughout the estimation process. Extensions of the proposed model should also consider the situation in which there are spatial patterns in the reporting process. By borrowing strength from spatial modeling and extreme learning machines, Prates [2019] introduce a hierarchical model to perform imputation over missing count data whose usage and adaptation for the context of underreporting is an interesting point for further investigation as well. Although not approached in this paper, the modeling of underreported count time series has been suggested in recent years, for instance, by Bracher and Held [2020] and Fernández-Fontelo *et al.* [2016]. Another related problem that may interest readers is the estimation of animal abundance with differential probability of detection (see, e.g., Dorazio and Royle [2005]; Hickey and Sollmann [2018]). In this context, hierarchical Poisson models are also used to model both the underlying process and the detection (reporting) probability.

References

Alexander, M., and Alkema, L. (2018). Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research*, **38**(15), 335–372.

Alkema, L., and New, J.R. (2014). Global estimation of child mortality using a Bayesian B-spline bias-reduction model. *The Annals of Applied Statistics*, **8**(4), 2122–2149.

Bailey, T. C., Carvalho, M. S., Lapa, T. M., Souza, W. V., and Brewer, M. J. (2005). Modeling of under-detection of cases in disease surveillance. *Annals of Epidemiology*, **15**(5), 335–343.

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.

Bracher, J. and Held, L. (2020). A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts. *arXiv:2003.05885 [stat.ME].*

Campos, D., Loschi, R. H., and França, E. (2007). Early neonatal hospital mortality in Minas Gerais: Association with healthcare variables and the issue of underreporting (in Portuguese). *Revista Brasileira de Epidemiologia*, **10**(2), 223–238.

Caudill, B. S. and Mixon Jr., F. G. (1995). Modeling Household Fertility Decisions: Estimation and Testing of Censored Regression Models for Count Data. *Empirical Economics*, **20**(2), 183–196.

Dorazio, R. M. and Royle, J. A. (2005). Estimating Size and Composition of Biological Communities by Modeling the Occurrence of Species. *Journal of the American Statistical Association*, **100**(470), 389–398.

Dvorzak, M. and Wagner, H. (2016). Sparse Bayesian modelling of underreported count data. *Statistical Modelling*, **16**(1), 24–46.

Fernández-Fontelo, A., Cabaña, A., Puig, P., and Moriña, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*, **35**(26), 4875–4890.

França, E., Abreu, D., Campos, D. and Rausch, M. C. (2006). Avaliação da Qualidade da informação sobre a mortalidade infantil em Minas Gerais: Utilização de uma metodologia simplificada (in Portuguese). *Revista Médica de Minas Gerais*, **16**(1 supl 2), S28–S35.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**(423), 881–889.

Gustafson, P., Gelfand, A. E., Sahu, S. K., Johnson, W. O., Hanson, T. E., Joseph, L., and Lee, J. (2005). On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables. *Statistical Science*, **20**(2), 111–140.

Hickey , J. R. and Sollmann, R. (2018). A new mark-recapture approach for abundance estimation of social species. *PLOS One*, **13**(12):e0208726.

Huang, G. H. (2005). Model Identifiability. *Encyclopedia of Statistics in Behavioral Science*, John Wiley & Sons, Ltd, Chichester. Volume **3**, pp. 1249–1251.

Ibrahim, J. G., Chen, M-H., and Sinha, D. (2001) *Bayesian Survival Analysis.* New York: Springer-Verlag; 2001. pp. 589.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A*, **186**, 453–461.

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika*, **56**, 331–347.

Moreno, E. and Girón J. (1998). Estimating with incomplete count data: A Bayesian approach. *Journal of Statistical Planning and Inference*, **66**(1), 147–159.

Oliveira, G. L., Loschi, R. H., and Assunção, R. M. (2017). A random-censoring Poisson model for underreported data. *Statistics in Medicine*, **36**(30), 4873–4892.

Oliveira, G. L., Argiento, R., Loschi, R. H., Assunção, R. M., Branco, M. D. and Ruggeri, F. (2020). Supplementary material for "Bias correction in clustered underreported data". *Bayesian Analysis.* DOI: 10.1214/20-BA1244SUPP.

Papadopoulos, G. and Silva, J. M. C. S. (2012). Identification issues in some double-index models for non-negative data. *Economics Letters*, **117**(1), 365–367.

Picci, G. (1977). Some connections between the theory of sufficient statistics and the identifiability problem. *SIAM Journal on Applied Mathematics*, **33**(3), 383–398.

Prates, M. O. (2019). Spatial extreme learning machines: An application on prediction of disease counts. *Statistical Methods in Medical Research*, **28**(9), 2583–2594.

R Core Team (2015). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, (2015). Available at https://www.R-project.org/.

Rothenberg, T. J. (1971). Identification on parametric models. *Econometrica*, **39**(3), 577–591.

Schmertmann, C. and Gonzaga, M. R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*, **55**(4), 1363–1388.

Silva, G.D.M. da, Bartholomay, P., Cruz, O.G. and Garcia, L.P. (2017). Evaluation of data quality, timeliness and acceptability of the tuberculosis surveillance system in Brazil's microregions. *Ciênc. Saúde Coletiva [online]*, **22**(10), 3307–3319.

Stamey, J. D., Young, D. M., and Boese, D. (2006). A Bayesian hierarchical model for Poisson rate and reporting-probability inference using double sampling. *Australian & New Zealand Journal of Statistics*, **48**(2), 201–212.

Stoner, O; Economou, T; Drummond, G. (2019). A Hierarchical Framework for Correcting Under-Reporting in Count Data. *Journal of the American Statistical Association*, **114**(528), 1481–1492.

Whittemore, A. S. and Gong, G. (1991). Poisson Regression with Misclassified Counts: Application to Cervical Cancer Mortality Rates. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **40**(1), 81–93.

World Health Organization (WHO). (2006). Neonatal and perinatal mortality: Country, regional and global estimates. *World Health Organization (WHO) Library Cataloguing-in-Publication Data.*

Acknowledgements

This research was partially funded by CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*), Brazil, and CNR (*Consiglio Nazionale delle Ricerche*) Italy, Scientific Cooperation Agreement, Grant 490233/2011-2. R. H. Loschi, Renato M. Assunção, and G. L. de Oliveira additionally thank the Brazilian funding aggencies CNPq, CAPES (*Coordenação de Aperfeiçoamento de Pessoal Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*), and FAPEMIG (*Fundação de Amparo à Pesquisa do Estado de Minas Gerais*) for partially support their research. We also thank the valuable comments by the editor and three reviewers which substantially contributed to the improvement of the paper.

Supplementary Material

S.1 Web Appendix A: Prior unconditional means and variances for the reporting probabilities $\boldsymbol{\epsilon}$ under the conditional uniform prior distribution

It is expected that in regions where the data quality is poor one will have a small value for the reporting probability. The opposite is expected for regions where the data quality is good. The conditional uniform prior specification for $\boldsymbol{\gamma}$ (given in expression (2.6) of the main paper), as well as the generalized Beta prior specification (expression (2.5) in the main paper), can capture this prior behavior if the hyperparameters a_l and a_l^* are appropriately chosen. The unconditional prior expectation and variance of γ_j may be useful when eliciting these parameters. Especially, these statistics are helpful whenever the specification of an informative prior distribution for all components of the parameter vector $\boldsymbol{\gamma}$ is needed. If the area i is classified in the jth data quality cluster, such prior summaries are, respectively, given by:

$$\mathrm{E}(\gamma_j) = \begin{cases} \frac{a_j+a_j^*}{2}\prod_{l=1}^{j-1}\left[1-\frac{(a_l+a_l^*)}{2}\right] & j \geq 2 \\ \frac{(a_1+a_1^*)}{2} & j = 1, \end{cases}$$

and

$$\mathrm{V}(\gamma_j) = \begin{cases} \left[\frac{(a_j^*-a_j)^2}{12}+\frac{(a_j^*+a_j)^2}{4}\right]\mathrm{V}\left(\sum_{l=1}^{j-1}(\gamma_l)\right)+\frac{(a_j^*-a_j)^2}{12}\left[1-\sum_{l=1}^{j-1}\mathrm{E}(\gamma_l)\right]^2 & j \geq 2 \\ (a_1^*-a_1)^2/12 & j = 1. \end{cases}$$

Consequently, the prior expectation for ϵ_i is $\mathrm{E}(\epsilon_i) = \prod_{l=1}^{j}\left[1-\frac{(a_l+a_l^*)}{2}\right]$ if $i \in A_j$, $i = 1, \ldots, A$, where A_j represents the jth data quality cluster for $j = 1, ..., K$. If the experts believe that data are fully recorded in the best cluster, then one should set $P(\gamma_1 = 0) = 1$ and the prior expectation of γ_j, for all $j \geq 2$, would be modified by removing the factor $1-\frac{a_1+a_1^*}{2}$. When moving from one specific cluster to the next and worse one, the factor $\left[1-\frac{(a_j+a_j^*)}{2}\right]$ represent the experts knowledge on how the recording probability in latter decreases if compared to the former. Such decrease depends on the sum $a_j + a_j^*$ rather than on the distinct values of a_j and a_j^*. This existent identifiability issue may be solved by also considering the variances $V(\epsilon_i)$, which describe how confident the experts are on their assessment of the expected value $E(\epsilon_i)$. These variances are given by

$$\mathrm{V}(\epsilon_i) = \begin{cases} \left[\frac{(a_j^*-a_j)^2}{12}+\frac{(2-a_j^*-a_j)^2}{4}\right]\mathrm{V}\left(\sum_{l=1}^{j-1}(\gamma_l)\right)+\frac{(a_j^*-a_j)^2}{12} & \left[1-\sum_{l=1}^{j-1}\mathrm{E}(\gamma_l)\right]^2 \\ & \text{if } i \in A_j,\ j = 2, ..., K, \\ (a_1^*-a_1)^2/12 & \text{if } i \in A_1. \end{cases}$$

S.2 Web Appendix B: Posterior full conditional distributions

This section provides the posterior full conditional (p.f.c.) distributions needed for sampling from the joint posterior distribution. Consider the model for the observed data defined by expressions (2.1) and (2.2) in the main paper with $\log(\theta_i) = \beta_0 + \mathbf{X}_i'\boldsymbol{\beta} + u_i + s_i$, Denote $\boldsymbol{\Psi} = (\beta_0, \boldsymbol{\beta}, \boldsymbol{s}, \boldsymbol{u}, \sigma_s^2, \sigma_u^2)$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)$, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_K)$ and $\boldsymbol{y} = (y_1, ..., y_A)$. To simplify the notation, consider $\boldsymbol{\Psi}_{-\kappa}$ as the vector $\boldsymbol{\Psi}$ without the coordinate κ and let $\phi_p(. \mid \mathbf{M}, \mathbf{V})$ be the probability density function (p.d.f.) of the p-variate Gaussian distribution with mean vector $\mathbf{M}$ and covariance matrix $\mathbf{V}$. In the univariate case, p is omitted.

Assuming the prior distributions discussed in Section 2.2.2 of the paper, the joint posterior distribution for $(\boldsymbol{\Psi}, \boldsymbol{\gamma}, \boldsymbol{\omega})$ is not known in closed form. Posterior inference for all these parameters can be carried out through a Markov chain Monte Carlo (MCMC) scheme. The p.f.c. distributions for parameters $\boldsymbol{u}$, $\boldsymbol{s}$, σ_u^2, σ_s^2 and β_0 are, respectively, given by

$$\begin{aligned}
\pi(\boldsymbol{u} \mid \boldsymbol{\Psi}_{-\boldsymbol{u}}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{y}) &\propto \phi_A(\boldsymbol{u} \mid \sigma_u^2 \boldsymbol{y}', \sigma_u^2 \mathbf{I}_A) \exp\left\{ -\sum_{i=1}^{A} E_i(1 - h_i\boldsymbol{\gamma}) e^{u_i} \right\}, \\
\pi(s_i \mid \boldsymbol{\Psi}_{-s_i}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{y}) &\propto \phi\left(s_i \mid \frac{\sum_j \delta_{ij} s_j - y_i \sigma_s^2}{\sum_j \delta_{ij}}, \frac{\sigma_s^2}{\sum_j \delta_{ij}} \right) \exp\left\{ -E_i(1 - h_i\gamma) e^{s_i} \right\}, \ \forall\ i, \\
\pi(\sigma_u^2 \mid \boldsymbol{\Psi}_{-\sigma_u^2}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{y}) &\sim \text{IG}\left(\sum_{i=1}^{A} u_i^2 + a_u,\ A + d_u - 3 \right), \\
\pi(\sigma_s^2 \mid \boldsymbol{\Psi}_{-\sigma_s^2}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{y}) &\sim \text{IG}\left(\sum_{i \sim j} \delta_{ij}(s_i - s_j)^2 + a_s,\ A + d_s - 3 \right), \\
\pi(\beta_0 \mid \boldsymbol{\Psi}_{-\beta_0}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{y}) &\propto \phi\left(\beta_0 \mid \sum_{i=1}^{A} y_i \sigma_{\beta_0}^2, \sigma_{\beta_0}^2 \right) \exp\left\{ -\sum_{i=1}^{A} E_i(1 - h_i\boldsymbol{\gamma}) e^{\beta_0} \right\},
\end{aligned}$$

where δ_{ij} is related to the neighborhood structure inherent to the region such that $\delta_{ii} = 0$ and, for all $i, j = 1, \ldots, A$, $i \neq j$, $\delta_{ij} = 1$ if i and j are first order neighboring areas and, $\delta_{ij} = 0$ otherwise.

Assuming the SSVS prior in expression (2.8) of the paper, the p.f.c. distribution for each fixed effect $\beta_m,\ m = 1, \ldots, p$, is given by

$$\begin{aligned}
\pi(\beta_m \mid \boldsymbol{\Psi}_{-\beta_m}, \boldsymbol{\gamma}, \omega_m, \boldsymbol{y}) \propto\ & \phi\left(\beta_m \mid \sum_{i=1}^{A} y_i X_{mi}[\sigma_{spike}^2 \omega_m + \sigma_{slab}^2(1 - \omega_m)], \sigma_{spike}^2 \omega_m + \sigma_{slab}^2(1 - \omega_m) \right) \\
\times\ & \exp\left\{ -\sum_{i=1}^{A} E_i(1 - h_i\boldsymbol{\gamma}) e^{\beta_m X_{mi}} \right\}.
\end{aligned}$$

The p.f.c. distribution for the weights $\omega_m,\ m = 1, ..., p$, is a Bernoulli distribution with parameter $\rho_m^* = \rho_m \phi\left(\beta_m \mid 0, \sigma^2_{spike}\right)\left[\rho_m \phi\left(\beta_m \mid 0, \sigma^2_{spike}\right) + (1-\rho_m)\phi\left(\beta_m \mid 0, \sigma^2_{slab}\right)\right]^{-1}$.

If the generalized Beta prior distribution (given in expression (2.5) of the main paper) is assumed for $\boldsymbol{\gamma}$, then the p.f.c. distribution of γ_j, for $j = 2, \ldots, K$, depends on $\gamma_1, \ldots, \gamma_{j-1}$ and it is given by

$$\begin{aligned}\pi(\gamma_j \mid \boldsymbol{\Psi}, \boldsymbol{\gamma}_{-j}, \omega, \boldsymbol{y}) \quad &\propto \quad \exp\left\{\gamma_j \sum_{l=j}^{K} \sum_{i \in A_l} E_i \theta_i\right\} \prod_{l=j}^{K} [L(l+1)]^{\sum\limits_{i \in A_l} y_i} \\ &\times \quad [\gamma_j - a_j L(j)]^{\alpha_j - 1}\left[a_j^* L(j) - \gamma_j\right]^{\nu_j - 1} \text{ if } \gamma_j \in \left(a_j L(j), a_j^* L(j)\right). \qquad (2.9)\end{aligned}$$

where $L(j) = 1 - \sum_{k=1}^{j-1} \gamma_k$. The p.f.c. distribution of γ_1 will depend on $L(2), \ldots, L(K+1)$ but not on $L(1)$. Thus, $\pi(\gamma_1 \mid \boldsymbol{\Psi}, \boldsymbol{\gamma}_{-1}, \omega, \boldsymbol{y})$ is obtained by deleting the term $L(1)$ from $\pi(\gamma_j \mid \boldsymbol{\Psi}, \boldsymbol{\gamma}_{-j}, \omega, \boldsymbol{y})$ wherever it appears. If, instead, the conditional uniform prior distribution (expression (2.6) of the main paper) is assumed for for $\boldsymbol{\gamma}$, the p.f.c. distributions of $\gamma_1, \ldots, \gamma_K$ are obtained from expression (2.9) by simply setting $\alpha_j = \nu_j = 1$, for all $j = 1, \ldots, K$.

S.3 Web Appendix C: Further simulated data studies

In this section, we provide additional simulation studies exploring the potential of proposed model in further scenarios than those presented in the main manuscript.

S.3.1 Simulation Study IV: Effect of wrongly defining the number K of data quality categories

Another important feature to be analyzed in the proposed model is the effect of misspecifying the number of data quality categories on the posterior inference. We consider again the $R = 100$ datasets analyzed in Section 2.3.1 of the main paper, which were generated assuming $K = 4$ categories. Besides the analysis with the correct number of categories ($K = 4$) presented in Section 2.3.1 of the paper, we analyze these datasets considering $K = 2$ and $K = 6$. To fit the models whenever $K = 2$, we merge the two best and the two worst data quality clusters creating two new groups composed by a total 44 and 31 areas, respectively. When assuming $K = 6$, we divide both the best and the worst original clusters into two new groups with balanced number of areas. To simplify the analysis, we only assume the partially informative conditional uniform prior for $\boldsymbol{\gamma}$ in which $\gamma_1 \sim U(0, 0.10)$. Note that the prior mean induced for γ_1 equals the true value of such parameter in the best data quality cluster when assuming $K = 4$ and $K = 6$. However, it underestimates γ_1, at least for part of the areas, if only $K = 2$ clusters are assumed to fit the generated data.

Table 2.8 shows that the misspecification of K introduces more bias as well as higher variability in the posterior estimates for $\boldsymbol{\theta}$. Bias and RMSE are much higher if the number of clusters assumed in the proposed model is smaller than the true value of K. This is probably occurring because, in this simulation study, the number of misclassified areas is higher if $K = 2$. Despite of this, we still have better posterior estimates than the ones obtained by fitting the standard Poisson model (see Table 2.1 of the main paper), indicating that the proposed methodology is an attractive approach to model underreported count data even when the number K of data quality categories is not precisely known.

Table 2.8: Bias and relative mean squared error (RMSE) for the estimated relative risks $\boldsymbol{\theta}$ under proposed model with $K = 2,\ 4$ and 6 data quality categories; Simulation Study IV.

	RMSE	Bias	RMSE	Bias	RMSE	Bias
	$K = 2$		$K = 4$		$K = 6$	
Model 1	0.016	-0.285	0.001	-0.000	0.002	0.055
Model 2	0.015	-0.276	0.001	-0.002	0.002	0.051
Model 3	0.021	-0.295	0.002	-0.003	0.003	0.060

S.3.2 Simulation Study V: Effect of the number of areas within the best and worst data quality categories

To fit the proposed model, we must first classify the A areas into K data quality clusters/categories. This section aims at evaluating whether the number of areas within the best and worst data quality categories significantly affects the posterior inference for relative risks $\boldsymbol{\theta}$. For doing that, $R = 100$ datasets are generated for a region containing $A = 75$ areas as described in the introduction of Section 2.3 of the main paper. We assume a total of $K = 4$ clusters but varying the number of areas within the best and the worst ones. For all scenarios, we consider a total of 16 and 14 areas within the second and third clusters, respectively. In relation to the best and worst groups, we assume the following cases: 28 areas (best) and 17 areas (worst) in Case 1; 17 areas (best) and 28 areas(worst) in Case 2; 40 areas (best) and 5 areas (worst) in Case 3; and 5 areas (best) and 40 areas(worst) in Case 4. The prior distributions are the same considered in Section 2.3.1 of the paper.

Table 2.9 shows that, under both partially informative and fully informative prior distributions for $\boldsymbol{\gamma}$, the models produced very similar results in terms of bias and RMSE. We note that having a greater number of areas within the worst data quality cluster (Case 4) makes the bias in the posterior estimates of $\boldsymbol{\theta}$ increases without substantially affecting the RMSE. This is an expected behavior since, whenever the number of areas within the best group is bigger, the model induce an informative prior for a greater number of areas. Under the standard Poisson model (results not shown) the rates $\boldsymbol{\theta}$ are always underestimated. The bias and the RMSE get higher as the number of areas in the worse data category increase, becoming even higher if the difference between the number of areas in the best and worse categories increases. That is not an unexpected result since as greater the number of areas within the best data quality cluster the greater is the underreporting severity in the simulated data. In general, such a behavior is also observed for estimates obtained under the partially informative and fully informative proposed models but biases and RMSEs under such models tend to be closer than in the standard Poisson model.

Table 2.9: Bias, relative mean squared error (RMSE) and nominal coverage of 95% credible intervals (Cov.) for the estimated relative risks $\boldsymbol{\theta}$ under proposed model; Simulation Study V.

	RMSE	Bias	Cov.	RMSE	Bias	Cov.
			Case 1			
	partially informative			fully informative		
Model 1	0.001	-0.000	0.988	0.001	-0.000	0.989
Model 2	0.001	-0.001	0.993	0.001	-0.002	0.992
Model 3	0.002	-0.003	0.997	0.002	-0.003	0.996
			Case 2			
	partially informative			fully informative		
Model 1	0.002	0.003	0.988	0.002	0.002	0.989
Model 2	0.001	0.003	0.992	0.001	0.003	0.992
Model 3	0.002	0.005	0.996	0.002	0.004	0.997
			Case 3			
	partially informative			fully informative		
Model 1	0.001	-0.002	0.987	0.001	-0.002	0.989
Model 2	0.001	-0.002	0.989	0.001	-0.003	0.987
Model 3	0.001	-0.002	0.995	0.002	-0.003	0.996
			Case 4			
	partially informative			fully informative		
Model 1	0.002	-0.036	0.970	0.002	-0.023	0.980
Model 2	0.002	-0.035	0.991	0.002	-0.020	0.992
Model 3	0.002	-0.029	0.998	0.002	-0.020	0.998

S.3.3 Simulation Study VI: Data perfectly recorded

To evaluate the performance of the proposed model when data is free of underreporting, we generate $R = 100$ datasets from the Poisson distribution $Y_i|\theta_i \overset{ind}{\sim} \mathcal{P}(E_i\theta_i),\ i = 1, \ldots, 75$, where the expected number of cases E_i is known and equal to that one available for case study presented in Section 4 of the main paper. We assume five independent variables as potential regressors $\boldsymbol{X}$ and the rates $\boldsymbol{\theta}$ are such that $\log(\theta_i) = \beta_0 + \boldsymbol{\beta}\boldsymbol{X}_i,\ i = 1, \ldots, A$, where $\beta_0 = 0.50$ and $\boldsymbol{\beta} = (-0.25, -0.25, 0, 0, 0.25)$.

To fit the proposed model, we assume $K = 4$ data quality categories composed by a total of 28, 16, 14 and 17 areas from the best to the worst category, respectively, as done in the study presented in Section 2.3.1 of the paper. We address this situation assuming different degrees of information about γ_1. In Case 1, we consider the same prior distributions given in Section 2.3.1 of the paper. In Case 2, when eliciting the prior distribution for $\boldsymbol{\gamma}$ in the partially informative case, we assume the conditional uniform distribution in which $\gamma_1 \sim U(0, 0.01)$ thus, *a priori*, $E(\gamma_1) = (0.005)$. For the fully informative case, we additionally assume that $\gamma_2|\gamma_1 \sim$

$U(0, 0.1005(1-\gamma_1))$, $\gamma_3|\gamma_1, \gamma_2 \sim U(0, 0.2116(1-\gamma_1-\gamma_2))$ and $\gamma_4|\gamma_1, \gamma_2, \gamma_3 \sim U(0, 0.2367(1-\gamma_1-\gamma_2-\gamma_3))$. Consequently, the prior expectation of γ_2, γ_3 and γ_4 are, respectively, given by 0.05, 0.10 and 0.10. Under this prior specification, the prior means for the reporting probabilities $\boldsymbol{\epsilon}$ approximate better the situation generated in this simulation study if compared to the prior specification considered in Case 1.

As expected, the standard Poisson model performs very well regardless of the structure used to model the rates $\boldsymbol{\theta}$ (see Table 2.10). Estimates for the rates $\boldsymbol{\theta}$ under the proposed model assuming both the partially informative and the fully informative prior distributions for $\boldsymbol{\gamma}$ present a small RMSE but a high bias. This is an expected result since, by construction, the proposed model imposes some correction on the relative risks. Models 1 and 2 present smaller bias than Model 3 but the nominal coverage for the credible interval is worse for such models. If the prior specification for $\boldsymbol{\gamma}$ is more informative about the true value considered to generate the datasets (Case 2), then an improvement on estimates of $\boldsymbol{\theta}$ is observed as the biases and RMSE are substantially reduced.

Table 2.10: Bias, relative mean squared error (RMSE) and nominal coverage of 95% credible intervals (Cov.) for the estimated relative risks $\boldsymbol{\theta}$ under proposed model; Simulation Study VI.

	RMSE	Bias	Cov.	RMSE	Bias	Cov.
	proposed model - Case 1					
	partially informative			fully informative		
Model 1	0.008	0.192	0.501	0.008	0.192	0.501
Model 2	0.007	0.189	0.477	0.007	0.190	0.478
Model 3	0.009	0.209	0.823	0.009	0.209	0.824
	proposed model - Case 2					
	partially informative			fully informative		
Model 1	0.002	0.094	0.750	0.002	0.094	0.752
Model 2	0.002	0.092	0.742	0.002	0.091	0.743
Model 3	0.003	0.110	0.945	0.003	0.109	0.943
	standard Poisson model					
Model 1	0.000	0.000	0.958	-	-	-
Model 2	0.000	0.000	0.967	-	-	-
Model 3	0.000	0.000	0.996	-	-	-

Appendix

A.1 Application to Brazilian tuberculosis data

Tuberculosis (TB) is one of the world's major public health problems. According to the World Health Organization (WHO), TB is the ninth leading cause of death worldwide and the leading cause from a single infectious agent, ranking above HIV/AIDS. In 2016, an estimated 1.7 million people died from TB, including nearly 400,000 people who were co-infected with HIV. Brazil is among the top twenty countries by absolute mortality [World Health Organization, 2017]. Ending the TB epidemic by 2030 is among the health targets of the Sustainable Development Goals of the United Nations. To assess whether these targets are reached and to provide better estimates for the incidence rates, robust monitoring and evaluation of trends in the burden of TB are essential.

Estimation of TB incidence is a major challenge in many countries due to underreporting and under-diagnosis of TB cases. Tackling the epidemic requires action to close gaps in care and availability of financial resources. The WHO Global Tuberculosis Report 2017 also evidences that underreporting and underdiagnosis of TB cases continues to be a challenge, especially in countries with large unregulated private sectors and weak health systems [World Health Organization, 2017]. The WHO has performed some inventory studies to measure the level of TB underreporting in civil registration systems, especially in endemic countries [World Health Organization, 2012].

In Brazil, the Notifiable Diseases Information System (SINAN) provides information about the tuberculosis occurrence, patterns and trends. The notification of TB cases in SINAN is mandatory and, despite its high spatial coverage, the system is not able to report all TB cases [Stoner *et al.*, 2019]. Santos *et al.* [2018] showed that the variables associated with underreporting of TB were mostly related to the healthcare system rather than to individual characteristics of the patients, which indicates the need for training the health professionals in order to correctly notify the information in the systems. As pointed out by the Brazilian Ministry of Health [Ministério da Saúde do Brasil, 2016], underreporting of TB represents a major loss as it leads to a delay in starting the TB treatment.

In this paper, we apply the model proposed in Oliveira *et al.* [2020] to estimate the TB incidence rates in the $A = 557$ mainland Brazilian microregions considering SINAN's data from 2012 to 2014. We consider usual clustering techniques to define the required data quality groups. That dataset were previously analyzed by Stoner *et al.* [2019], from which we obtained all variables considered in our analysis. Results are compared to those obtained by fitting Stoner *et al.* [2019]'s model.

The content of this appendix have contributions of Guilherme Oliveira, Rosangela Loschi and Renato Assunção. At the time of dissertation submission, the analysis were being improved for posterior submission to a peer-reviewed journal.

Based on our final dataset (after usual cleaning for missing data and inconsistent information), we found that between 2012 and 2014 there were 208,901 TB cases notified in the 557 Brazilian microregions. As we aim to map the TB incidence in Brazilian territory, we exclude the microregion of *Fernando de Noronha* from our analysis since it is an island with no contiguous neighboring area.

A.1.1 Model specification

We assume $Y_i \mid \theta_i, \epsilon_i \overset{ind}{\sim} \mathcal{P}oisson(n_i\theta_i\epsilon_i),\ i = 1, ..., 557$, where n_i is an offset representing the total population in the ith area. The TB relative risk assumes a log-linear regression structure which includes local and spatial random effects, that is, $\log(\theta_i) = \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + u_i + s_i,\ \forall\ i$, where u_i and s_i represent the usual local and spatial effects, respectively. Five covariates are introduced in this regression model: the proportion of economically active adults without employment (Unemployment), the the proportion of people residing in households with more than two persons per room (Density), the proportion of people living in an urban setting (Urbanisation), the proportion of the population made up by indigenous groups (Indigenous) and average monthly coverage (%) of the Brazil's Family Health Strategy (ESF) from 2012 to 2014 in relation to the total population (ESF).

The approach for the compound Poisson model proposed in Oliveira *et al.* [2020] requires the prior specification of data quality groups for the microregions. We will refer to such model as the *Clustering Model*. To define the clustering indicator variable, we performed an usual clustering method with basis on a set of numerical indicators proposed in Silva *et al.* [2017] to evaluate the quality of data recorded in the Brazilian TB surveillance system. These authors considered 14 indicators to measure four attributes for the TB data recorded in SINAN from 2012 to 2014: completeness, consistency, timeliness and acceptability. More specifically, we collected from Silva *et al.* [2017] the indicators of *consistency* (percentage of cases with notification date greater or equal to diagnosis date), *completeness* (median for the percentage of completeness measured in five attributes of the SINAN registration form), *timeliness of notification* (percentage of cases with an interval between notification date and diagnosis date smaller or equal to 7 days) and *timeliness of treatment* (percentage of cases with an interval between the date of starting treatment and diagnosis of less than 1 day). Besides these four indicators, we included in the clustering analysis the information of two other covariates originated from distinct data sources: the *percentage of general deaths with ill-defined cause* (collected from the Brazilian DATASUS repository for period 2012-2014, available at http://www2.datasus.gov.br/) and the *estimated registration coverage for the Brazilian mortality information system* (available from Schmertmann and Gonzaga [2018]'s companion website http://mortality-subregistration.schmert.net/). Although these two last variables may be more likely related to underreporting of TB deaths rather than TB incidence, we consider they are relevant proxies for general quality of the civil systems for collecting health data in Brazil. As such, they can be helpful in our data quality clustering definition.

The six previous variables was applied to the usual Ward linkage clustering method with the squared Euclidean distance measure. By comparing the similarity measures in the clustering algorithm steps, we found that using $K = 23$ groups is an interesting strategy to analyze our TB data in period 2012-2014. The groups were labeled into hierarchical data quality categories according to the resulting clusters' centroid (mean). As all variables considered for grouping are measured in an increasing quality scale, we assumed that the greater the cluster mean (centroid), the best the data quality. The best group (Cluster 1) ended with 32 microregions whereas only 3 microregions were allocated to the worst data quality cluster (Cluster 23). Then, following Oliveira *et al.* [2020] we model the TB reporting probabilities in each area i as being

$$\epsilon_i = 1 - \sum_{j=1}^{23} h_{ji}\gamma_j, \tag{2.10}$$

where $h_{ji} = 1$ if area i belongs to cluster j and $h_{ji} = 0$ otherwise, for $i = 1, \ldots, A$ and $j = 1, \ldots, 23$. Parameters $\gamma_1, \ldots, \gamma_{23}$ are related to the clustering underreporting probabilities which are discussed in details in Oliveira *et al.* [2020].

For comparison purposes, we also fitted the Brazilian TB data using the modeling strategy proposed in Stoner *et al.* [2019]. The relative risks $\boldsymbol{\theta}$ are modeled using the same log-regression structure previously mentioned. Stoner *et al.* [2019] assumes that the reporting probabilities $\boldsymbol{\epsilon}$ have the logistic-regression structure given by

$$\text{logit}\,(\epsilon_i) = \alpha_0 + \boldsymbol{g}(w_i)\boldsymbol{\alpha} + \delta_i, \quad \text{for } i = 1, \ldots, A, \tag{2.11}$$

where w represents the covariate *timeliness of treatment* previously mentioned; $\boldsymbol{g}$ is a function defining an orthogonal polynomial of degree 3 introduced to reduce multiple-collinearity and, at the same time, it ensures that $\boldsymbol{g}(w) = 0$ when $w = \bar{w}$, so that (at the logistic scale) α_0 is the mean reporting rate for a region with mean treatment timeliness (for more details, see Section 3 of Stoner *et al.* [2019]); $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$; and δ_i is a local random effect. We will refer to such model as the *Pogit Model.*

A.1.2 About the prior elicitation

To fit the Clustering Model we adopt the conditional uniform prior distribution given in expression (2.6) to model the parameter vector $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{23})$, eliciting an informative prior distribution only for parameter γ_1 (called partially informative prior distribution). To build such informative prior distribution, we consider available studies on TB underreporting in Brazil. As discussed in Stoner *et al.* [2019], in 2017 the WHO reported point estimates for overall TB detection rate in Brazil for years 2012, 2013 and 2014 [World Health Organization, 2017]. The results are related to inventory study-derived estimates [World Health Organization, 2012] and revels that, respectively for those years, 91% (78%, 100%), 84% (73%,99%), and 87% (75%,100%) of TB cases was detected in the Brazilian microregions, where the quantities between parenthesis are the associated 95% confidence intervals. The reporting probability in those areas experiencing the best data quality (areas within Cluster 1) is likely greater than the overall detection level. With basis on the findings of previous studies regarding TB underreporting in Brazil [Sousa and Pinheiro, 2011; Sousa *et al.*, 2012; Oliveira *et al.*, 2012; Silva *et al.*, 2017; Stoner *et al.*, 2019], we assume that $\gamma_1 \sim U(0.0, 0.05)$ appropriately reflects our prior belief about ϵ_i, for all $i \in$ Cluster 1.

When considering the Pogit Model, the prior specification for the reporting probabilities $\boldsymbol{\epsilon}$ given in equation (2.11) are the same considered in Stoner *et al.* [2019]. Namely, it is assumed a Gaussian $\text{N}(0, 100)$ for the fixed effects $\boldsymbol{\alpha}$ and a Gaussian $\mathbf{N}(0, \sigma^2_\delta)$ for each addictive local effect $\delta_i,\ i = 1, ..., A$. The required informative distributive for parameter α_0, which represents the overall mean reporting rate, was assumed to be a Gaussian distribution $\text{N}(2, 0.36)$ with basis on the information provided by the WHO. It is worth noticing that all Gaussian distributions described in this section are parametrized in terms of mean and variance.

Regarding the prior specification for the log-linear structure of the relative risks $\boldsymbol{\theta}$, in both models we assume mean-centered covariates with Gaussian prior $\text{N}(0, 100)$ for their fixed effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)$. Following Stoner *et al.* [2019], we assume *a priori* that $\beta_0 \sim \text{N}(-8, 1)$. This was specified by those authours using a prior predictive checking and it reflects the belief that very high values (such as over 1 million) for the total number of TB cases are unlikely. Additionally, we assume that $u_i \overset{iid}{\sim} \text{N}(0, \sigma^2_u)$ and that $\boldsymbol{s} = (s_1, \ldots, s_A)$ have the ICAR prior distribution [Besag *et al.*, 1991] with precision parameter $\tau_s = \sigma_s^{-2}$. Following Stoner *et al.* [2019], the prior distributions for variances σ^2_u, σ^2_s and σ^2_δ are truncated $\text{N}(0, 1)$ with domain restrict to $(0, \infty)$. Such a choice reflects the belief that low variance values are more likely than higher ones.

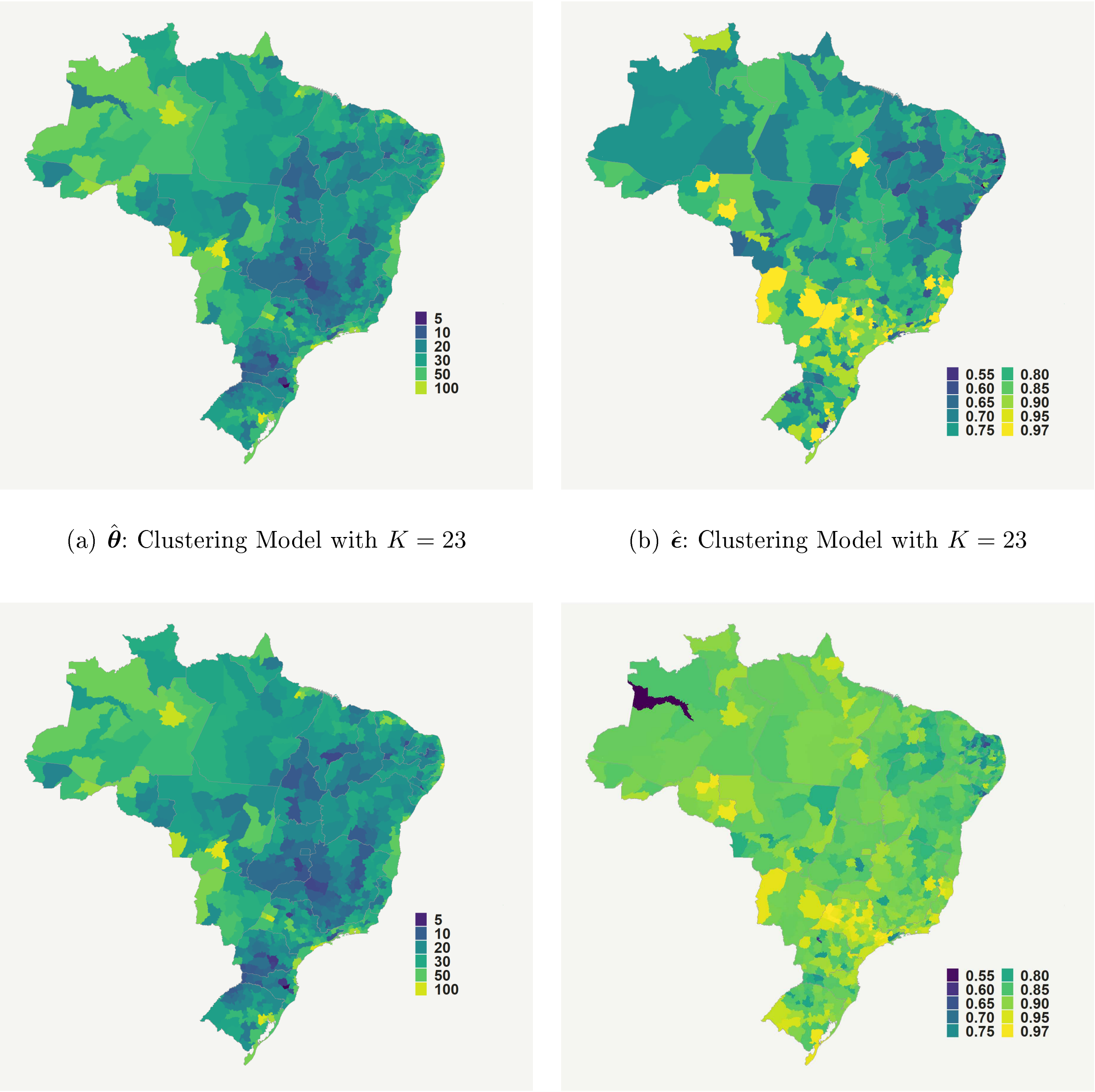
The MCMC scheme is performed using package Nimble [de Valpire *et al.*, 2017] from software R [R Core Team , 2015]. The basic script for running each model is available in the supplementary materials of Stoner *et al.* [2019] and Oliveira *et al.* [2020]. For both models two chains were considered each with a total of 3,000,000 iterations, being the first 1,000,000 discarded as a burn-in period and a lag of 3,000 iterations was selected in order to avoid autocorrelated posterior samples. Trace plots for the MCMC samples were inspected and the potential scale reduction factor (PSRF) [Brooks and Gelman, 1998] was calculated as less than 1.04 and 1.06 for all regression coefficients and variance parameters, respectively, in the Clustering and the Pogit Models, thus indicating convergence.

A.1.3 Posterior Results

Figure 2.4 displays the spatial distribution of the estimated TB incidence rates per 100,000 inhabitants and the respective reporting probabilities throughout the 557 mainland Brazilian microregions under the Censoring Model and the Pogit Model. The models provided a quite similar spatial structure for the tuberculosis incidence (Panels (a) and (c)), with highest values mainly concentrated in the North and Central-West regions of the country. Clusters of microregions with elevated values for the disease incidence can also be observed along the coast of

Brazil (right edge of the map).

Regarding to the reporting probabilities, from Panels (b) and (d) of Figure 2.4, it can be seen that, except for an specific microregion in the Northwest of Brazil, both models provide a similar spatial pattern. Although the reporting probabilities showed to be more homogeneous under the Pogit model, there is an agreement in relation to the clusters of areas with the highest and smallest values for the posterior estimates of $\boldsymbol{\epsilon}$, specially regarding the highest values. In general, estimates under the Pogit Model are more concentrated in greater values than those observed under the Clustering Model.



(a) $\hat{\boldsymbol{\theta}}$: Clustering Model with $K = 23$

(b) $\hat{\boldsymbol{\epsilon}}$: Clustering Model with $K = 23$

(c) $\hat{\boldsymbol{\theta}}$: Pogit Model

(d) $\hat{\boldsymbol{\epsilon}}$: Pogit Model

Figure 2.4: Posterior mean for the tuberculosis incidence rates per 100,000 inhabitants (left) and the reporting probability (right) under the model proposed in Oliveira *et al.* [2020] with $K = 23$ data quality clusters (top), denoted by Clustering Model, and also under the modeling strategy proposed in Stoner *et al.* [2019] (bottom), denoted by Pogit Model; Brazilian data, 2012-2014.

The minimum estimate for the reporting probabilities under the Pogit Model is 0.5352, the only estimate smaller than 0.60. Under such a model, the first quartile, mean, third quartile and maximum estimates were, respectively, 0.8515, 0.8722, 0.8978 and 0.9748. The four smallest estimates (all above 0.675) were observed in microregions in which the value for covariate w is zero. Furthermore, these four microregions figured out among the 10% smallest populations and 17% lowest TB counts. This analysis suggests that the estimation of the reporting probabilities is highly influenced by the reporting proxy variable considered in the logistic regression. In small populations, the variable *timeliness of treatment* (w) may not be measured properly, inducing to discrepant results.

A similar analysis under the Clustering model revels that the minimum estimate for the reporting probabilities is 0.4756 (within Cluster 23). This is the only cluster with an estimate for the reporting probability smaller than 0.60 and it composed by 3 microregions. Under such a model, the first quartile, mean, third quartile and maximum estimates were, respectively, 0.7128, 0.7865, 0.8416 and 0.9759. The three regions within the worst cluster figured out among the 13% smaller populations and 12% lower TB counts. The quite discrepant small value for the reporting probability in this cluster may be related to the fact that it only contains microregion with small populations.

Table 2.11 summarizes the results for relative risks $\boldsymbol{\theta}$. In both models, the 95% highest posterior density interval ($HPD_{95\%}$) for covariate *ESF* contains the value zero, thus indicating that this proxy for access to healthcare does not have a significant (non-null) effect in the tuberculosis incidence rate. Among the other covariates, only the effect of *Density* was estimated differently under the two fitted models. Result provided by the Clustering Model is more consistent with what is expected in practice for the effect of such covariate.

The log pseudo-marginal likelihood (LPML) criterion [Ibrahim, Chen, and Sinha, 2001] points that the TB data is better fitted by the Clustering Model.Such model is more parsimonious than the Pogit Model since only $K = 23$ parameters are estimated in the reporting mechanism instead of estimating the fixed effects α_0 and $\boldsymbol{\alpha}$ besides the local effects $\delta_1, \ldots, \delta_{557}$. In some sense, there are more data information available to estimate each unknown parameter associated to $\boldsymbol{\epsilon}$ under the Clustering Model than under the Pogit Model. This might be one of the reasons for the slightly better performance of the former model in relation to the latter.

Table 2.11: Posterior summaries for the regression effects β_0 and $\boldsymbol{\beta}$ under the model proposed in Oliveira *et al.* [2020] with $K = 23$ data quality clusters, denoted by Clustering Model, and also under the modeling strategy proposed in Stoner *et al.* [2019], denoted by Pogit Model; Brazilian tuberculosis data for period 2012-2014. We provide the posterior mean (Mean), the posterior standard deviation (St.Dev.) and the 95% highest posterior density interval ($HPD_{95\%}$).

Covariate	Mean	St.Dev.	$HPD_{95\%}$	.Mean	St.Dev.	$HPD_{95\%}$
	Clustering Model (**LPML=-2505.357**)			Pogit Model (LPML=-2528.662)		
Intercept	-8.254	0.052	(-8.353,-8.153)	-8.360	0.065	(-8.468,-8.245)
Unemployment	0.119	0.027	(0.064,0.169)	0.047	0.011	(0.026,0.068)
Density	0.125	0.044	(0.043,0.212)	-0.218	0.009	(0.003,0.016)
Urbanisation	0.206	0.030	(0.152,0.268)	0.014	0.002	(0.010,0.017)
Indigenous	0.056	0.018	(0.019,0.090)	0.015	0.005	(0.005,0.025)
ESF	-0.026	0.025	(-0.078,0.020)	-0.001	0.001	(-0.003,0.001)

A.2 Discussion

We addressed an important problem in Epidemiology and public health fields. Providing realistic estimates for the TB incidence rates is important to guide healthcare professionals in making their decisions to control the endemic disease.

The correction of underreporting bias in Brazilian TB counts, 2012-2014, was performed using the approaches introduced by Oliveira *et al.* [2020] and Stoner *et al.* [2019]. They provided a quite similar spatial pattern for the disease incidence rates. For the reporting probabilities, estimates under the Clustering Model [Oliveira *et al.*, 2020] showed a greater discrepancy throughout the country if compared to the ones obtained under the Pogit Model [Stoner *et al.*, 2019].

For the Clustering Model, information from six data quality indicators was taken into consideration to define the groups, including that one used as TB reporting proxy in the logistic regression assumed for the Pogit Model. The effort to define the grouping was compensated by a better data fitting, according to the LPML measure. It worth noting, however, that the findings of this applied analysis cannot be generalized to other examples without further exhaustive investigation.

It is intended to make a more robust comparison between the methods through simulated scenarios. In the TB data analysis, it is also of interest to perform a sensitivity analysis regarding effects of different clustering definitions under the Oliveira *et al.* [2020]'s approach. Likewise, we aim to fit the model of Stoner *et al.* [2019] using different proxies in the logistic regression. Some discussions on these regards are presented in the cited papers but we intend to do additional studies focusing on this specific application and possibly others.

Acknowledgements

G. L. de Oliveira , R. H. Loschi and Renato M. Assunção thank the Brazilian funding aggencies CNPq, CAPES (*Coordenação de Aperfeiçoamento de Pessoal Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) , and FAPEMIG (*Fundação de Amparo à Pesquisa do Estado de Minas Gerais*) for partially support their research. The authors also thank Gabriela Drummond Marques da Silva for supporting with the datasets and discussion about the prior specifications in the fitted models.

References

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.

Brasil. Ministério da Saúde. Programa Nacional de Controle da Tuberculose [Internet]. Brasília: Ministério da Saúde; 2016 [citado 23 abr. 2017]. Available at http://portalarquivos.saude.gov.br/images/pdf/2017/fevereiro/21/Apresentacao-sobre-os-principaisindicadores-da-tuberculose.pdf.

Brooks, S.P. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

de Valpine, P., Turek, D., Paciorek, C.J., Anderson-Bergman, C., Lang, D.T. and Bodik, R. (2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With Nimble. *Journal of Computational and Graphical Statistics*, **26**, 403–413.

Ibrahim, J. G., Chen, M-H., and Sinha, D. (2001) *Bayesian Survival Analysis*. New York: Springer-Verlag; 2001. pp. 589.

Oliveira, G.L., Argiento, R., Loschi, R.H., Assunção, R.M., Ruggeri, F. and Branco, M.D. (2020). Bias correction in clustered underreported data. To appear at *Bayesian Analysis*.

Oliveira, G.P., Pinheiro, R.S., Coeli, C.M., Barreira, D. and Codenotti, S.B. (2012). Mortality information system for identifying underreported cases of tuberculosis in Brazil. *Revista Brasileira de Epidemiologia*, **15**(3), 468–477.

R Core Team (2015). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, (2015). Available at https://www.R-project.org/.

Santos, M.L, Coeli, C.M., Batista, J.L., Braga, M.C., de Albuquerque, M.F.P.M. (2018). Factors associated with underreporting of tuberculosis based on data from Sinan Aids and Sinan TB. *Revista Brasileira de Epidemiologia* [online], **21**:(e180019).

Schmertmann, C. and Gonzaga, M. R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*, **55**(4), 1363–1388.

Silva, G.D.M., Bartholomay, L., Cruz, O.G. and Garcia, L.P. (2017). Evaluation of data quality, timeliness and acceptability of the tuberculosis surveillance system in Brazil's micro-regions. *Ciência & Saúde Coletiva* [online], **22**(10), 3307–3319.

Sousa, L.M., Pinheiro, R.S. (2011). Unnotified deaths and hospital admissions for tuberculosis in the municipality of Rio de Janeiro. *Revista de Saúde Pública*, **45**(1), 31–39.

Sousa, M.G.G., Andrade, J.R.S., Dantas, C.F. and Cardoso, M.D. (2012). Investigação de óbitos por tuberculose, ocorridos na Região Metropolitana do Recife (PE), registrados no Sistema de Informação de Mortalidade, entre 2001 e 2008 (in Portuguese). *Cadernos Saúde Coletiva*, **20**(2), 153–60.

Stoner, O; Economou, T; Drummond, G. (2019). A Hierarchical Framework for Correcting Under-Reporting in Count Data. *Journal of the American Statistical Association*, **114**(528), 1481–1492.

World Health Organization (WHO). (2012). Assessing tuberculosis under-reporting through inventory studies. *WHO Library Cataloguing-in-Publication Data.* France, WHO/HTM/TB/2012.12, ISBN 978-92-4-150494-2.

World Health Organization (WHO). (2017). Global Tuberculosis Report 2017. *WHO Library Cataloguing-in-Publication Data.* Switzerland, WHO/HTM/TB/2017.23, ISBN 978-92-4-156551-6.

Chapter 3

Bayesian Dynamic Estimation of Mortality Schedules

Abstract[1]

The determination of the shapes of mortality curves, the estimation and projection of mortality patterns over time, and the investigation of differences in mortality patterns across different small underdeveloped populations have received special attention in recent years. The challenges involved in this type of problems are the common sparsity and the unstable behavior of observed death counts in small areas (populations). These features impose many difficulties in the estimation of reasonable mortality schedules. In this chapter, we present a discussion about this problem and we introduce the use of relational Bayesian dynamic models for estimating and smoothing mortality schedules by age and sex. Preliminary results are presented, including a comparison with a methodology recently proposed in the literature. The analyzes are based on simulated data as well as mortality data observed in some Brazilian municipalities.

Keywords: Bayesian smoothing, dynamic model, mortality curves, relational model.

3.1 Introduction

The mortality rate, life expectancy and other indicators of longevity are of fundamental importance to measure the health and well-being conditions of human populations. Methods for describing mortality patterns are common in demography, but the understanding of mortality evolution plays an important role in many other fields, such as actuarial science, epidemiology and genetics.

Demographic studies often use data related to the entire population. Because of this mortality studies are commonly performed at an aggregate level. More specifically, mortality data are usually available as the number of deaths and the population exposure to risk in a particular population (e.g., country, state, municipality, counties) and age group (e.g., sequential 5-year age intervals). In such context, the focus is to estimate the mortality rate, i.e. deaths per population at risk and age in each population. Typically, the age-specific mortality rates are calculated separately by sex or other characteristics, providing the so-called *mortality schedules*.

In human populations, mortality rate curves generally display regular patterns. The left panel of Figure 3.1 illustrates the usual shape for human mortality schedules considering data selected from the life tables available in the Human Mortality Database (HMD) [Wilmoth *et al.*, 2020]. The curves were obtained by fitting spline functions to the observed values of mortality rates by 1-year age intervals for each selected life table. Spline smoothing is considered in order to clarify the underlying shape. It can be seen that, in general: (i) the mortality rate tends to decay quickly after birth; (ii) then, it apparently remains stable until age 35; and (iii) an exponential growth can be noticed from there. Indeed, with basis on historical evidence, the previous features are commonly observed for a large range of human mortality curves. Alternatively, mortality curves are often analyzed on the logarithmic scale (right panel of Figure 3.1). In this case, the typical variations of mortality rates across ages can be better noted. In particular: (1) the mortality is relatively high in the first years of life; (2) then, it shows a steep decline between birth until a regular minimum around age 10; (3) the mortality curve typically reveals an "accident hump” at young adult ages (around, roughly, ages 15 to 30); (4) above age 30 the curve is fairly linear (in the log scale); and (5) at the oldest ages (say, above age 90) there is often a deceleration of mortality.

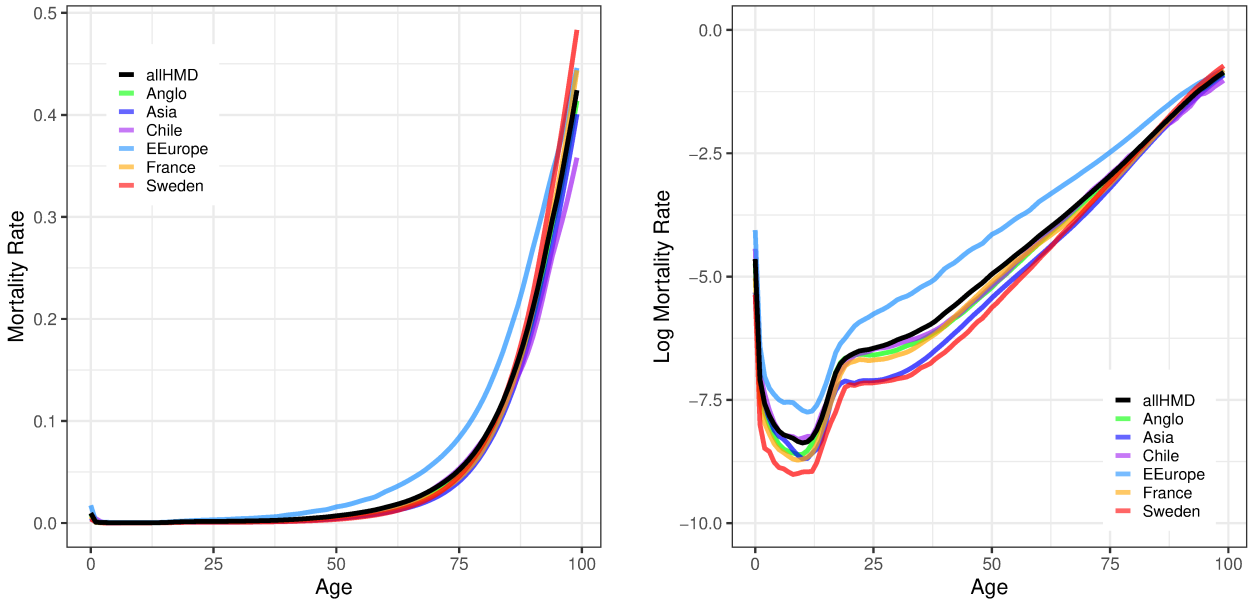


Figure 3.1: Seven mortality schedules by 1-year age intervals extracted from the life tables available in the Human Mortality Database: all countries, Chile, Sweden, France, Eastern European countries in HMD, Anglophone countries and Asian countries. Original scale (left) and logarithmic scale (right). Source: HMD [2015] through Gonzaga and Schmertmann [2016]'s website http://topals-mortality.schmert.net.

When dealing with large populations, specially in developed countries, the usual shape for the mortality curve is easily identified in general. That is the case for most life tables available in the HMD (e.g., see curves in Figure 3.1). However, that is not true when analyzing data from populations with poor data quality as, for instance, in developing countries that present incomplete coverage of the vital registration systems as well as errors in age declaration for both population and death counts.

Besides the data registration issues, the production of life tables is even more difficult when focusing in small populations. In this cases, the observed rates are often highly erratic and may have a great amount of null death counts (numerator of mortality rates) or even the lack of individuals exposure to risk in some age intervals (denominator of mortality rates).

In addition to the presence of extreme values, the mortality rates observed in small populations tend to present high variability across ages and sexes (male and female). Such data characteristics make it difficult to identify the true underlying mortality pattern. As an example, Figure 3.2 displays the log-mortality rates observed for some out of the 5,565 Brazilian municipalities. There is a higher variability in the smallest populations due to the greater occurrence of low and null counts, as a consequence of having a small number of individuals at risk in some ages. According to the national census of 2010, for 45% of the Brazilian municipalities the population is smaller than 10,000 inhabitants and for 90% of them it is smaller than 50,000. Figure 3.2 shows that it is possible to identify the shape of usual mortality schedules for municipalities in which population is greater than 100,000 inhabitants. It also shows that data noise is smaller for large populations. For small populations, besides the great variability, there are several age intervals with zero deaths count, specially at infant and young ages, making cumbersome the identification of the shape of the mortality schedules.

Naturally, the sparsity in observed data becomes even more severe when subnational groups are disaggregated by age and sex, usual procedure in life table estimation. Reliable measurements and comparative analysis of mortality levels, age patterns and sex differences for regional populations help to better understand health status at local levels and to guide policy definitions and changes in the targets for public investments. It may also help in the appropriate derivation of life expectancy and other measures used to perform population projections.

Some important goals of mortality modeling include describing the shape of mortality curves, estimating and projecting mortality patterns over time, and investigating differences in mortality patterns across different populations [Gompertz, 1825; Brass, 1971; Heligman and Pollard, 1980; Coale *et al.*, 1983; Lee and Carter, 1992; Dellaportas *et al.*, 2001; Dowd *et al.*, 2011; Li, 2014; Wilmoth *et al.*, 2012; Lima *et al.*, 2016; Alexopoulos *et al.*, 2019]. Because mortality schedules generally display regular patterns, smoothing approaches are a natural choice to analyze changes in mortality rates, age-structure decompositions and the construction of continuous life tables [Kashiwagi and Yanagimoto, 1992; Alexander *et al.*, 2017]. In this context, a common approach involves spline smoothing functions (see, e.g., Currie *et al.* [2004]; De Beer [2012]; Camarda [2012]; Gonzaga and Schmertmann [2016]; Alexander and Alkema [2018] and references there in).

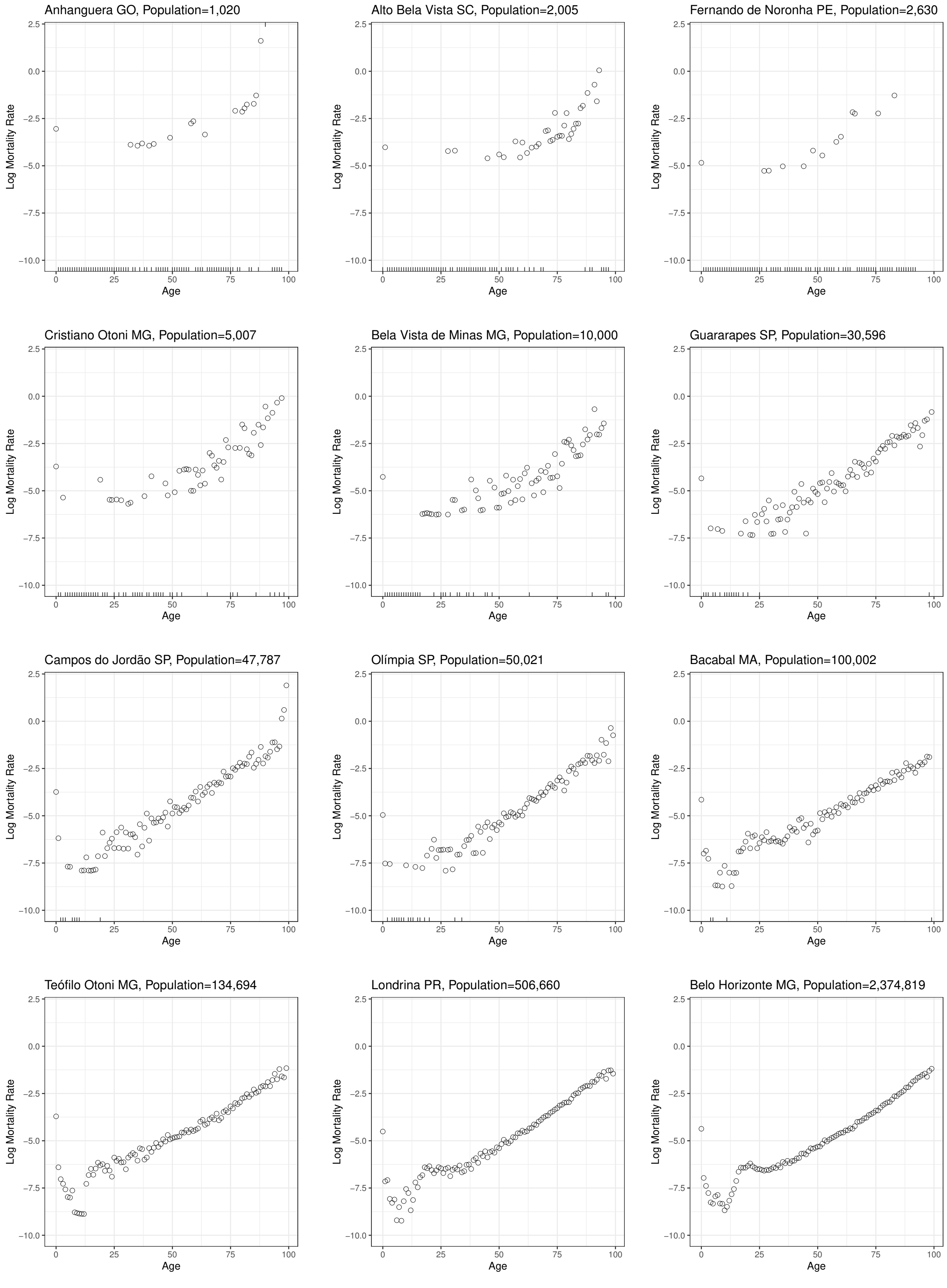


Figure 3.2: Observed mortality schedules (in the log scale) for selected Brazilian municipalities from 2009 to 2011, both sexes. Open circles represent the log-mortality rate for each single-year of age. Tick marks on the horizontal axis represent ages with no reported deaths or ages with no population at risk, which makes impossible to calculate the mortality rate. Data sources: IBGE (2010) and Brazilian Ministry of Health (http://www.datasus.gov.br).

For developed countries, where annual population updates tends to be available and vital registration systems have good quality, researchers have recently made important advances in statistical modeling and smoothness of complete mortality schedules in small areas (see Lima *et al.* [2016] and the references there in).

For less developed countries, the identification of discrepancies in mortality patterns across large regions (such as the states of a country) may not be a complicated task, especially because the populations size tend to be large. In the presence of abundant population, even simple parametric demographic models may produce good estimates for the target mortality rates. However, studies on complete age- and sex-specific mortality schedules at subnational levels are rare in developing countries due to lack of updated information. Because of that, demographers and statistical epidemiologists have proposed a voluminous literature for estimation of partial mortality schedules in developing countries, especially infant and child mortality indicators [Souza *et al.*, 2010; Walker *et al.*, 2012; Silva, 2013; Alkema and New, 2014; You *et al.*, 2015]. Some studies with focus in adult and old-age mortality levels in such areas can also be found [Kannisto, 1988; Timaeus, 1991; Hill *et al.*, 2009]. Many methods rely on indirect information from surveys or censuses to adequately address the volatility of these estimates due to data quality issues regarding vital registration systems.

As noted by Lima *et al.* [2016], a prominent and promising modeling approach for estimation of complete mortality schedules in less developed small populations involves the combination of statistical models with formal demography methods. The incorporation of demographic knowledge into statistical modeling frameworks is intended to ensures that estimates and projections have plausible patterns across ages.

Many authors have evaluated the efficacy of using parametric modeling structures combined with empirical regularities observed in mortality schedules obtained from external trustful sources [Brass, 1971; Coale *et al.*, 1983; Murray *et al.*, 2003; Wilmoth *et al.*, 2012; Alexander *et al.*, 2017; Gonzaga and Schmertmann, 2016; Clark, 2019]. Some authors refers to these sort of methods as relational models. The idea behind relational models is that complete age patterns in mortality, while inherently non-linear, exhibit strong similarities across different populations. Thus, patterns observed in high-quality data (e.g., mortality schedules in Figure 3.1) are used as a standard basis for producing estimates of mortality in populations where observed data are sparse or have poor quality, as in the small populations illustrated in Figure 3.2.

In this context of "borrowing strength" from an external mortality standard, recently, some authors have used data available in the Human Mortality Database [Wilmoth *et al.*, 2020] to build more flexible systems of life tables. For instance, the Bayesian hierarchical relational model proposed by Alexander *et al.* [2017] ensures a relatively smooth trend in mortality over time at the same time of sharing information across geographic areas. As an alternative, the approach of Gonzaga and Schmertmann [2016] linearly relates a mortality standard and penalized spline offsets to smooth mortality curves in small areas (populations).

The main goal of this work is to propose alternative models to estimate complete mortality schedules, specially in small underdeveloped populations. In order to estimate and smooth the associated mortality curves, we propose some Bayesian hierarchical dynamic models jointly with an underlying functional form that captures regularities in age patterns of mortality. The dynamic terms relates mortality rates across age intervals as an attempt to relatively smooth trend in the complete mortality curve. In turn, the use of a mortality standard is intended to penalize departures from the usual characteristic shapes of the target curves. We consider simulated and real datasets for different population sizes in order to exploit the advantages and drawbacks of the proposed methods. Competing approaches are compared in terms of adequacy and quality of smoothed mortality estimates.

3.2 Estimating and Smoothing Mortality Curves

For a given population i and age x assume that the total number of deaths is $Y_{i,x}$ and the total exposures individuals is $E_{i,x}$, $i = 1, \ldots, n$ and $x = 0, 1, \ldots, A$. Assume that the expected mortality rate is given by the product of exposures $E_{i,x}$ and an unknown parameter $\theta_{i,x}$ which represents the mortality risk in each population i and age x. A fully empirical (naive) estimate for the risk can be obtained by computing the observed death rates at the respective population and age $M_{i,x} = Y_{i,x}/E_{i,x}$ for all i, x. In terms of probabilistic modeling, a Poisson distribution with mean $E_{i,x}\theta_{i,x}$ is an usual choice to model the observed count $Y_{i,x}$ in which case an appropriate inferential method can be used to estimate the unknown parameter $\theta_{i,x}$.

To simplify the presentation, mortality data are commonly prepared as rectangular arrays. For each age in particular population, we have the number of deaths, the number of total exposures and the observed mortality rates arranged in $n \times A$ matrices $\boldsymbol{Y}$, $\boldsymbol{E}$ and $\boldsymbol{M}$, respectively. The rows are indexed by population and the columns are indexed by age. Without loss of generality, in this work we consider $A = 100$ corresponding to ages $0, 1, 2, \ldots, 98, 99$.

We aim to estimate a smoothed mortality curve for each population of interest. For doing that, we must estimate the related mortality rates at each age. We approach this problem by using a regression structure for the mortality rates in which the covariate corresponds to a mortality standard curve. Such a covariate is used to inform about the usual pattern observed for human mortality curves (see discussion in Section 3.1). In order to "borrow strength" from the relation of mortality rates in subsequent age intervals, we treat the sequence of age-specific mortality rates as a time series. Then, a dynamic structure across them is imposed through the model coefficients, thus providing a smoothed solution.

In our analysis, the standard mortality curve corresponds to the a mortality schedule calculated with basis in all life tables available in the Human Mortality Database in 2015 [HMD, 2015]. The information was obtained from the supplementary materials that were made available by Gonzaga and Schmertmann [2016] who also used such a standard in their model definition (http://topals-mortality.schmert.net). The standard mortality schedule is available separately for males and females. The standard schedule obtained by tacking the mean between the two sex-specific schedules is the ”allHMD” curve displayed in Figure 3.1.

In the following subsections we present the models we consider to analyze the datasets of interest. The first model is based on the assumption that observed death counts $\boldsymbol{Y}$ follows a Poisson distribution (Section 3.2.1) whereas the second approach models the observed log-mortality rates $\log(\boldsymbol{M})$ through a Gaussian distribution (Section 3.2.2).

3.2.1 Dynamic Poisson Model

Assume that the total number of deaths for population i and age x has the Poisson distribution

$$Y_{i,x} \sim Poisson(E_{i,x}\theta_{i,x}),$$

where $E_{i,x}$ is an offset corresponding to the total exposure individuals and $\theta_{i,x}$ denotes the mortality risk, $i = 1, , \dots, n$ and $x = 0, 1, \dots, A$. Consider that

$$\log(\theta_{i,x}) = \beta_{i,x} + \mu_{i,x} S_{i,x},$$

where $S_{i,x}$ is the mortality standard obtained from the HMD and $\beta_{i,x}$ and $\mu_{i,x}$ are the regression parameters for population i and age x. We assume the same standard for the n populations, that is, $S_{i,x} = S_x$, but a different standard can be applied for each population. The proposed model assumes a Markovian dependence among counts in neighboring age intervals. Such a dependence is included into the model through the Gaussian prior distributions elicited for $\boldsymbol{\beta}_i = (\beta_{i,1}, \dots, \beta_{i,A})$ and $\boldsymbol{\mu}_i = (\mu_{i,1}, \dots, \mu_{i,A})$ for all i. It is assumed that, given the associated precision parameters τ_β and τ_μ, the regression parameters are independent among populations, that is, $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n$ and $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_n$ are mutually independent for $i = 1, ..., n$. The proposed dynamic Poisson model is hierarchically represented as follows:

$$\begin{array}{rcll}
Y_{i,x}|\theta_{i,x} & \overset{ind}{\sim} & Poisson(E_{i,x}\ \theta_{i,x}), \quad for \;\; i = 1, \dots, n; \;\; x = 1, ..., A & (3.1) \\
\log(\theta_{i,x}) & = & \beta_{i,x} + \mu_{i,x} S_x & \\
\beta_{i,0}|\beta_i^0, \tau_\beta & \overset{ind}{\sim} & N\left(\beta_i^0, \tau_\beta\right) & \\
\beta_{i,x}|\beta_{i,x-1}, \tau_\beta & \overset{ind}{\sim} & N\left(\beta_{i,x-1}, \tau_\beta\right), \qquad for \;\; x = 1, \dots, A & \\
\mu_{i,0}|\mu_i^0, \tau_\mu & \overset{ind}{\sim} & N\left(\mu_i^0, \tau_\mu\right) & \\
\mu_{i,x}|\mu_{i,x-1}, \tau_\mu & \overset{ind}{\sim} & N\left(\mu_{i,x-1}, \tau_\mu\right), \qquad for \;\; x = 1, \dots, A & \\
\tau_\beta, \tau_\mu & \overset{iid}{\sim} & Gamma(0.01, 0.01), &
\end{array}$$

where $\boldsymbol{\beta}^0 = (\beta_1^0, ..., \beta_n^0)$ and $\boldsymbol{\mu}^0 = (\mu_1^0, ..., \mu_n^0)$ are vectors of initial values for the dynamic structure such that $\boldsymbol{\mu}^0 \sim N_n(\boldsymbol{0}, 0.1I_n)$ and $\boldsymbol{\beta}^0 \sim N_n(\boldsymbol{0}, 0.1I_n)$; with $N_n(\boldsymbol{\alpha}, \Psi)$ denoting the n-variate Gaussian distribution with mean vector $\boldsymbol{\alpha}$ and precision matrix Ψ (if $n = 1$ then n is removed from the notation) and I_n denotes the identity matrix of order n. The Normal distributions are parameterized in terms of mean and precision. The model was implemented in package *Nimble* [de Valpire *et al.*, 2017] from software R [R Core Team , 2015].

We propose to jointly model the n mortality schedules of interest. By doing that, a greater amount of sample information is guaranteed to estimate the precision parameters τ_β and τ_μ, which are shared by the n populations. We noted that such a strategy improve smoothness of the resulting mortality curves.

3.2.2 Gaussian Dynamic Linear Model

Although the response $Y_{i,x}$, the number of deaths at population i and age x is a count variable, another modeling strategy is the analysis of the associated observed mortality rate $M_{i,x} = Y_{i,x}/E_{i,x}$, where $E_{i,x}$ is the offset corresponding to the total exposure individuals. That is considered, for instance, by Lee and Carter [1992]; Dowd *et al.* [2011]; Wilmoth *et al.* [2012]; Clark [2019]. We consider the well-known Gaussian dynamic linear model (GDLM) to fit the observed mortality rates in the log scale. The GDLM is widely discussed in the statistical literature to model multivariate time series/longitudinal data (see e.g. West, Harrison and Migon [1985]; Migon *et al.* [2005]; Campagnoli *et al.* [2009] and references there in).

For each population $i = 1, ..., n$, the GDLM is considered to decompose the sequence of age-specific log-mortality rates $\log(M_{i,x})$, for $x = 1, ..., A$, as the sum of two components: an overall age-varying trend, $\beta_{i,x} + \mu_{i,x}S_{i,x}$, where $S_{i,x}$ is the standard mortality schedule obtained from the HMD, and an error component, $\epsilon_{i,x}$, that follows a zero mean Gaussian distribution with precision τ_ϵ. The dynamic evolution is taken across the age intervals $x = 0, 1, ..., A$. As in the Poisson case, we assume the same standard for the n populations, that is, $S_{i,x} = S_x$. Finally, the GDLM considered in this work is hierarchically specified as follows:

$$\begin{aligned}
\log(M_{i,x}) &= \beta_{i,x} + \mu_{i,x}S_x + \epsilon_{i,x}, \quad for \quad i = 1, \ldots, n; \; x = 1, ..., A \qquad (3.2)\\
\beta_{i,x} &= \beta_{i,x-1} + \delta_{i,x} \quad for \quad x = 1, \ldots, A \quad \text{and} \quad \beta_{i,0} = \mu_i^0 + \delta_{i,0}\\
\mu_{i,x} &= \mu_{i,x-1} + \omega_{i,x} \quad for \quad x = 1, \ldots, A \quad \text{and} \quad \mu_{i,0} = \mu_i^0 + \omega_{i,0}\\
\epsilon_{i,x}|\tau_\epsilon &\overset{iid}{\sim} Normal\,(0, \tau_\epsilon)\\
\delta_{i,x}|\tau_\delta &\overset{iid}{\sim} Normal\,(0, \tau_\delta)\\
\omega_{i,x}|\tau_\omega &\overset{iid}{\sim} Normal\,(0, \tau_\omega)\\
\tau_\epsilon, \tau_\delta, \tau_\omega &\overset{iid}{\sim} Gamma(0.01, 0.01),
\end{aligned}$$

where S, $\boldsymbol{\beta}^0$ and $\boldsymbol{\mu}^0$ and all other notations are as defined in Section 3.2.1. To perform posterior inference under this GDLM, we consider the Kalman filter and Kalman smoother algorithms available within the *dlm* package from software R [Campagnoli *et al.*, 2009]. The original function was modified in order to include a discount factor term. In the literature of dynamic linear models the *discount factor* is a known technique which has a crucial role in determining the influence of past observations throughout the estimation and forecasting of dynamic structure (for an in-depth discussion, see West and Harrison [1997], Section 6.3). In practice, the value of the discount factor is usually fixed between 0.9 and 0.99, or it is chosen by model selection diagnostics (for details, see Section 4.3.2 in Campagnoli *et al.* [2009]). We consider a discount factor with value 0.99 based on a predictive checking of model performance.

An important modeling feature regarding the GDLM approach is that the response variable $\log(M_{i,x})$ cannot be calculated in any population i and time x in which there were no observed death, that is, whenever $Y_{i,x} = 0$. The information is then treated as *missing observation*. It is known that occurrence of null counts is quite common when analyzing age-specific mortality data from small populations. The structure of dynamic linear models (DLM) is such that missing observations can be easily accommodated in the filtering recursion (for details, see Section 2.7.3 in Campagnoli *et al.* [2009]). As pointed out by these authors, in order to improve estimation for the mortality rates at ages x with observed null counts, we must consider to perform the inference process for multiple mortality schedules instead of estimating the model parameters for each population at a time. The idea multiple analysis is that, for each individual age x, the information contained in the non-missing components of vector $\boldsymbol{M}_x = (M_{1,x}, \ldots, M_{n,x})$ is accounted when estimating the dynamic parameters associated to the missing components. We performed the posterior estimation for the n mortality schedules jointly as an attempted to improve estimation for the age mortality rates in the presence of null death counts.

3.2.3 TOPALS model from Gonzaga and Schmertmann [2016]

Gonzaga and Schmertmann [2016] introduce a new method to estimate age-specific mortality rates in small populations. It corresponds to a Poisson regression model based on TOPALS, a relational model developed by De Beer [2012] for smoothing and projecting age-specific probabilities of death. Their approach estimates a complete schedule of log-mortality rates by adding a linear spline function to a pre-specified standard schedule. The spline represent additive offsets in specific ages. As in Section 3.2.1 denote by $Y_{i,x}$ and $E_{i,x}$ the death count and the total exposure population in an population i and age x, for $x = 0, 1, \ldots, A$, respectively. The TOPALS model from Gonzaga and Schmertmann [2016] is defined as follows:

$$\begin{aligned} Y_{i,x}|\theta_{i,x} &\overset{ind}{\sim} Poisson(E_{i,x}\exp\{\theta_{i,x}\}) \\ \theta_{i,x}|\boldsymbol{\alpha} &= S_{i,x} + \boldsymbol{B}_x\boldsymbol{\alpha}, \end{aligned} \tag{3.3}$$

where S is a standard mortality schedule; $\boldsymbol{B}_x$ is a 1×7 vector of constants corresponding to the xth line of a $A \times 7$ matrix $\boldsymbol{B}$ in which each column is a linear B-spline basis (for details, see Gonzaga and Schmertmann [2016] and their references); and $\boldsymbol{\alpha}$ is a 7×1 vector of parameters representing offsets to the standard schedule for specific knots $t_0, ..., t_6$ at ages $(0, 1, 10, 20, 40, 70, 100)$, respectively. For ages $x \in 0, 1, ..., 99$ and columns $k \in 0, ..., 6$ the basis functions in $\boldsymbol{B}$ are

$$B_{x,k} = \begin{cases} \frac{x-t_{k-1}}{t_k-t_{k-1}} & \text{if } \; x \in [t_{k-1}, t_k]\,, \\ \frac{t_{k+1}-x}{t_{k+1}-t_k} & \text{if } \; x \in [t_k, t_{k+1}]\,, \\ 0 & \text{otherwise.} \end{cases}$$

The authors argue that under such a parameterization, the $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_7)$ values represent additive offsets $(\theta_{i,x} - S_{i,x})$ to the log-rate schedule at exact ages $(0, 1, 10, 20, 40, 70, 100)$ and offsets change linearly with age between those knots. Parameters $\alpha_1, ..., \alpha_7$ are estimated by maximizing a penalized Poisson likelihood function for age-specific deaths, conditional on age-specific exposures. The penalty term added to the log-likelihood function increases as the linear spline offsets become less smooth. Gonzaga and Schmertmann [2016] considered such approach in order to avoid implausible fitted schedules for very small populations with very low numbers of deaths. TOPALS is applied to estimate mortality schedules in Brazilian municipalities and an illustration of the method is provided in Figure 1 of Gonzaga and Schmertmann [2016]..

All datasets and functions related to TOPALS adjustment were made accessible on the supplementary website http://topals-mortality.schmert.net. We consider the available material to fit the TOPALS model in the data analysis presented in the following sections and we compare its performance with the proposed dynamic models.

3.3 Simulation Experiments

In this section, we consider simulated datasets to compare the performance of the dynamic Poisson model introduced in Section 3.2.1, the Gaussian DLM presented in Section3.2.2 and the TOPALS model briefly reviewed in Section 3.2.3. To generate a complete mortality schedule, we have to determine a true underlying mortality mechanism, which must include information about the mortality rate and the associated total exposure population by age. For this purpose, we consider the mortality schedule observed for the São Paulo State (SP), Brazil, available from the case study presented in Section 3.4. As SP has a large population size, it provides a sufficiently smooth mortality curve to be used as a reference in our simulation study. In addition, the mortality curve of São Paulo State characterizes a modern population and it guarantees the simulation of mortality patterns potentially conformable to those we are interested in our application to Brazilian data.

We simulate nine populations with sizes 1,000; 2,000; 5,000; 10,000; 20,000; 50,000; 100,000; 500,000 and 1,000,000. In each case, the total exposures per 1-year age interval, $E_{i,x}$, was proportionally calculated to mimic the population pattern of SP. By considering the mortality rates observed for São Paulo as the true underlying risks, $\theta_{i,x}$, we generate the death counts $Y_{i,x}$ from a Poisson distribution such that $Y_{i,x} \sim Poisson(E_{i,x}\theta_{i,x})$, for $i = 1, , \dots, 9$ and $x = 0, 1, \dots, 99$.

The generated datasets were fitted under the three models and the estimates for the associated mortality rates, $\hat{\theta}_{i,x}$, were compared in terms of relative bias (RBias), square root of the mean squared error ($\sqrt{MSE}$) and mean absolute percentage error (MAPE) averaged over the $A = 100$ age-mortality rates in each simulated population. Such measures are calculated, respectively, as $RBias = \frac{1}{A}\left[\sum_{x=1}^{A}\left(\frac{\theta_{i,x}-\hat{\theta}_{i,x}}{\theta_{i,x}}\right)\right]$; $MSE = \frac{1}{A}\left[\sum_{x=1}^{A}\left(\theta_{i,x}-\hat{\theta}_{i,x}\right)^2\right]$ and $MAPE = \frac{1}{A}\left[\sum_{x=1}^{A}\left|\frac{\theta_{i,x}-\hat{\theta}_{i,x}}{\theta_{i,x}}\right|\right]$. Under the Poisson and Gaussian DLM models, for each generated dataset, two chains were run in the MCMC scheme considering a burn-in period of 100,000 iterations and a lag of 5,000 iterations was selected to avoid autocorrelated posterior samples, ending with a posterior sample of size 2,000 for each chain. The TOPALS model was fitted by using the functions available at http://topals-mortality.schmert.net. The three models were fitted on an Intel (R) Core (TM) i7-8550U 1.80GHz CPU with 8GB RAM. Respectively, the computational effort to run the algorithm for the TOPALS, Poisson and Gaussian models was around 1, 16 and 13 minutes proportionally per population.

The log-mortality rates observed for each simulated population are showed in Figure 3.3 along with the estimates provided by the three models. The same standard mortality schedule obtained from the HMD is considered when fitting the models. All models provided mortality schedules that evolves smoothly across ages, even for populations with a high frequency of null counts and a high noise in the observed data. The Gaussian DLM only provides a reasonable fit in large populations (greater than 100,000 inhabitants) with quite smooth mortality curves. In small populations, there are a great number of age intervals for which the number of deaths is zero, which are treated as missing data when fitting the Gaussian model. Such data feature imposes a restriction on the Gaussian model when applied for small populations.

Figure 3.3 shows that, as expected, Poisson and TOPALS models perform better than the Gaussian DLM in small populations. In general, these models tend to present similar shapes for the mortality schedules with more remarkable discrepancies in young ages, around the ”bathtub” pattern, and in the oldest ages. Such discrepancies are more evident in populations sized 1,000 and 5,000 with a visually better performance of the TOPALS model in such cases. For the population of size 2,000 the discrepancy between the two estimates is only evident in the latest ages, with a better performance of the Poisson model. For populations with 10,000 and 100,000 inhabitants the results are less similar for ages around the ”bathtub” pattern of the mortality schedules. Poisson and TOPALS models provided almost the same fit for populations with sizes 20,000, 50,000 and 1,000,000 .

The Poisson model seems to be more influenced by outliers than TOPALS model, specially in small populations. For instance, in the population with size 1,000 a single point was observed for younger ages. This point seems to be very influential, pulling the curve upwards in previous ages. Similarly, the mortality curve tends to be pulled downwards whenever a high frequency of null counts is observed in the latest ages. The influence of the mortality standard seems to be stronger in TOPALS model, making smaller the influence of atypical observations. That is a point which worth further investigation.

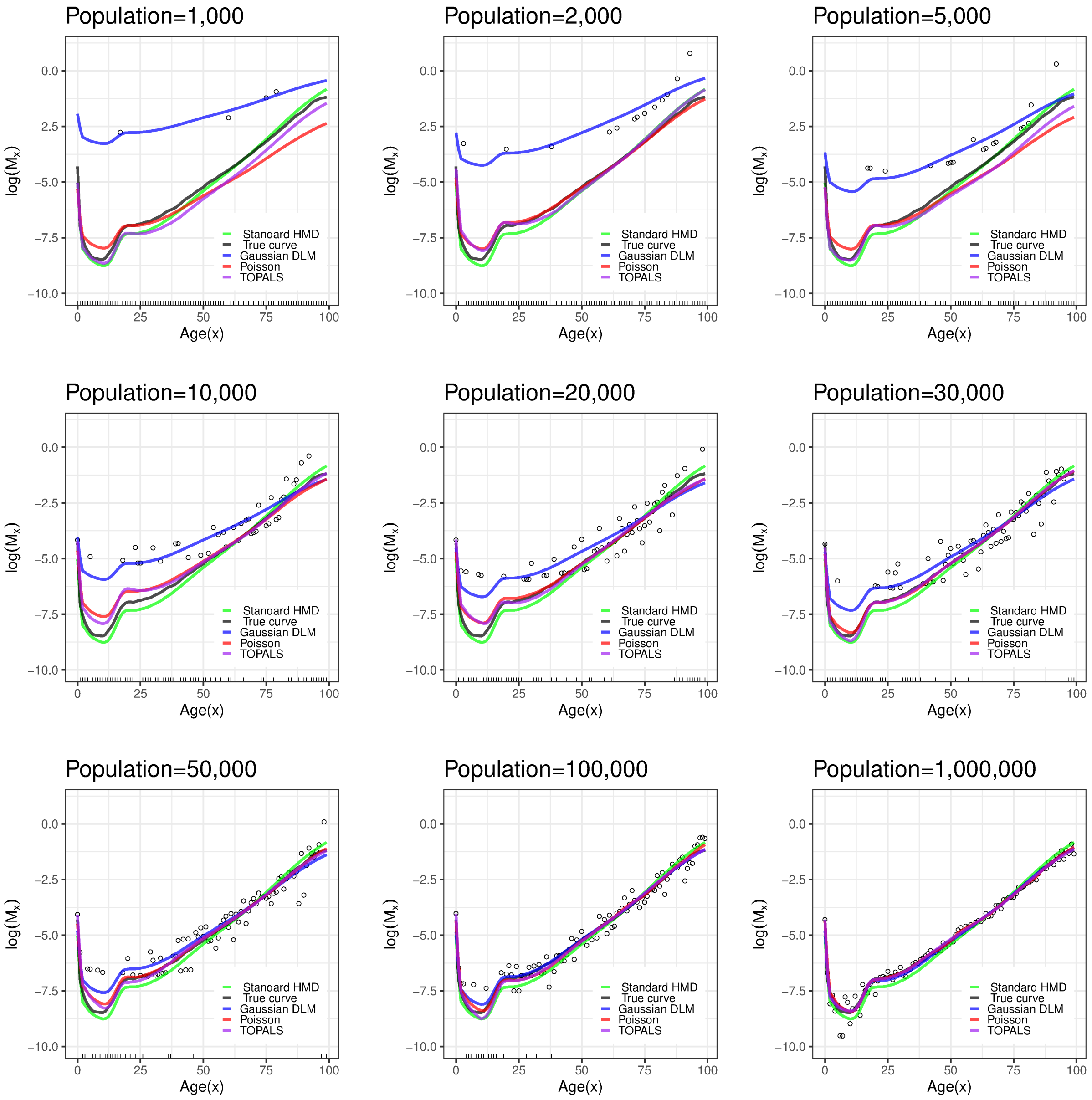


Figure 3.3: Simulated mortality schedules (in the log scale) for selected population sizes. Open circles represent the log-mortality rate for each single-year of age. Tick marks on the horizontal axis represent ages with no observed deaths. Black curve represents the true underlying mortality rates considered to generate the datasets. The simulated datasets were fitted under the three models described in Section 3.2: the dynamic Poisson model (red curve), the Gaussian DLM (blue curve) and the TOPALS model (purple curve). The green curve represents the standard mortality schedule assumed in all models (HMD, 2015).

In addition to the visual analysis available through Figure 3.3, we present in Table 3.1 some evaluation metrics comparing the three fitted models. We highlighted in bold the model presenting the smallest value for each metric. In general, Poisson and TOPALS models present the smallest values for the evaluation metrics. TOPALS performs better in the most populations according to $Rbias$ and $\sqrt{MSE}$ whereas the Poisson model figures as the best fitting model according to $MAPE$.

Table 3.1: Relative bias ($RBias$), square root of the mean squared error ($\sqrt{MSE}$) and mean absolute percentage error ($MAPE$) for the estimated log-mortality schedules in each simulated populations.

Population	$RBias$	$\sqrt{MSE}$	$MAPE$	$RBias$	$\sqrt{MSE}$	$MAPE$	$RBias$	$\sqrt{MSE}$	$MAPE$
	Poisson			TOPALS			Gaussian DLM		
1,000	-0.183	0.640	**0.078**	**-0.108**	**0.437**	0.105	0.601	3.304	0.598
2,000	**-0.007**	**0.201**	**0.012**	0.025	0.238	0.032	0.503	2.636	0.485
5,000	-0.143	0.512	**0.063**	**-0.111**	**0.390**	0.077	0.262	1.684	0.273
10,000	**0.004**	0.394	0.049	0.020	**0.309**	**0.020**	0.159	1.344	0.212
20,000	**-0.012**	**0.254**	**0.011**	-0.034	0.269	0.019	0.056	0.856	0.131
30,000	0.006	0.104	**0.007**	**0.004**	**0.091**	0.008	0.029	0.531	0.075
50,000	0.008	0.179	**0.010**	**-0.001**	**0.174**	0.011	0.013	0.396	0.050
100,000	0.007	**0.133**	**0.014**	**-0.005**	**0.133**	0.018	0.008	0.156	0.009
1,000,000	0.004	0.087	**0.006**	**-0.001**	**0.057**	**0.006**	**-0.001**	0.103	0.014

3.4 Application to Brazilian Municipalities

In this section we analyze data from selected Brazilian municipalities over calendar years 2009-2011. As some of them have a small population, we only consider Poisson and TOPALS models to fit the data. In both models, we assume the same standard mortality schedule considered in the simulation study. Population and deaths for the 5,565 Brazilian municipalities were obtained from the Gonzaga and Schmertmann [2016]'s project website http://topals-mortality.schmert.net. The data is available by 100 single-year ages and separately by sex and they was collected from the Demographic Census (2010) and from the Ministry of Health's Mortality Information System (SIM/Datasus), respectively. The 2010 census populations is used to estimate age- and sex-specific exposure populations over 2009-2011. It is worth noting that in the complete dataset, despite using three years of exposure, 49.2% of the 1,113,000 combinations of municipality, age and sex cells are null in number of deaths.

We present results in each selected municipality for both sexes (Figure 3.4) and separately for females (Figure 3.5) and males (Figure 3.6). By comparing Figures 3.5 and 3.6 it can be noticed the usual differences between mortality patterns for females and males. As noted in the simulation study, Poisson and TOPALS models tends to present similar shapes for mortality schedules. Both models provided smooth mortality curves even for populations with highly erratic data. For the three smallest selected populations, the models provided almost the same fit, except for females in Alto Bela Vista SC (population=1,000) and Fernando de Noronha PE (population=1,338). In the most cases, the biggest discrepancies between the two models, when they can be visually noted, occur at latest ages or at ages around the "bathtub pattern” of the mortality schedules. In agreement with findings of the simulation study, the Poisson model seems to be more influenced by outliers. This can be specially noted for high observed values for the mortality rates at the end of the ages' scale, which apparently pulls the estimates upwards (see graphs for Campos do Jordão SP, both sexes). A similar effect is noted when occurs a high frequency of sequential null counts in the ages' range, which tends to pull the curve downwards (see graphs for females in Alto Bela Vista SC and Fernando de Noronha PE).

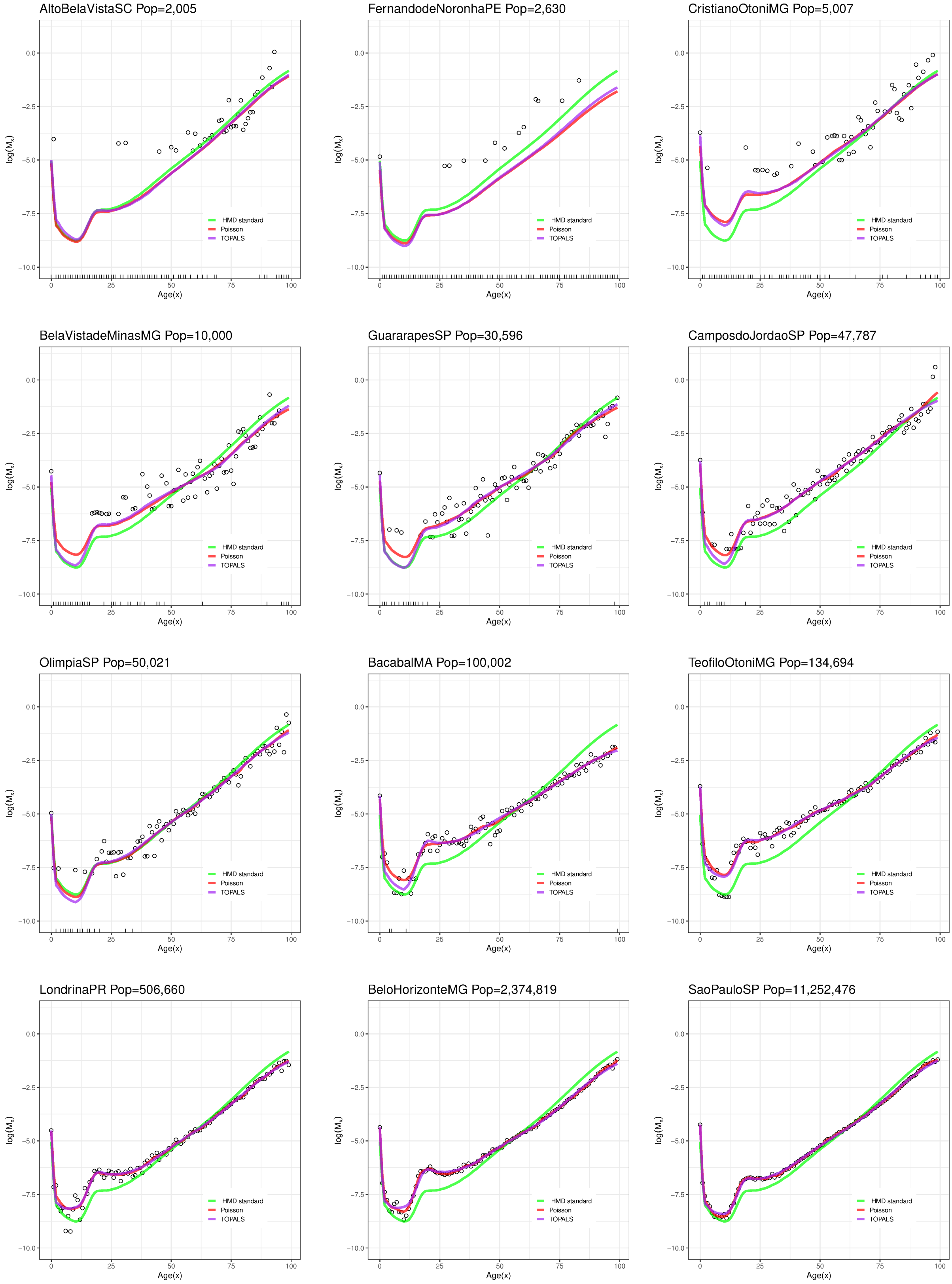


Figure 3.4: Estimated mortality schedules (in the log scale) for selected Brazilian municipalities, both sexes. Open circles represent the observed log-mortality rate for each single-year of age. Tick marks on the horizontal axis represent ages with no observed deaths. The observed data were fitted under the dynamic Poisson model (red curve) and the TOPALS model (purple curve). The green curve represents the standard mortality schedule assumed in both models (HMD, 2015).

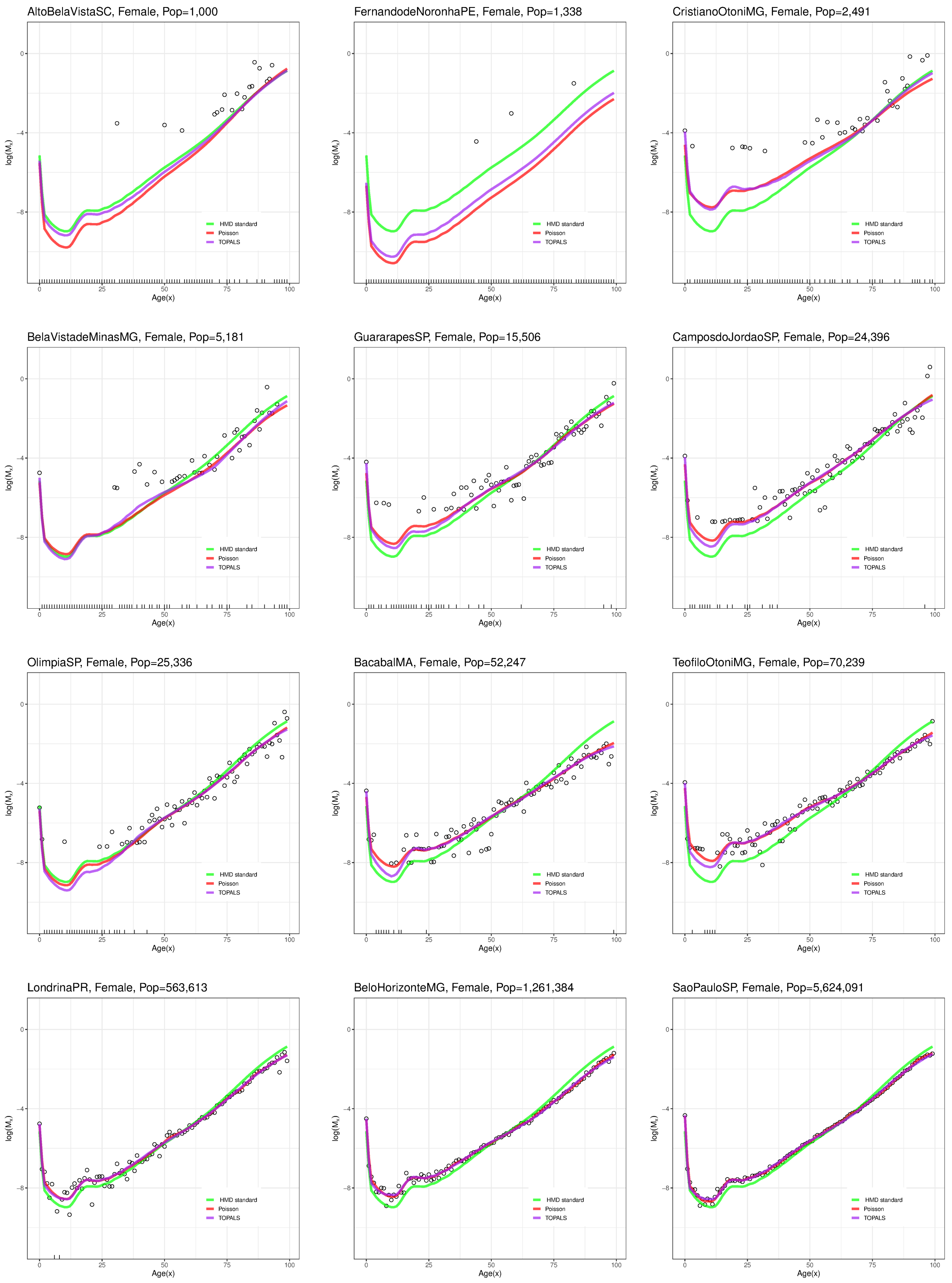


Figure 3.5: Estimated mortality schedules (in the log scale) for selected Brazilian municipalities, females only. Open circles represent the observed log-mortality rate for each single-year of age. Tick marks on the horizontal axis represent ages with no observed deaths. The observed data were fitted under the dynamic Poisson model (red curve) and the TOPALS model (purple curve). The green curve represents the standard mortality schedule assumed in both models (HMD, 2015).

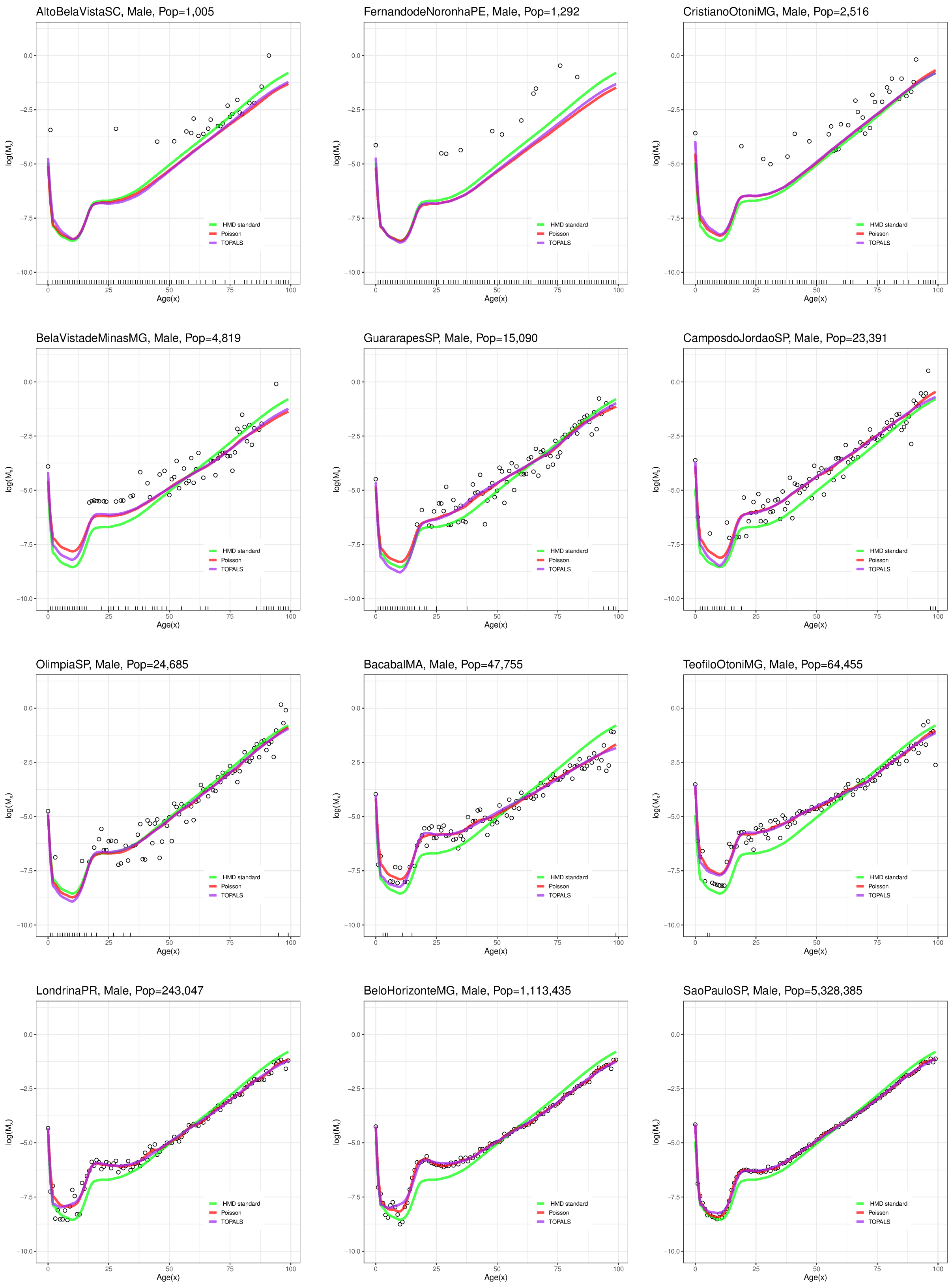


Figure 3.6: Estimated mortality schedules (in the log scale) for selected Brazilian municipalities, males only. Open circles represent the observed log-mortality rate for each single-year of age. Tick marks on the horizontal axis represent ages with no observed deaths. The observed data were fitted under the dynamic Poisson model (red curve) and the TOPALS model (purple curve). The green curve represents the standard mortality schedule assumed in both models (HMD, 2015).

3.5 Concluding Remarks

The reliable measurement and comparative analysis of mortality schedules for different populations helps to highlight differences among groups of people and guide analysts to understand what drives health disparities. For developing countries, specially in subnational geographic populations that do not have the resources to establish reliable death registration, this goal could be approached through the incorporation of empirical information to improve the estimates provided by usual parametric models. That is a common approach in the called relational demographic models. The problem with data coming from small and underdeveloped populations is the high occurrence of low or null counts, which impairs the estimation of the true underlying mortality rates by usual methods.

In this work we proposed two regression models with dynamic parameters to estimate the complete mortality schedules per 1-year age intervals. Inference is made under the Bayesian paradigm. Since mortality curves generally have a specific pattern, to prevent unrealistic estimates, we use a standard schedule as a covariate in the regression model, similar to the idea of relational models which are common in demography. Dynamic evolution across ages is included in order to provide smoothed estimates for the mortality schedules.

We consider a dynamic Poisson model to directly fit the observed mortality counts as well as the Gaussian dynamic linear model to model the observed log-mortality rates. The TOPALS model proposed by Gonzaga and Schmertmann [2016] is also fitted for comparison purpose. TOPALS allows the derivation of complete schedules of age mortality rates via mathematical adjustments to a specified standard schedule through penalized splines function. A simulation study is performed leading to interesting initial visualizations of the models performance.

In general, the Poisson and TOPALS models demonstrated to be more efficient as some of the analyzed data are sparse. Despite the Gaussian model showed to be a competitive model in large populations, it is not appropriated for small areas. Poisson and TOPALS models were applied to fit data from Brazilian municipalities. The Poisson model showed to be promising to estimate smoothed mortality schedules influenced to discrepant values in the observed mortality rates.

We believe that an implementation of the model as a generalized dynamic linear model [West, Harrison and Migon, 1985] may improve estimates, for instance, with the inclusion of discount factors in the covariance matrix of the dynamic parameters throughout the well-known Kalman filtering and Kalman smoothing estimation algorithms [Campagnoli *et al.*, 2009] or appropriate Markov chain Monte Carlo techniques [Gamerman, 1988; Schmidt and Pereira, 2011]. In this context, the use of the generalized dynamic Poisson model proposed by Schmidt and Pereira [2011] to fit time series of count data in epidemiological studies can be investigated. Their model have a time-dependent parameter that captures possible extra-variation present in the data and also zero-inflated versions are proposed. As such, their approach may provide interesting results in the context of mortality schedules estimation.

The simulation study presented in Section 3.3 is no longer exhaustive to support the preference for one specific model in a general case. Thus, the limited results must not be generalized. A broader simulation study must be performed in order to provide more confident evidence about the models performances in different scenarios and to support the previous findings discussed in this work. It includes a Monte Carlo study with more populations sizes and particular characteristics. Also, the application to a greater range of municipalities and other geographic aggregation levels can be included. A sensitivity study on the standard schedule choice must be performed as well. In this context, as addressed by Alexander *et al.* [2017], we aim to investigate the use of multiple mortality patterns, in particular the consideration of principal components analysis with basis on a large set of trustful mortality curves such as those available in the HMD. The definition of a multiple regression model based on the main principal components may increases flexibility in the estimates, potentially reducing the effect of outliers in the Poisson model evidenced by the studies developed so far.

Finally, we note that the excess of null death counts in some localities can indicate that the events is really rare or that there is a high level of underreporting of death counts. Therefore, the appropriate use of a zero-inflated Poisson model (Lambert [1992]; Piancastelli and Barreto-Souza [2019]; Gonçalves and Barreto-Souza [2020]), or other models which account for overdispersion, can be investigated as in Lima *et al.* [2016] besides the consideration of underreporting bias correction. We intend to formulate a zero inflated Poisson model which, in addition to taking into account a standard mortality schedule, is capable of accounting for the occurrence of under-registration. In particular, we aim to explore the clustering model present in Chapter 2 [Oliveira *et al.*, 2020] within the problem of mortality schedule estimation. In the demography literature, it is known that the level of underreporting varies between ages intervals. Particularly, specialists argue that such a problem is worse in infant ages than in old ages which, in turn, tends to present a higher level of under-registration than young and adult ages (see e.g. Schmertmann and Gonzaga [2018] and references there in). A clustering structure between subsequent ages intervals may be determined in order to apply the clustering model jointly with an adequate usage of a standard mortality schedule to ensure usual mortalities patterns.

References

Alexander, M., Zagheni, E. and Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography*, **54**(6), 2025–2041.

Alexander, M. and Alkema, L. (2018). Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research*, 38, 335–372.

Alexopoulos, A., Dellaportas, P., and Forster, J.J. (2019). Bayesian forecasting of mortality rates by using latent Gaussian models. *Journal of the Royal Statistical Society (Series A)*, **182**(2), 697–711.

Alkema, L. and New, J.R. (2014). Global estimation of child mortality using a Bayesian B-spline bias-reduction model. *The Annals of Applied Statistics*, 8(4), 2122–2149.

Brass, W. (1971). On the scale of mortality. In *Biological aspects of demography*, pp. 69–110. Taylor & Francis.

Camarda, C.G. (2012) MortalitySmooth: An R Package for Smoothing Poisson Counts with P-Splines *Journal of Statistical Software*, **50**(1), 1–24.

Campagnoli, P., Petrone, S. and Petris, G. (2009). Dynamic Linear Models with R. Springer-Verlag New York.

Clark, S.J. (2019). A General Age-Specific Mortality Model With an Example Indexed by Child Mortality or Both Child and Adult Mortality. *Demography* **56**, 1131–1159.

Coale, A. J., Demeny, P. G. and Vaughan, B. (1983). Regional Model Life Tables and Stable Populations (2nd edition). New York: Academic Press.

Currie, I.D., Durban, M. and Eilers, P.H.C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**(4), 279–298.

De Beer, J. Smoothing and projecting age-specific probabilities of death by TOPALS. (2012) *Demographic Research*, **27**(20), 543–592.

Dellaportas, P., Smith, A.F., and Stavropoulos, P. (2001). Bayesian analysis of mortality data. *Journal of the Royal Statistical Society (Series A)*, **164**(2), 275–291.

Dowd, K., Andrew, J.G.a., Blake, D., Coughlan, G.D., and Marwa, K-A. (2011). A Gravity Model of Mortality Rates for Two Related Populations. *North American Actuarial Journal*, **15**(2), 334–356.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalized linear models. *Biometrika*, **85**(1), 215–227.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality. *Philosophical Transactions*, 27, 513–585.

Gonçalves, J.N., and Barreto-Souza, W. (2020). Flexible regression models for counts with high-inflation of zeros. *METRON*, **78**, 71–95.

Gonzaga, M.R. and Schmertmann, C.P. (2016). Estimating age- and sex-specific mortality rates for small areas with TOPALS regression: an application to Brazil in 2010. *Revista Brasileira de Estudos de População*, **33**(3), 629–652.

Heligman, L. and Pollard, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, **107**(1), 49–80.

Hill, K., You, D. and Choi, Y. (2009) Death Distribution Methods for Estimating Adult Mortality: sensitivity analysis with simulated data errors. *Demographic Research*, **21**(9), 235–254.

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany), 2015. Available at http://www.mortality.org/. Accessed on: 01 Oct. 2015 by Gonzaga and Schmertmann (2016).

Kannisto, V. (1988). On the survival of centenarians and the span of life. *Population Studies*, **42**(3), 389–406.

Kashiwagi, N. and Yanagimoto, T. (1992). Smoothing Serial Count Data Through a State-Space Model. *Biometrics*, **48**(4), 1187–1194.

Lambert, D.(1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1), 1–14.

Lee, R. and Carter, L.R. (1992). Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, **87**(419), 659–671.

Li, N. (2014). Estimating Life Tables for Developing Countries. *United Nations Population Division, Department of Economic and Social Affairs*, Technical Paper No. 2014/4. United Nations New York. Available at www.unpopulation.org.

Lima, E., Queiroz, B.L., Missov, T. and Lenart, A. (2016). Estimate Mortality Curves in Small Areas : an Application to Municipality Data in Brazil. *Proceedings of the Annual Meeting of the Population Association of America 2016*, 1–17.

Migon, H., Gamerman, D., Lopez, H. and Ferreira, M. (2005). Bayesian dynamic models, in D. Day and C. Rao (eds), Handbook of Statistics, Vol. 25, Elsevier B.V., chapter 19, pp. 553–588.

Murray, C.J.L., Fergunson, B.D., Lopez, A.D., Guillot, M., Salomon, J.A. and Ahmad, O. (2003). Modified logit life table system: principles, empirical validation, and application. *Population Studies*, **57**(2), 165–182.

Piancastelli, L.S.C., and Barreto-Souza, W. (2019). Inferential aspects of the zero-inflated Poisson INAR(1) process. *Applied Mathematical Modelling*, **74**, 457–468.

Schmertmann, C.P. and Gonzaga, M.R. (2018). Bayesian Estimation of Age-Specific Mortality and Life Expectancy for Small Areas With Defective Vital Records. *Demography*, **55**, 1363–1388.

Schmidt, A.M. and Pereira, J.B.M. (2011) Modelling Time Series of Counts in Epidemiology. *International Statistical Review*, **79**(1), 58–69.

Silva, R.M.A. (2013). Papers on Indirect Mortality Estimation & Analysis in Low-Resource Settings. UC Berkeley Electronic Theses and Dissertations. Available at https://escholarship.org/uc/item/8xz5n880.

Souza, A., Hill, K. and Dal Poz, M. (2010). Sub-national assessment of inequality trends in neonatal and child mortality in Brazil. *International Journal for Equity in Health*, **9**(21).

Timaeus, I. M. (1991). Measurement of adult mortality in less developed countries: A comparative review. *Population Index*, **57**(4), 552–568.

You, D., Hug, L., Ejdemyr, S., Idele, P., Hogan, D., Mathers, C., Gerland, P., New, J.R. and Alkema, L. (2015). Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation. *The Lancet*, **386**(10010), 2275–2286.

Walker, N., Hill, K. and Zhao, F. (2012, 08). Child Mortality Estimation: Methods Used to Adjust for Bias due to AIDS in Estimating Trends in Under-Five Mortality. *PLOS Medicine*, **9**(8), 1–7.

West, M., Harrison, J. and Migon, H. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, **80**, 73–83.

West, M., and Harrison, J. (1997). Bayesian Forecasting and Dynamic Models, 2nd edition, Springer, New York.

Wilmoth, J., Zureick, S., Canudas-Romo, V., Inoue, M. and Sawyer. C. (2012) A flexible two-dimensional mortality model for use in indirect estimation *Popul Stud (Camb)*, **66**(2), 1–28.

Wilmoth, J.R., Andreev, K., Jdanov, D., Glei, D.A. and Riffe, T. with the assistance of Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., Vachon, P., Winant, C. and Barbieri, M. (2020). Methods Protocol for the Human Mortality Database. University of California at Berkeley (United States) and the Max Planck Institute for Demographic Research (Rostock, Germany). Last Revised: August 8, 2020 (Version 6). Available at https://www.mortality.org/Public/Docs/MethodsProtocol.pdf.

Chapter 4

A Random Censoring Poisson Model for Underreported Data

Abstract[1]

A major challenge when monitoring risks in socially deprived areas of under developed countries is that economic, epidemiological and social data are typically underreported. Thus, statistical models that do not take the data quality into account will produce biased estimates. To deal with this problem, counts in suspected regions are usually approached as censored information. The censored Poisson model (CPM) can be considered but all censored regions must be precisely known a priori, which is not a reasonable assumption in most practical situations. We introduce the random censoring Poisson model (RCPM) which accounts for the uncertainty about both, the count and the data reporting processes. Consequently, for each region we will be able to estimate the relative risk for the event of interest as well as the censoring probability. To facilitate the posterior sampling process, we propose a Markov chain Monte Carlo (MCMC) scheme based on the data augmentation technique. We run a simulation study comparing the proposed RCPM with two competitive models. Different scenarios are considered. RCPM and CPM are applied to account for potential underreporting of early neonatal mortality counts in regions of Minas Gerais State, Brazil, where data quality is known to be poor.

Keywords: Bayesian inference; Censoring; data augmentation; infant mortality; underreporting.

[1] The content of this chapter has been published in *Statistics in Medicine* [Oliveira *et al.*, 2017].

4.1 Introduction

Vital statistics such as fertility and mortality rates are typically calculated using deaths and births counts from a civil registration system. In less developed countries this method produces estimates that do not reflect the reality due to the large amount of those events that go underreported. Recognizing the low quality of data, the revision of population projections published by the United Nations in 2000 used the official registry counts only in 14% of the developing world [Hill, Choi and Timaeus, 2005]. As a fundamental statistic to monitor the population health condition, the infant mortality rate in such countries suffers from severe underestimation bias to the point of making the uncorrected statistics useless. For example, consider the infant mortality rates in Brazil. The worse off regions in terms of health care and economic development are also the best regions with respect to the infant mortality rates calculated directly from registry data. This is obviously incorrect because these regions have more obstacles to gain health care access and, once assessed, the health care system has low quality. A correction based on special information collected in the Census changes this completely, reverting the order and making them compatible with the obvious correct pattern [Simões, 1999].

Even with the advances achieved in recent years with relation to data collection systems, the underreporting of infant mortality and disease incidence has been high in the most of the underdeveloped and developing countries, such as Afghanistan [Viswanathan *et al.*, 2010], China [Xu *et al.*, 2014] and several other countries in African, Asia, Latin America and the Caribbean according to the World Health Organization [World Health Organization, 2006]. Although on a smaller scale, underreporting of mortality and disease cases can also be present in more developed countries such as Japan [Campbell *et al.*, 2011], United States of America Gould *et al.* [2002] and Norway [Alfonso *et al.*, 2015]. Also, underreporting of events is quite common in criminology data sets [Tibbetts and Hemmens, 2010].

An alternative for correcting vital statistics are the so called indirect methods which were developed during the 60's and 70's by demographers due to the need of studying the population patterns in Africa, Latin America and in the poor Asian countries [Brass *et al.*, 1968; Brass , 1996]. They are based on stable population theory and on decennial Census collected answers to questions about, for example, children ever born and their survival [Heligman, Finch and Kramer, 1978]. These methods rely on assumptions that may not be true such as the time invariance mortality and fertility rates. Also, these methods were developed for large populations such as countries or state-level populations, where statistical variability of rates were of less concern. Furthermore, the Census are carried out once every 10 years which is a long time before an update can be made. A second possibility to correct vital statistics is to resource to special data collection such through sample surveys or active search in hospitals and households [Heligman, Finch and Kramer, 1978]. However, these are very occasional and expensive to be used in a regular basis for monitoring a large number of regions.

From the statistical point of view, data recording problem (e.g., under or over-reporting) is a bias problem. We are used to fix estimators' bias by techniques such as Bartlett correction and bootstrap [Cordeiro and Cribari-Neto, 2014]. However, techniques to correct bias induced by data collection problems are less common. One major exception was the development of models to deal with censored or truncated observations, which is the main influence for this work.

Our motivation is the need to provide frequent and regular mortality risk estimates in small areas. We focus on the relative risk (RR) of early neonatal mortality (ENM) in 853 municipalities of Minas Gerais State (MG), Brazil. We consider data from two periods of time, from 1999 to 2001 and from 2012 to 2014, available at the hospital information system (Sistema de Informações Hospitalares, in Portuguese, abbreviated as SIH) of the Brazilian Health Ministry. Such data set is of major interest because, although data from SIH are more reliable than other available data sources for monitoring infant mortality in Brazilian municipalities, in Minas Gerais the ENM data recorded in such a system are still underreported [Campos *et al.*, 2007]. This was a problem in the past, and it still plagues the epidemiological analysis. The quality of information produced in Minas Gerais is inadequate, mainly in the north and northeast regions [Schramm and Szwarcwald, 2000], which are the socio-economically more deprived areas and present the worst social indicators of the state.

The standardized mortality ratio (SMR), given by the ratio between the observed and the expected counts in each region, is often a starting point to estimate the relative risk in epidemiological studies. Figure 4.1 displays the SMR associated to the ENM (left) and the Human Development Index (HDI) in 2000 (right) for the $n = 75$ regions of Minas Gerais. To avoid very small or zero counts, which leads to unstable estimates for the mortality rates, the 853 municipalities of Minas Gerais were previously grouped into 75 regions (see Figure 4.1). These 75 regions were created grouping contiguous municipalities such that at least one of them is able to provide high complexity health assistance (for further details see the Regionalization Directive Plan for health assistance proposed by the Government of MG in 2001/2004, available at https://www.nescon.medicina.ufmg.br/biblioteca/imagem/3022.pdf. The SMRs given in Figure 4.1 indicate that ENM risks in northern regions of the state are very low and comparable to those observed in highly developed countries. This result contradicts what it is expected by the epidemiologists since those regions experience the lowest HDIs in the state (see the rightmost map in Figure 4.1). This underestimation occurs because SMR does not take into account the underreporting in the ENM counts.

As an alternative to SMR mapping, Clayton and Kaldor [1987] proposed estimate the RR using a Poisson model and an empirical Bayes method that borrows information from all regions to obtain the posterior estimates. Such an approach, however, does not consider the spatial dependence inherent to the map. To explicitly take such dependence into account, Besag [1974] assumes that the relative risks are functions of random effects that follow a conditional autoregressive (CAR) distribution. Then, in order to achieve better estimates for the risks of ENM in Minas Gerais, the map in the middle in Figure 4.1 shows the analysis of our dataset using the intrinsic CAR Poisson model [Besag, York and Molliè, 1991]. Despite being a more sophisticated model, it does not overcome the underestimation of the relative risk in the poorest

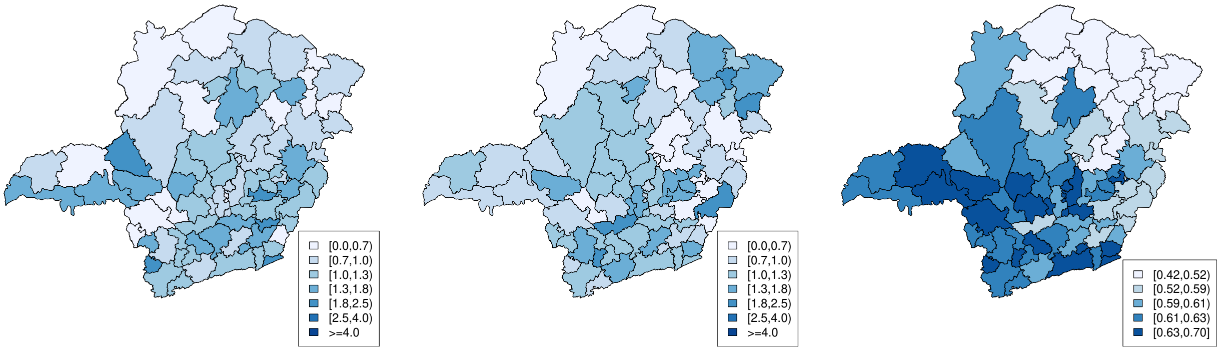


Figure 4.1: Early neonatal mortality RR estimates using SMR (left) and CAR model (middle); and the HDI in MG (right), 1999-2001.

regions in north and northeast of Minas Gerais. In fact, such a model aim at better fitting the spatial and non-spatial Poisson overdispersion but not to reduce the bias induced by the underreporting [Assunção, Potter and Cavenaghi, 2002; Assunção *et al.*, 2005].

To properly account for potential underreporting, Bailey *et al.* [2005] suggest treating data from suspected regions as censored information. Thus, models usually assumed to handle censored count data provide interesting approaches. Terza [1985] introduced the censored Poisson model (CPM) for data that experiences a constant censoring threshold. Caudill and Mixon Jr. [1995] extended such a model by assuming that the censoring threshold vary among areas. In both models, all censored counts are known and fixed *a priori*. Extending the models proposed by Besag, York and Molliè [1991] and Caudill and Mixon Jr. [1995], Bailey *et al.* [2005] introduced the censored CAR Poisson model for dealing with underreported data. A challenge in building this model is the precise specification of the regions where data are underreported - say, the censored regions. Information about such regions are usually obtained indirectly and *ad-hoc* procedures are considered to determine them. In their particular application, Bailey *et al.* [2005] defined a censoring criterion based on a social deprivation indicator.

Another way to overcome problems generated by potentially underreported data (briefly reviewed in Section 4.2.2) was considered by Winkelmann [1996] and Moreno and Girón [1998] - hereafter called Moreno and Girón model (MGM). They jointly model the uncertainty about the data generating and the data reporting processes by assuming an area-specific probability of each event being recorded. An approach quite similar to MGM is considered in Whittemore and Gong [1991], Powers, Gerlach and Stamey [2010] and Dvorzak and Wagner [2015] for estimating the rates of cervical cancer with data subject to underreporting. These three latter models assume that the intensity of the count process and the reporting probability are both dependent on covariates. Because of this, additional validation data giving information on the proportion of underreporting are required for the model identification.

Good performance for the MGM-type model is attained whenever we assume informative prior distributions for the reporting probability. The elicitation of such informative distributions requires the prior knowledge about the percentage of non-reporting that take place in each area. In the public health applications we deal with, this prior knowledge is unavailable as well as validation datasets as required in Dvorzak and Wagner [2015]. The only previous information we can count with is the fact that some areas are more prone to experience underreporting than others, but not how much censoring one can expect there.

Motivated by these applied situations, we develop a new censored Poisson model (Section 4.2.1). We propose a random mechanism to specify which regions are censored or, equivalently, regions where counts are underreported. The proposed model is named the random censoring Poisson model (RCPM). Using RCPM, we are able to estimate both, the probability of each area being censored and its relative risk for the event of interest. We elicit different prior distributions for the censoring probabilities. By assuming degenerate prior distributions for them, CPM arises as a particular case of RCPM. We develop an algorithm to sample from the posterior distribution which relies on the data augmentation strategy [Chib, 1992], simplifying the posterior sampling process.

We compare the posterior estimates for the relative risks provided by the proposed RCPM, the CPM and the MGM through simulation (Section 4.3). Data sets with different proportions of censoring are assumed. In addition, we perform a sensitivity analysis on the prior specifications for the reporting process under RCPM and MGM. Since the CPM consider that censored regions are known, we also evaluate the effect of assuming different censoring criteria - say, different degenerate prior distributions on the reporting process. We fit the proposed model to estimate the relative risk of early neonatal mortality in each region of the Minas Gerais State (Section 4.4). Results are compared with those obtained by fitting the CPM. Section 4.5 closes the paper with some final comments and main conclusions.

4.2 Models for Underreported Data

Suppose a map formed by n regions. Let N_i and Y_i be, respectively, the true but unobserved total number of events and the total number of events recorded (observed) at region i, $i = 1, \ldots, n$. To model underreported data, two approaches have being frequently considered. One of them assumes that there are two independent probabilistic mechanisms underlying the total number of events N_i. One of these mechanisms is associated with Y_i and the other is associated to the total number U_i of unrecorded data such that $N_i = Y_i + U_i$. It is assumed that each event j at region i is independently recorded with probability ϵ_i. Assuming these hypotheses and a distribution for N_i, the distribution of Y_i is obtained. Moreno and Girón model, presented in Section 4.2.2, assumes this model with a Poisson distribution for N_i. Another approach is the censored Poisson model (CPM) [Caudill and Mixon Jr., 1995]. It assumes that the map is composed by regions where N_i is not completely observed being these regions considered as censored regions. Under this approach, it is assumed that $Y_i \overset{D}{=} N_i$ in non-censored regions, and $Y_i \leq N_i$, otherwise. Moreno and Girón model requires the prior knowledge about the percentage of non-reporting that takes place in each area, which is not available in our case study. The CPM requires the precise prior specification of all censored regions, which is not a simple task in many practical situations. Our goal in this section is to introduce a model to account for potential underreporting, whenever none of this prior knowledge is available. We extend the CPM by assuming that the mechanism to define the censored regions is random. In our model we assume that $Y_i \leq N_i$ with probability π_i which is also estimated.

4.2.1 Random Censoring Poisson Model

Considere Y_i and N_i as previously defined. Denote by E_i the expected number of cases for the event of interest at region i, $\ i = 1, ..., n$. Assume that

$$N_i|\theta_i \overset{ind}{\sim} \text{Poisson}(E_i\theta_i), \tag{4.1}$$

where θ_i is the relative risk associated to region i. As in Caudill and Mixon Jr. [1995], instead of assuming that all variables N_i are completely observed, we consider that some of them may be right-censored (underreported), which means that for some regions the true non-observed value of N_i is higher or equal than the observed value Y_i. Let γ_i be the censoring indicator assuming value 1 if the count at region i is underreported and 0 otherwise. For our purpose, γ_i is a latent random variable having a Bernoulli distribution with censoring probability $\pi_i \in (0, 1)$, that is, we consider that the count N_i at region i is underreported with probability π_i. Since only Y_i is observable in each region, modelling is built based on the observed counts $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$.

Assume that, given $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$, the observed counts $\boldsymbol{Y}$ are independent. Consider also that the random censoring mechanism is independent of the counts in each area, that is, it is assumed that underreporting may occur either in regions with low or high counts. Under these assumptions, the joint model for $(\boldsymbol{y}, \boldsymbol{\gamma})$ is hierarchically obtained as

$$f(\boldsymbol{y} \mid \boldsymbol{\gamma}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \left[\frac{e^{E_i\theta_i}(E_i\theta_i)^{y_i}}{y_i!} \right]^{1-\gamma_i} \left[\sum_{y \geq y_i} \frac{e^{E_i\theta_i}(E_i\theta_i)^{y}}{y!} \right]^{\gamma_i} \right\}, \tag{4.2}$$

$$f(\boldsymbol{\gamma} \mid \boldsymbol{\pi}) = \prod_{i=1}^{n} \pi_i^{\gamma_i} \ (1 - \pi_i)^{1-\gamma_i}.$$

If the censored regions are fixed *a priori* such that the vector $\boldsymbol{\gamma}$ is known, expression (4.2) gives the likelihood related to the censored Poisson model by Caudill and Mixon Jr. [1995]. That is equivalent to assume the proposed model with a degenerate prior distribution for each π_i, which puts all positive probability mass in one, if region i is censored, or zero, if it is a non-censored region. Under this approach, however, Cromwell's rule is not followed and the information about the censoring mechanism cannot be updated by data information.

4.2.1.1 On the prior specifications

To complete our model specification, prior distributions must be elicited for the relative risks $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and for the censoring probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$. Obviously, several structures can be chosen for modeling the uncertainty about $\boldsymbol{\theta}$. Due to conjugacy properties, it is usual to assume that $\theta_i|\alpha_i, \phi_i \overset{ind}{\sim} Gamma(\alpha_i, \phi_i)$. Although not considered in this paper, other structures for the mean $\mu_i = E_i\theta_i$ of the Poisson model can be considered to account for extra variations in each region $i = 1, ..., n$. Besag, York and Molliè [1991], for instance, assumed a spatial structure for the relative risk by considering $\log \mu_i = \log E_i + \upsilon_i + s_i$, where υ_i and s_i denote, respectively, the non-spatially and the spatially structured random effects. The effects υ_i usually account for the dependence among the counts $\boldsymbol{Y}$ induced by unmeasured covariates while the s_i effects account for an explicit spatial dependence among them. A suitable linear combination of L available covariates $\boldsymbol{W} = (\boldsymbol{W}_1, ..., \boldsymbol{W}_L)$, where $\boldsymbol{W}_k^T = (W_{1k}, ..., W_{nk})$ for $k = 1, ..., L$; related to suspected risk factors can also be included for modeling the relative risks, so that $\log \mu_i = \log E_i + \sum_{k=1}^{L} \omega_k W_{ik} + \upsilon_i + s_i$.

A key point in the proposed RCPM is the modeling of the uncertainty about the censoring probabilities $\boldsymbol{\pi}$. A possible approach is to assume that $\pi_1, \ldots, \pi_n$ are independent and identically distributed with a non-informative uniform prior distribution, that is, $\pi_i \overset{ind}{\sim} \mathcal{U}(0, 1),\ i = 1, ..., n$. *A priori*, it is not realistic to assume that the censoring probabilities are uniformly distributed among the regions of interest. Instead, it is expected that regions with the worse social deprivation indicators have π_i close to 1.0 whereas regions with the best ones have π_i close to 0.0. If such prior information is available, one can elicit beta prior distributions for $\boldsymbol{\pi}$ in such a way that its hyperparameters reflect this expected behavior, for instance. We propose to assume that the censoring probability π_i can be modeled using a logistic regression model such that

$$\text{logit}(\pi_i) = \log\left(\pi_i(1 - \pi_i)^{-1}\right) = \beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta}, \tag{4.3}$$

where $\boldsymbol{\beta}^T = (\beta_1, ..., \beta_J)$ represents the fixed effects associated to J covariates measured in each region i, so that $\boldsymbol{X}_i^T = (X_{i1}, ..., X_{iJ})$. The set of covariates $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_J)$, where $\boldsymbol{X}_j^T = (X_{1j}, ..., X_{nj})$ for $j = 1, ..., J$; may be related to the socioeconomic/educational level or quality of health services in each region, for instance. To ensure the expected prior behavior of $\boldsymbol{\pi}$ among the regions, the prior specification of $\beta_j,\ j = 1, ..., J$, must take into consideration that the highest values for the censoring probabilities are associated to the regions with the worst social deprivation indicators. Sometimes it is possible to consider the covariates $\boldsymbol{X}_j$ to create an index representing the socio-economic/educational quality of the regions. If that is the case, it can be easier to order the values associated to such index and then the prior distribution for the β must be built putting positive probability mass in the appropriate part (positive or negative) of the real axis $\mathcal{R}$ depending on the ordering chosen for the index. For example, if the highest values for the social deprivation index are associated with the worse regions, then the parametric space of β must be $\mathcal{R}^+$. Similarly, if the highest values for the social deprivation indicator are associated with the better regions, then the parametric space of β must be $\mathcal{R}^-$.

4.2.1.2 Posterior inference

Under the structure assumed for the prior distribution of π_i in (4.3) and additionally assuming that β_0 and $\boldsymbol{\beta}$ are independent, the joint distribution of the complete model is given by

$$\begin{aligned} \mathbb{P}(\boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{\psi}) &= f(\boldsymbol{Y} \mid \boldsymbol{\gamma}, \boldsymbol{\theta}) f(\boldsymbol{\gamma} | \beta_0, \boldsymbol{\beta}) \mathbb{P}(\boldsymbol{\theta}) \mathbb{P}(\beta_0) \mathbb{P}(\boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} \Big\{ \left[\pi_i \left(1 - F_{Y_i|\mu_i}(y_i - 1)\right)\right]^{\gamma_i} \\ &\times \left[(1 - \pi_i) f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i} \mathbb{P}(\theta_i) \Big\} \; \mathbb{P}(\beta_0) \mathbb{P}(\boldsymbol{\beta}), \end{aligned} \tag{4.4}$$

where $\pi_i = (1 + \exp\{-(\beta_0 + \boldsymbol{X}_i^T \boldsymbol{\beta})\})^{-1}$, $\boldsymbol{\psi} = (\boldsymbol{\theta}, \beta_0, \boldsymbol{\beta})$ and $f_{Y_i|\mu_i}$ and $F_{Y_i|\mu_i}$ denote, respectively, the probability function and the cumulative distribution function (c.d.f.) of a Poisson distribution with mean $\mu_i = E_i \theta_i$.

The main focus is to infer about the relative risks $\boldsymbol{\theta}$ and about the censoring indicator latent variables $\boldsymbol{\gamma}$. It follows from Equation (4.4) that the posterior distributions for these parameters do not have a closed form and computational approaches should be used to approximate them. For parameter γ_i the posterior full conditional distribution is given by

$$\gamma_i | \boldsymbol{Y}, \boldsymbol{\gamma}_{-i}, \boldsymbol{\psi} \;\sim\; \text{Bernoulli}\left(A_i [A_i + B_i]^{-1}\right), \tag{4.5}$$

where $A_i = \pi_i \left[1 - F_{Y_i|\mu_i}(y_i - 1)\right]$, $B_i = [1 - \pi_i] f_{Y_i|\mu_i}(y_i)$ and $\boldsymbol{\gamma}_{-i}$ denotes the vector $\boldsymbol{\gamma}$ without the component i. As a particular case, if it is assumed that $\theta_i \overset{ind}{\sim} \text{Gamma}(\alpha_i, \phi_i)$, the posterior full conditional distribution of this parameter is given by

$$\theta_i | \boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{\psi}_{-\theta_i} \sim \begin{cases} \text{Gamma}\left(y_i + \alpha_i, \phi_i [E_i \phi_i + 1]^{-1}\right) & \text{if } \gamma_i = 0; \\ \text{Gamma}\left(\alpha_i, \phi_i\right) \times \left[1 - F_{Y_i|\mu_i}(y_i - 1)\right] & \text{if } \gamma_i = 1. \end{cases} \tag{4.6}$$

Sampling from the posterior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ by using a regular Gibbs sampler algorithm is highly inefficient because their f.c.d. in (4.5) and (4.6) depend on the cumulative distribution function of a Poisson distribution.

To obtain a more efficient sampling procedure in the presence of censored data we adopt the data augmentation technique [Tanner and Wong, 1987; Chib, 1992]. For the well-known Tobit model, in which the censored observations are fixed and known *a priori*, Chib [1992] proved that, whenever the data augmentation technique is used, the posterior inference for parameters of interest remains the same as in the initial model. In next section we extend Chib's proposal and develop an algorithm to sample from the posterior distributions whenever the censoring mechanism is random.

4.2.1.3 Data Augmentation for Posterior Sampling

In the context of censored models, data augmentation technique consists on including latent variables or unobserved data into the model in order to facilitate the computation procedures. To simplify the structure of the likelihood function in a right-censored model as the proposed RCPM, basically, the censored counts are replaced by augmented values, which are generated from a suitable truncated distribution in order to represent the true non-observed counts.

Consider a sample $\boldsymbol{Y} = (y_1, ..., y_n)$ in which n_c observations are censored (underreported) and $n_o = (n - n_c)$ observations are non-censored (correctly observed). Denote by $\boldsymbol{y}^c$ and $\boldsymbol{y}^o$ the set of censored and non-censored observations of $\boldsymbol{Y}$, respectively. Suppose that along with the censored observations, $\boldsymbol{y}^c$, we have available the corresponding latent data $\boldsymbol{Z}$, which is a vector of dimension $n_c \times 1$. Let $\mathcal{C}$ denote the index set of the censored observations. Following the approach of Chib [1992], we assume that, given $(\boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{\psi})$, $\boldsymbol{Z}$ is a collection of independent random variables such that, for all $i \in \mathcal{C}$, Z_i has a Poisson distribution with mean $\mu_i = E_i\theta_i$ truncated from below at the underreported value y_i, whose conditional probability function is given by

$$f(Z_i = z_i|\boldsymbol{Y}, \boldsymbol{\psi}, \gamma_i = 1) = f_{Z_i|\mu_i}(z_i)[1 - F_{Z_i|\mu_i}(y_i - 1)]^{-1}, \tag{4.7}$$

for $z_i = y_i, y_i + 1, ...,$ where $\boldsymbol{\psi}$, $f_{Z_i|\mu_i}$ and $F_{Z_i|\mu_i}$ are as defined in (4.4).

By doing that, we now have a vector of augmented data $\boldsymbol{Y}^z = (y_1^z, ..., y_n^z)$, which corresponds to the original collection of data $\boldsymbol{Y}$ with $\boldsymbol{y}^c$ replaced by $\boldsymbol{z}$ generated from (4.7). Consequently, the joint distribution under the complete data-augmented model is given by

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{\psi}) &= f(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\gamma})f(\boldsymbol{\gamma}|\beta_0, \boldsymbol{\beta})f(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\gamma})\mathbb{P}(\boldsymbol{\theta})\mathbb{P}(\beta_0)\mathbb{P}(\boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} \left\{ \left[\pi_i f_{Z_i|\mu_i}(z_i)\right]^{\gamma_i} \left[(1 - \pi_i)\, f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i} \mathbb{P}(\theta_i) \right\} \mathbb{P}(\beta_0)\mathbb{P}(\boldsymbol{\beta}),\end{aligned}$$

where π_i, $\boldsymbol{\psi}$, μ_i, $f_{Y_i|\mu_i}$ and $F_{Z_i|\mu_i}$ are as previously defined in (4.4).

We now must obtain posterior samples of $(\boldsymbol{\psi}, \boldsymbol{\gamma}, \boldsymbol{Z})$. The most important point is that the conditional probability function of the latent data is available in a tractable form (see expression (4.7)) and the data-augmented posterior distribution $\mathbb{P}(\boldsymbol{\psi}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{Z})$ do not involves a cumulative probability function. Therefore, it has a more simple form than $\mathbb{P}(\boldsymbol{\psi}, \boldsymbol{\gamma}|\boldsymbol{Y})$ and generates the same posterior inference for the parameters of interest. In a general case, the posterior full conditional distributions of γ_i, θ_i, β_0 and $\boldsymbol{\beta}$ under the new structure become

$$\mathbb{P}(\gamma_i|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\gamma}_{-i}, \boldsymbol{\psi}) \propto \left[\pi_i f_{Z_i|\mu_i}(z_i)\right]^{\gamma_i} \left[(1 - \pi_i)\, f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i}, \tag{4.8}$$

$$\mathbb{P}(\theta_i|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{\psi}_{-\theta_i}) \propto \left[f_{Z_i|\mu_i}(z_i)\right]^{\gamma_i} \left[f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i} \mathbb{P}(\theta_i), \tag{4.9}$$

$$\mathbb{P}(\beta_j|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\psi}_{-\beta_j}) \propto \prod_{i=1}^{n} f(\gamma_i|\beta_0,\boldsymbol{\beta})\mathbb{P}(\beta_j),\ j=0,1,...,J. \tag{4.10}$$

A Gibbs sampler strategy that can be considered is to sequentially sample from the distributions given in (4.7), (4.8), (4.9) and (4.10). Implementing such a MCMC scheme (results not shown), we verified that the algorithm is inefficient due to slow convergence. Because of the strong dependence between $\boldsymbol{Z}$, $\boldsymbol{\gamma}$, β_0 and $\boldsymbol{\beta}$, especially the dependence between $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, convergence is improved if such parameters are jointly sampled. Hence, we consider their joint posterior full conditional distribution which is given by

$$\begin{aligned}\mathbb{P}(\boldsymbol{Z},\boldsymbol{\gamma},\beta_0,\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{\theta}) \ &\propto \ \prod_{i=1}^{n}\Big\{\left[\pi_i f_{Z_i|\mu_i}(z_i)\right]^{\gamma_i} \\ &\times \ \left[(1-\pi_i)\,f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i}\Big\}\,\mathbb{P}(\beta_0)\mathbb{P}(\boldsymbol{\beta}),\end{aligned} \tag{4.11}$$

and a Metropolis-Hastings (M-H) step is thus needed to sample from (4.11). Therefore, we suggest the following MCMC scheme for posterior sampling:

1. Sample $(\boldsymbol{Z},\gamma,\beta_0,\boldsymbol{\beta})$ from $\mathbb{P}(\boldsymbol{Z},\boldsymbol{\gamma},\beta_0,\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{\theta})$ in (4.11) using a M-H step;

2. Sample θ_i from $\mathbb{P}(\theta_i|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\psi}_{-\theta_i})$ in (4.9), for $i=1,...,n$.

The proposal distribution in the Metropolis-Hastings (M-H) step assumes that γ, β_0, and $\boldsymbol{\beta}$ are independent. We assume normal distributions for β_0, and $\boldsymbol{\beta}$ and a Bernoulli distribution for γ_i all of them centered in the value generated in the previous step. Particularly, if we assume that, *a priori*, $\theta_i \overset{ind}{\sim}$ Gamma(α_i,ϕ_i), then its posterior full conditional distribution under the augmented model is $\theta_i|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\psi}_{-\theta_i} \sim \text{Gamma}\left(z_i+\alpha_i,\phi_i[E_i\phi_i+1]^{-1}\right)$ if $\gamma_i = 1$ and, if $\gamma_i = 0$, $\theta_i|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\psi}_{-\theta_i} \sim \text{Gamma}\left(y_i+\alpha_i,\phi_i[E_i\phi_i+1]^{-1}\right)$. In opposite to what was presented in Equation (4.6), under this particular gamma prior distribution the posterior full conditional distribution of θ_i has now closed form for both censored and non-censored regions.

4.2.2 Moreno and Girón's Model

Let N_i be the true but unobserved number of events at region i, $i=1,\ldots,n$. In the MGM [Moreno and Girón, 1998], it is assumed that $N_i \overset{ind}{\sim}$ Poisson$(E_i\theta_i)$, where E_i and θ_i are as previously defined. It is also assumed that an event occurring at region i is independently recorded with probability $(1-\epsilon_i)$. Given N_i and ϵ_i, the observed (recorded) count Y_i is assumed to be sampled from the true number of events by binomial trials such that $Yi|Ni,\epsilon_i \overset{ind}{\sim}$ Binomial$(N_i, 1-\epsilon_i)$. Consequently, given θ_i and ϵ_i, the distribuition of the observed count Y_i, $i=1,...,n$ is

$$Y_i|\theta_i,\epsilon_i \overset{ind}{\sim} \text{Poisson}(E_i\theta_i(1-\epsilon_i)).$$

Inference is done under the Bayesian paradigm. Therefore, to complete the model specification, it is required the elicitation of prior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$. It is assumed prior independence between $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ to simplify the elicitation of prior distributions for such parameters. For the relative risk $\boldsymbol{\theta}$ one can elicit that, for instance, $\theta_i|\alpha_i, \phi_i \overset{ind}{\sim} Gamma(\alpha_i, \phi_i)$, $i = 1, ..., n$.

Moreno and Girón [1998] discuss some ways to set prior distributions for $\boldsymbol{\epsilon}$. Informative prior distributions for ϵ_i can be built if we obtain from the experts pieces of information about the proportion of unrecorded events at region i. We may assume that $\text{logit}(\epsilon_i) = \beta_0 + \boldsymbol{\beta X}_i$, where $\boldsymbol{X}_i$ represents information from covariates related to the socio-economic or educational level in each region and the prior distribution for $(\beta_0, \boldsymbol{\beta})$ should reveals how much such factors influence the probability of any event being recorded. We can also elicit $\epsilon_i \overset{ind}{\sim} \text{Beta}(\kappa_i, \Psi_i)$, where κ_i and Ψ_i are chosen in a way that the prior expectation of ϵ_i reflects reasonably well our prior knowledge about the proportion of unrecorded counts.

4.3 Simulation Study

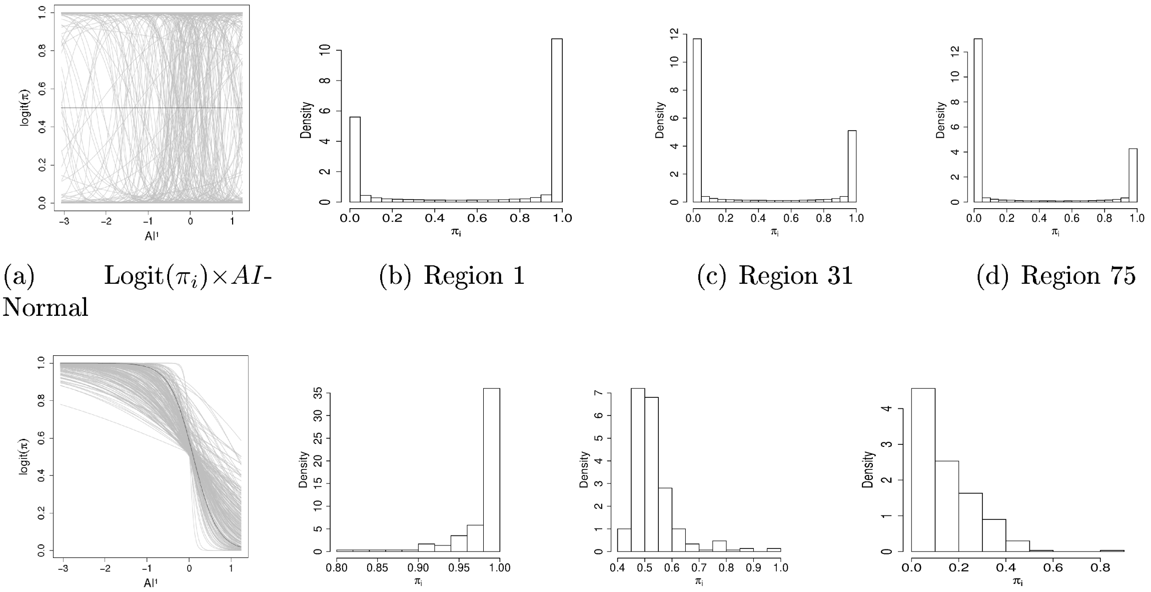
We perform a Monte Carlo study in order to compare the performance of the proposed model (RCPM), the censored Poisson model (CPM) [Caudill and Mixon Jr., 1995] and the model presented by Moreno and Girón (MGM) [Moreno and Girón, 1998]. Different scenarios are considered. To mimic the case study to be presented in Section 4.4, we consider the map of Minas Gerais State with the goal of estimating the relative risk associated to an event in the $n = 75$ regions of the state. We also evaluate the performance of the proposed model in correctly identifying the regions that experience underreporting (the censored regions).

We consider 100 replications of each data set in all scenarios. Data are generated from Poisson distributions so that $N_i|\theta_i \overset{ind}{\sim} \text{Poisson}(E_i\theta_i)$. The expected number of cases E_i is assumed to be known and equal to that available for the case study to be discussed in Section 4.4. However, differently from what occurs in that case study, here E_i is free of underreporting. The true relative risk in each region is given by $\theta_i = \exp\{5.71 + 0.31Lat_i\}$, where Lat_i is the latitude of the centroid of region i. Consequently, the relative risk increases from the South to the North varying from 0.3 to 3.0.

Scenarios differ by the percentage of censored regions and the censoring level in each observation. The censoring level $\delta \in [0, 1]$ is the proportion of the generated count N_i that is reported in each censored region, that is, in each censored region, the observed count Y_i is taken as the smallest integer greater or equal than $N_i \times \delta$. It means that $(1-\delta) \times 100\%$ of the true generated count will be missed (unreported) in the censored regions. The smallest the δ, the lowest is the data information available to estimate the relative risk in the censored areas. In order to evaluate the effect of the amount of underreporting, we consider two different censoring levels: $\delta = 0.8$ and $\delta = 0.6$. Also to mimic our case study, to define the censored regions, we built three criteria based on the available adequacy index (AI) introduced by França *et al.* França *et al.* [2006]. The AI assesses the quality on infant mortality data collected in Minas Gerais and was built considering four available indicators of the quality on birth and mortality data recorded in each regions of the state: the standartized global mortality rate and its standard deviation, the standard deviation of the global natality rate and the proportion infant deaths with ill-defined cause.

The adequacy index ranges from -83.74 to 100.0 assuming negative values for eight regions of Minas Gerais, all of then located in the North and Northeast part of the state. The smallest the AI, the worst is the data quality in the region. Because the AI summarizes all available information about the data reporting process in our case study, it is used as the sole covariate in (4.3) for modeling the censoring probabilities under RCPM. We consider as censored, the regions for which the $AI \leq 0.0$ (Scenario 1), $AI \leq 20.0$ (Scenario 2) and $AI \leq 45.0$ (Scenario 3) thus establishing that the proportion of censored regions are, respectively, 11%, 23% and 36% for Scenarios 1, 2 and 3. The censored regions are shown in the top maps in Figure 4.4. Results for scenarios with different levels of censoring δ among the censored regions and also scenarios with proportions of censored regions equals to 0% and 75% are presented in the supplementary material.

To analyse the data sets, we fit ten different models, one of them being the regular Poisson model. The three approaches for the proposed model (RCP1-RCP3) differ in the way the prior distribution for the censoring probabilities $\boldsymbol{\pi}$ are assigned. In RCP1, we assume a distribution with flat prior information about the censored regions by eliciting $\pi_i \overset{iid}{\sim} \mathcal{U}(0,1)$. Thus, with this non-informative prior, it is expected *a priori* that around 50% of the regions are censored. In RCP2 and RCP3 the censoring probabilities are modelled by a logit function such that $\text{logit}(\pi_i) = \beta_0 - \beta_1 \text{AI}_i^1$ in which AI_i^1 represents the standardized adequacy index. An usual approach in regression models is to consider flat normal prior distributions for the coefficients β_0 and β_1. Following this strategy, in RCP2 we assume $(\beta_0, \beta_1) \sim \mathcal{N}_2(\mathbf{0}, 100\boldsymbol{I}_2)$. The top row of plots in Figure 4.2 exemplifies this prior choice. The first plot shows the logistic curves resulting from sampling from the prior distribution for (β_0, β_1). The solid line represents the curve with the parameters set at their expected values (zero, in this case). The non-informative normal prior distributions for β_0 and β_1 induce a poor prior distribution for π_i which is symmetric with bathtub shape and concentrating the most of its probability mass around one and zero in all regions. The three other plots in the first row in Figure 4.2 show this induced prior distribution for region 1, with the lowest AI, region 31, with the average AI, and region 75, with the highest AI. In RCP3, we consider informative log-normal prior distributions for the coefficients β_0 and β_1 by assuming $\beta_0 \sim LN(-1.65; 0.95)$ and $\beta_1 \sim LN(0.63; 0.60)$ independently. The results are shown in the bottom row of plots in Figure 4.2. The prior distribution for π_i induced by such informative log-normal prior distributions for β_0 and β_1 concentrate probability mass in different values depending on the adequacy index of the region. For instance, for region 75, with the highest AI, the prior π_i concentrates its probability mass around zero, pointing out that this region is most probably non-censored. For region 1, with the lowest AI, the induced prior distribution of π_i concentrates the most of its probability mass around 1, as expected. In the case study presented in Section 4.4 this distribution is capable of describing reasonably well our prior knowledge about the censored regions.



(a) Logit(π_i)×AI-Normal

(b) Region 1

(c) Region 31

(d) Region 75

(e) Logit(π_i) ×AI-Log-Normal

(f) Region 1

(g) Region 31

(h) Region 75

Figure 4.2: Logit(π_i) *versus* AI and the induced prior distribution for π_i in regions with the lowest (Region 1), the median (Region 31) and the highest (region 75) adequacy index assuming normal (top) and lognormal (botton) prior distributions for β_0 and β_1.

The three different formulations for Moreno and Girón's model (MG1-MG3) differ in the way we assigned the prior distributions for the probabilities of each event being recorded at region i, ϵ_i. In MG1, we assumed that $\epsilon_i \overset{iid}{\sim} \mathcal{U}(0,1)$ which means that around 50% of the counts are not recorded in each region. For MG3 we elicit $\epsilon_i \overset{ind}{\sim} \text{Beta}(\kappa_i, AI_i^2)$, where AI_i^2 represents the adequacy index rescaled to be positive and κ_i in chosen in a way that the prior expectation of the components of $\boldsymbol{\epsilon}$ belongs to the interval $(0.02; 0.33)$ and the prior variance vary approximately from 0.002 to 0.060. This assumption guarantee that each ϵ_i is not far from the true $1-\delta$ associated to the region i for all scenarios. In MG2 we assume that $\text{logit}(\epsilon_i) = \beta_0 - \beta_1 \text{AI}_i^1$ in which, *a priori*, $(\beta_0, \beta_1) \sim \mathcal{N}_2(\mathbf{0}, 100\boldsymbol{I}_2)$. The prior for ϵ_i induced by this specification also has the bathtub shape shown in the top row of Figure 4.2 and it is less informative than the one considered in MG3 which concentrate the most of its probability mass around the true δ.

We also consider three approaches for the censored Poisson model (CP1-CP3) by assuming different criteria for prefixing the censored regions. The use of such fixed criteria for censoring the regions is equivalent to assign a degenerate prior distribution for the censoring probability π_i in the proposed RCPM. The CP3 correctly prefix 100% of the truly censored and non-censored areas considered in the data generation. It is the most informative scenario and clearly unrealistic. However, it serves as a golden standard about the prior information on the censored status for the regions. In CP2, we correctly prefix only 75% randomly selected of the censored areas and, at the same time, 25% of truly uncensored areas receive the censored status. In CP1

100% of the truly censored regions are wrongly informed to the model and, at the same time, an equal number of truly uncensored areas receive the censored status.

In all models, the prior distribution for the relative risk is $\theta_i|\alpha_i, \phi_i \overset{ind}{\sim} \text{Gamma}(\alpha_i, \phi_i)$. The hyperparameters α_i and ϕ_i are fixed at each region in such way that prior distributions for θ_i have $\text{Var}[\theta_i] = 3.0$ and $E[\theta_i]$ equal to the true relative risk used to generate the data sets.

For the MCMC, we run one chain of 39,500 iterations and, after convergence being reached, we discarded the first 20,000 iterations as the burn-in period. The number of iterations was suggested by the CODA functions [Plummer *et al.*, 2016]. To avoid correlation among the generated samples, we consider a lag of length 13 obtaining a posterior sample of total size 1,500. The algorithm was implemented using the software *R* 3.1.3 [R Core Team , 2015] and MCMC convergence was monitored considering diagnosis tests available on CODA package. The code can be obtained from the authors upon request. These specifications for the MCMC parameters are also assumed in the case study presented in Section 4.4.

The posterior estimates for the relative risk $\boldsymbol{\theta}$ are obtained under the square loss function. To evaluate the quality of the estimates provided by the posterior means, in the Monte Carlo simulation study, we consider the bias (BIAS) and the mean squared error (MSE). We also obtained the coverage percentage of the 95% highest posterior density (HPD) intervals for $\boldsymbol{\theta}$. Results are shown in Table 4.1. To evaluate the capacity of the proposed model in correctly identifying the censored areas, we consider the sensitivity ($Sensit. = TP/(TP + FN)$), specificity ($Specif. = TN/(TN+FP)$) and the accuracy ($Accur. = (TP+TN)/(TP+FP+TN+FN)$) rates given in Table 4.2, where TP, FP, TN and FN represent, respectively, the true positive, false positive, true negative and false negative cases. We classify the region i as censored whenever the posterior mean of π_i is higher than 0.5. The true censored regions are shown in Figure 4.4.

Tables 4.1 and 4.2 show the resulting evaluation metrics for the estimates of $\boldsymbol{\theta}$, separately in censored and non-censored regions, for datasets with censoring levels $\delta = 0.8$ and $\delta = 0.6$, respectively. The analysis in the following will take into consideration that, in the epidemiological context, the underestimation of the relative risk is not desirable since it may lead to the establishment of inappropriate healthy policies.

It can be noticed that the regular Poisson model produces undesirable relative risk estimates in censored regions by always underestimating them. Despite of this, it tends to provide good estimates for the relative risk in non-censored regions. Comparing models that consider the presence of underreporting, CP3 tends to provide the less biased estimates and the smallest MSE in the non-censored regions for almost all scenarios. It must occur in fact, since CP3 is equivalent to the regular Poisson model in these regions. However, to fit CP3 we need to perfectly know the censored and the non-censored regions which may not be a simple task. Among the other models, RCP3 is the one that better estimate $\boldsymbol{\theta}$ in non-censored regions while RCP1, MG1, MG2 and CP1 are the worst ones, tending to highly overestimated the relative risk. This overestimation of $\boldsymbol{\theta}$ also occurs in censored regions under RCP1 and MG1, for all scenarios, showing that these two models are non-competitive and pointing out that it is not a reasonable strategy to elicit vague prior distributions for $\boldsymbol{\pi}$ and $\boldsymbol{\epsilon}$ (parameters that brings information about the data quality to the model).

In censored regions, the models that, in general, provide the best estimates for $\boldsymbol{\theta}$ are RCP3 and MG3 in terms of both MSE and bias. For scenarios where $\delta = 0.8$ (Table 4.1), MG3 presents slightly better estimates, but the opposite occurs when the data quality gets worse and we have $\delta = 0.6$ (Table 4.2). In non-censored regions, RCP3 fits better than MG3 in all scenarios. For those regions, RCP2 provides smaller MSE but more biased estimates than MG2. Despite of this, in the non-censored regions, RCP2 has better performance presenting less bias and variability for the estimates independently of the scenario. One undesirable feature of these two models is their tendency to underestimate the relative risk in censored regions.

Comparing only the CP models in censored regions, the relative risks are overestimated in all scenarios under CP3 and CP2, mainly, in those scenarios with few data information. CP1 presents the same results as the regular Poisson model since they are equivalent in these regions. In non-censored regions, the performance of CP models gets worse as the information about actually censored regions also gets worse. It is worthy to mention, however, that the quality of the estimates provided by CP models are quite sensitive to the information about censored regions considered by the model (see also Figure 4.3).

The biases obtained by fitting all the ten models considered in our study tend to be closer to the bias obtained under the regular Poisson model as the proportion of censored regions and the censoring level δ increase. In scenarios where $\delta = 0.6$, even the models RCP3 and MG3, which present the best performances, underestimate the relative risk. In these cases, the underestimation of $\boldsymbol{\theta}$ under MG3 tends to be greater than that observed by fitting RCP3. The prior distributions for $\boldsymbol{\pi}$ and $\boldsymbol{\epsilon}$ play an even more important role in the quality of the posterior estimates as data information gets poorer.

The coverage percentage of the 95% HPD intervals (see Tables 4.1 and 4.2) in censored regions tends to be smaller under MG2, CP1 and the regular Poisson model. In Scenario 1 with $\delta = 0.6$, for instance, when fitting MG2, the posterior distribution puts significant probability mass in the true region of the parametric space only for 42.1% of the generated samples. For CP1 and the regular Poisson model, such percentage is even worse being around 0.006%. For RCP1, MG1 and CP3, the coverage percentage for the HPD intervals is equal to 100.0% in all scenarios, which may indicate that the posterior uncertainty about the relative risk is still large. Similar results can be observed for RCP2, RCP3 in some scenarios with $\delta = 0.8$.

The proposed model is more sensitive than specific (see Table 4.3) being more capable of precisely classifying truly censored regions as censored. Model RCP3, which consider informative prior about the censored regions, provides better classifications for both the censored and non-censored regions in all scenarios. The accuracy for such a model is above 0.73. Regarding the correct identification of censored regions, RCP2 presents the worst performance while RCP1 is the worst model to identify the non-censored ones. Although RCP1 and RCP2 produced relative risk estimates quite different in terms of bias and variability (see Table 4.1), the accuracy rate of such models are comparable in all scenarios.

Table 4.1: Evaluation of the relative risk estimates with censoring level $\delta = 0.8$.

Model	censored regions			non-censored regions		
	MSE	BIAS	%HPD	MSE	BIAS	%HPD
	Scenario 1 ($AI \leq 0.0$, 11% of censoring)					
RCP1	0.460	0.648	1.000	2.544	1.454	0.995
RCP2	0.057	-0.166	1.000	0.033	0.131	0.994
RCP3	0.113	0.224	1.000	0.020	0.049	0.969
MG1	0.633	0.768	1.000	1.007	0.951	0.969
MG2	0.353	-0.072	0.704	1.173	0.354	0.957
MG3	0.080	0.177	1.000	0.032	0.083	0.961
CP1	0.162	-0.369	0.384	0.555	0.242	0.943
CP2	1.453	0.906	0.818	0.105	0.054	0.942
CP3	1.775	1.316	1.000	0.011	0.002	0.942
Poisson	0.162	-0.369	0.382	0.011	0.002	0.943
	Scenario 2 ($AI \leq 20.0$, 23% of censoring)					
RCP1	0.716	0.774	1.000	2.806	1.504	0.995
RCP2	0.049	-0.126	1.000	0.030	0.126	0.995
RCP3	0.073	0.077	1.000	0.012	0.025	0.965
MG1	0.617	0.758	1.000	0.962	0.929	0.971
MG2	0.240	-0.111	0.728	1.159	0.405	0.949
MG3	0.047	0.071	0.998	0.016	0.051	0.966
CP1	0.130	-0.313	0.416	1.559	0.639	0.940
CP2	1.405	0.910	0.856	0.360	0.169	0.941
CP3	1.787	1.319	1.000	0.382	0.216	0.941
Poisson	0.130	-0.313	0.422	0.010	0.000	0.939
	Scenario 3 ($AI \leq 45.0$, 36% of censoring)					
RCP1	0.985	0.887	1.000	2.768	1.508	0.996
RCP2	0.041	-0.107	1.000	0.029	0.123	0.994
RCP3	0.060	0.001	0.980	0.010	0.014	0.961
MG1	0.601	0.744	1.000	0.939	0.912	0.974
MG2	0.152	-0.100	0.764	1.454	0.484	0.941
MG3	0.040	0.007	0.984	0.013	0.036	0.963
CP1	0.106	-0.272	0.447	3.697	1.360	0.947
CP2	1.781	1.065	0.904	0.789	0.326	0.945
CP3	2.742	1.587	1.000	0.009	0.000	0.941
Poisson	0.106	-0.272	0.449	0.009	0.000	0.940

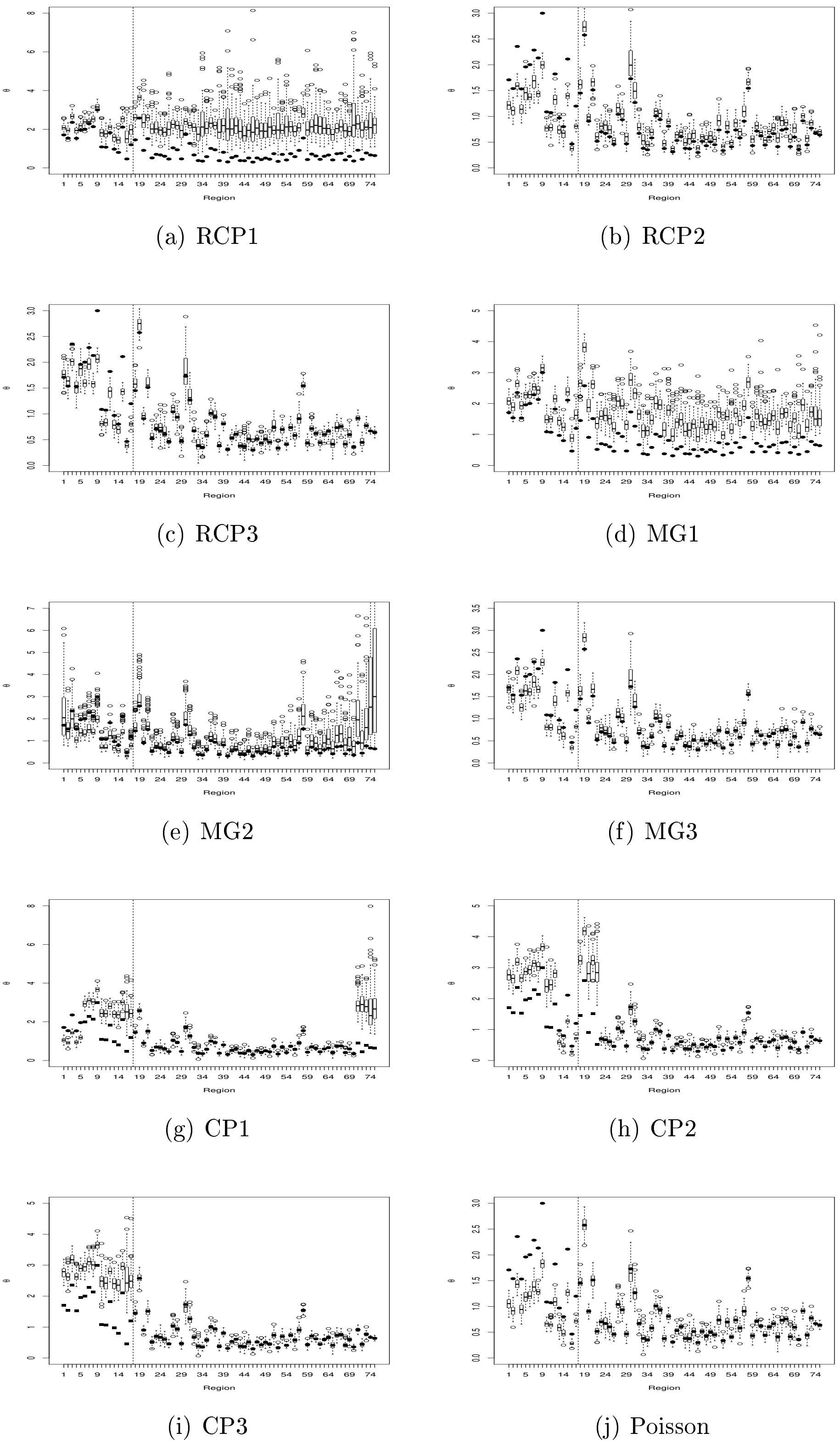
Table 4.2: Evaluation of the relative risk estimates with censoring level $\delta = 0.6$.

Model	censored regions			non-censored regions		
	MSE	BIAS	%HPD	MSE	BIAS	%HPD
	Scenario 1 ($AI \leq 0.0$, 11% of censoring)					
RCP1	0.163	0.346	1.000	2.590	1.447	0.996
RCP2	0.335	-0.551	0.821	0.034	0.134	0.995
RCP3	0.104	-0.193	0.996	0.021	0.053	0.970
MG1	0.141	0.330	1.000	0.999	0.948	0.969
MG2	0.877	-0.311	0.421	0.955	0.330	0.968
MG3	0.125	-0.284	0.998	0.032	0.083	0.965
CP1	0.595	-0.754	0.006	0.614	0.249	0.943
CP2	0.961	0.530	0.751	0.102	0.053	0.942
CP3	0.965	0.960	1.000	0.011	0.001	0.941
Poisson	0.595	-0.754	0.006	0.011	0.002	0.944
	Scenario 2 ($AI \leq 20.0$, 23% of censoring)					
RCP1	0.354	0.483	1.000	2.731	1.500	0.995
RCP2	0.296	-0.481	0.778	0.031	0.125	0.994
RCP3	0.163	-0.287	0.841	0.013	0.023	0.963
MG1	0.164	0.357	1.000	0.955	0.924	0.975
MG2	0.440	-0.420	0.428	1.091	0.381	0.930
MG3	0.146	-0.320	0.882	0.016	0.051	0.965
CP1	0.492	-0.648	0.032	1.486	0.625	0.946
CP2	0.856	0.588	0.738	0.339	0.162	0.946
CP3	1.493	1.143	1.000	0.010	-0.002	0.947
Poisson	0.492	-0.648	0.036	0.010	-0.003	0.946
	Scenario 3 ($AI \leq 45.0$, 36% of censoring)					
RCP1	0.647	0.641	1.000	3.063	1.575	0.995
RCP2	0.238	-0.409	0.800	0.028	0.122	0.995
RCP3	0.166	-0.308	0.684	0.010	0.014	0.962
MG1	0.191	0.388	1.000	0.981	0.918	0.972
MG2	0.135	-0.249	0.513	3.438	0.753	0.945
MG3	0.074	-0.243	0.681	0.033	0.055	0.963
CP1	0.391	-0.555	0.060	3.611	1.347	0.949
CP2	1.161	0.778	0.771	0.790	0.326	0.943
CP3	1.979	1.305	1.000	0.009	0.000	0.942
Poisson	0.391	-0.555	0.061	0.009	0.000	0.940

Table 4.3: Evaluating the inference about the censored areas

Model	$\delta = 0.8$			$\delta = 0.6$		
	Sensit.	Specif.	Accur.	Sensit.	Specif.	Accur.
	Scenario 1 ($AI \leq 0.0$, 11% of censoring)					
RCP1	0.728	0.235	0.482	0.752	0.235	0.494
RCP2	0.501	0.496	0.499	0.507	0.495	0.501
RCP3	0.943	0.557	0.750	0.943	0.557	0.750
	Scenario 2 ($AI \leq 20.0$, 23% of censoring)					
RCP1	0.737	0.231	0.484	0.755	0.232	0.493
RCP2	0.507	0.494	0.500	0.504	0.496	0.500
RCP3	0.880	0.616	0.748	0.881	0.616	0.748
	Scenario 3 ($AI \leq 45.0$, 36% of censoring)					
RCP1	0.746	0.230	0.488	0.761	0.227	0.494
RCP2	0.504	0.496	0.500	0.504	0.495	0.500
RCP3	0.799	0.674	0.736	0.798	0.623	0.735

Figure 4.3 present the box-plots for the relative risk estimates obtained under all models for Scenario 2 with $\delta = 0.6$. Similarly to what is observed whenever a regular Poisson model is fitted (Figure 4.3 (j)), if the CP model does not assume as censored the regions where information are in fact underreported, the relative risks are underestimated in that regions. If the CP models are assumed, the relative risk tends to be overestimated in all regions fixed in the model as being censored. This issue is even more evident in regions that are wrongly considered as censored in the model. In these cases, there is also a greater variability in the estimates. Models RCP2, RCP3, MG3 and CP3 provide quite similar estimates for the relative risk in non-censored regions. If models RCP1 and MG1 are fitted, the relative risks is highly overestimated in all non-censored regions and great variability on the estimates is also observed. These results can be possibly explained by the estimates obtained for the censoring probability π and the each event non-recording probability ϵ_i under such models. The estimates for π are above 0.70 under RCP1 and the estimates for ϵ_i are above 0.15 under MG1, for almost all regions. In turn, the posterior estimates of π obtained under RCP2 model are around 0.5 for all regions. Because of this, the relative risk estimates provided by RCP2 are comparable to those observed by fitting the regular Poisson model, mainly in censored regions. In general, for the scenario shown in Figure 4.3, the estimates provided by the RCP3 and MG3 model are the best ones presenting the smallest MSE (Table 4.2).



(a) RCP1

(b) RCP2

(c) RCP3

(d) MG1

(e) MG2

(f) MG3

(g) CP1

(h) CP2

(i) CP3

(j) Poisson

Figure 4.3: Box-plots of $\boldsymbol{\theta}$ posterior means for Scenario 2 and $\delta = 0.6$. The true values are represented by black squares for regions considered as censored in the CP models and by black circles for the other regions. The truly censored areas are in the left side of the vertical dashed line. Note that graphs have different scales for the y-axis.

4.4 Case Study: Mapping the ENM rate in Minas Gerais State, Brazil

In this section we consider the proposed random censoring Poisson (RCPM) and the censored Poisson (CPM) models to map the relative risk of early neonatal mortality (ENM) in Minas Gerais State (MG), Brazil. Although there was an improvment in the Brazilian socio-economic conditions in the last decade and some investment in improving the information gathering, data in Minas Gerais still suffer of under-recording. To show that we consider data of two different periods, 1999-2000 and 2012-2014, and described in Section 4.1. Because we do not have any prior information about the proportion of underreporting experienced in each region, which is needed in the construction of informative prior distribution for $\boldsymbol{\epsilon}$, we do not fit the model proposed by Moreno and Girón [1998] (MGM). As noted from the simulation studies (Section 4.3), eliciting informative prior distribution for $\boldsymbol{\epsilon}$ is a key point to obtain good estimates for the relative risks under MGM.

Let Y_i and E_i be, respectively, the observed and the expected counts of ENM at region $i = 1, ..., 75$, where $E_i = B_i \times T$, B_i is the number of live births in area i and T denotes the death rate for all Minas Gerais State. As in Section 4.2.1, assume that in some regions the observed counts Y_i may be right censored (underreported) with respect to the true but unobserved count N_i which, given θ_i, has a Poisson distribution with parameter $E_i\theta_i$. Despite the counts may be underreported, the fitted models assumes that the available E_i is free of underreporting because the possible bias is negligible due to the small number of missing deaths relatively to the large total number for the whole state. Also, based on information provided by experts we believe that the ENM data set is similar to the data considered in Scenario 2 of the simulation study (Section 4.3) in relation to the proportion of censored areas.

4.4.1 On the prior specification for π_i and θ_i

To build an informative prior distribution for the censoring probabilities we consider the logit function in (4.3) with the adequacy index (AI) França *et al.* [2006] as covariate. As mentioned in Section 4.3, the AI ranges from -83.74 to 100.0 assuming negative values for eight regions of Minas Gerais, all of then located in the North and Northeast part of the state. The smallest the AI, the worst is the data quality in the region. We elicit the prior distributions of β_0 and β_1 taking into consideration that the highest values for the censoring probabilities are associated to the regions with the worse adequacy index.

Attempting to make this elicitation a simpler task, we conveniently ordered the observed values of AI from the smallest to the highest and let $\text{logit}(\pi_i) = \beta_0 - \beta_1 \text{AI}_i^1$, where AI^1 represents the standardized adequacy index and $i = 1, ..., 75$. Because of this construction, to ensure the expected behavior of $\boldsymbol{\pi}$ among the regions, the prior distribution β_1 should put probability mass in positive values. To appropriately describe the expert prior knowledge about the censored regions in Minas Gerais, we elicit: $\beta_0 \sim LN(-2.06; 0.95)$ and $\beta_1 \sim LN(1.52; 0.60)$, where LN is the Log-normal distribution. Under such prior distributions, it follows that $E[\beta_0] = 0.20$, $Var[\beta_0] = 0.06$, $E[\beta_1] = 5.50$, $Var[\beta_1] = 13.11$. This prior elicitation reveals that, in average, the underreporting probability in the eight regions with worse data quality is greater than 0.99 while for the twelve regions with highest values of AI it is smaller than 0.01. We named the model with such log-normal prior distributions on β_0 and β_1 by RCPM3.

To provide a sensitivity analysis on the reporting process, similarly to what is considered in Section 4.3, we consider here the case in which $\pi_i \overset{iid}{\sim} \mathcal{U}(0,1)$ (named RCPM1) and also that one in which $(\beta_0, \beta_1) \sim \mathcal{N}_2(\mathbf{0}, 100\boldsymbol{I}_2)$ (named RCPM2).

For fitting the CPM to our data set, we consider the three criteria based on the adequacy index (AI) given in Section 4.3, that is, we consider as censored the regions for which the $AI \leq 0.0$ (named CPM1), $AI \leq 20.0$ (named CPM2) and $AI \leq 45.0$ (named CPM3) thus establishing that the proportion of censored regions are, respectively, 11%, 23% and 36% for Scenarios 1, 2 and 3. The censored regions induced by these criteria are presented in Figure 4.4 (row 1).

In epidemiological studies, the use of informative prior distributions for the relative risks based on information provided by experts is important and it has been encouraged by many authors as Bernardinelli, Clayton and Montomoli [1995]. To build the prior specification for the relative risks in our study we obtained information from experts in the study of the early neonatal mortality in Minas Gerais and we summarized their knowledge through the mean relative risks over the $n = 75$ regions by doing

$$E[\theta_i] \ = \begin{cases} 5.0, & \text{if } AI_i \leq 0.0, \ i = 1, ..., 8, \quad \text{(8 regions)} \\ 3.5, & \text{if } 0.0 < AI_i \leq 20.0, \ i = 9, ..., 17, \quad \text{(9 regions)} \\ 2.5, & \text{if } 20.0 < AI_i \leq 45.0, \ i = 18, ..., 27, \quad \text{(10 regions)} \\ 1.5, & \text{if } 45.0 < AI_i \leq 60.0, \ i = 28, ..., 40, \quad \text{(13 regions)} \\ 1.0, & \text{if } AI_i > 60.0, \ i = 41, ..., 75. \quad \text{(35 regions)} \end{cases}$$

We then elicit Gamma distributions in which the hyperparameters α_i and ϕ_i are chosen so that the prior means are these given by the experts and the prior variance $Var[\theta_i]$ is common for all regions and equal to 3.0. By eliciting this prior, we assume that the expected relative risk is below 3 in the majority of the regions and it is above 3 in 17 regions.

4.4.2 The posterior results

Figures 4.4 and 4.5 displays the posterior means and the 95% highest posterior density (HPD) intervals for the relative risks of early neonatal mortality in Minas Gerais, under the CP models in the periods 1999-2001 and 2012-2014, respectively. Compared to the standardized mortality ratio (SMR) and the estimates provided by the CAR model (see Figure 4.1), the relative risks estimates for the period 1999-2001 in non-censored regions remains essentially the same. The same occurs for data from period 2012-2014 if we compare Figure 4.5 and Figure 1 of the supplementary material, in which is shown the SMR estimates for this period. However, when underreporting is considered, the relative risk estimates in North and Northeast regions of the Minas Gerais State are much higher under all CP models. This result is closer to what is expected by the epidemiologists. A critical point by using the CP model is that the risk estimates are highly affected by the censoring criterion established *a priori*. As an example, we can highlight what occurs with the region having the highest latitude in the State, i.e., the region further North in the map. That region is not censored in CPM1 but censored in CPM2 and CPM3. When models CPM2 and CPM3 are fitted, the relative risk estimates belong to the interval $[2.5, 4.0)$ while such estimate is smaller than 0.70 under the model CPM1. For such region the 95% HPD interval also reveals a great uncertainty about the relative risk under CPM2 and CPM3, in which it is a censored region, but great certainty if CPM1 model is fitted, where it is a non-censored one. Usually, the range of the HPD interval is smaller in non-censored areas.

The posterior means and the 95% HPD intervals for the relative risk obtained by fitting the three different configurations for the proposed RCP models are exhibited in Figures 4.6 and 4.7 for periods 1999-2001 and 2012-2014, respectively. As can be noticed, the posterior estimates of the relative risk are influenced by the prior distribution of the censoring probabilities $\boldsymbol{\pi}$. However, one advantage of using the proposed model is that we quantify the posterior uncertainty about $\boldsymbol{\pi}$. Thus we can decide about the censored regions in a probabilistic way.

If the prior uniform distribution describes the uncertainty about $\boldsymbol{\pi}$, RCPM1, then *a posteriori* we have that the probability of censoring is, in average, above 0.67 in all regions (see Figure 4.6), similarly to what is observed for this model in the simulation studies. As a consequence, we obtained that the posterior means for the relative risk in almost all areas is greater than the standardized mortality ratio and the estimates provided by the CAR model (see Figure 4.1). In this case the HPD intervals also reveal a great posterior uncertainty about the relative risks. By fitting RCPM2, the posterior probability of censoring is, in average, between 0.48 and 0.55 for all regions. Under this model the estimates for the relative risks are similar to that obtained under the standardized mortality ratio. In both cases, RCPM1 and RCPM2, the posterior estimates for the relative risk is far from what is expected by the experts. RCPM1 overestimated the relative risk in Midwestern, Southeastern and Southern regions of the state, whereas RCPM2 underestimated the risks in the North and Northern regions. RCPM3 produced estimates that are more compatible to what is expected (Figure 4.6). This model indicates that areas having posterior probability of censoring, in average, above 0.95 are situated in the north, northeast and northwest of the state. In fact, such areas are usually pointed out by the experts as being the region with the worst data quality. The posterior relative risk estimates for that regions tend to be higher than for regions with low probability of censoring, being above 4.0 only for regions in the Northeast region of Minas Gerais. Despite the great uncertainty about the relative risk in regions with high posterior censoring probability, the HPD intervals obtained under RCPM3 disclose less posterior variability than obtained under the CP models in almost all regions (see Figure 4.4).

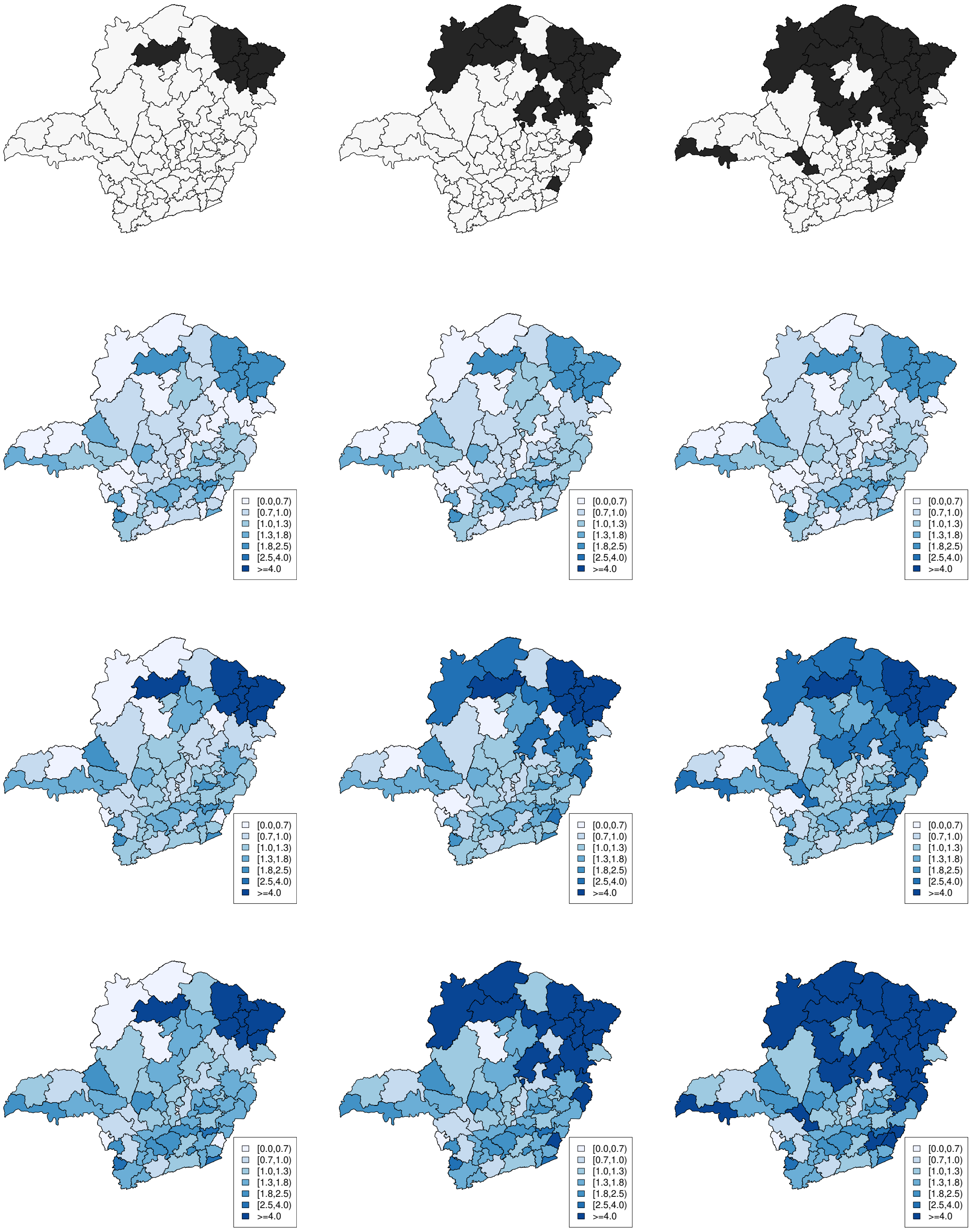


Figure 4.4: Case study results under CPM1 (column 1), CPM2 (column 2) and CPM3 (column 3). Censored regions are highlighted in row 1. Relative risk estimates for the early neonatal mortality in Minas Gerais 1999-2001: the lower limit of the 95% HPD interval (row 2), the posterior mean (row 3) and the upper limit of the 95% HPD interval (row 4).

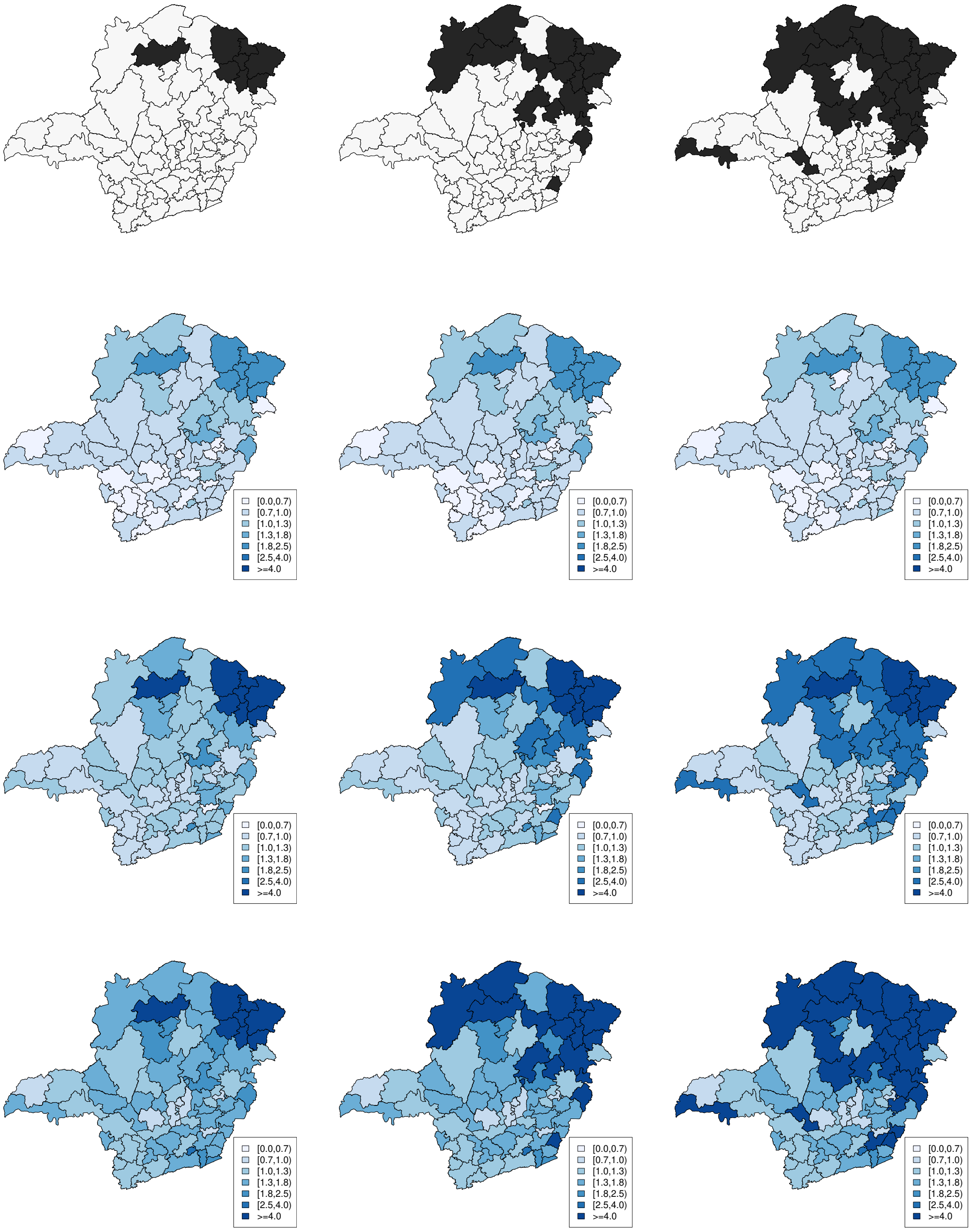


Figure 4.5: Case study results under CPM1 (column 1), CPM2 (column 2) and CPM3 (column 3). Censored regions are highlighted in row 1. Relative risk estimates for the early neonatal mortality in Minas Gerais 2012-2014: the lower limit of the 95% HPD interval (row 2), the posterior mean (row 3) and the upper limit of the 95% HPD interval (row 4).

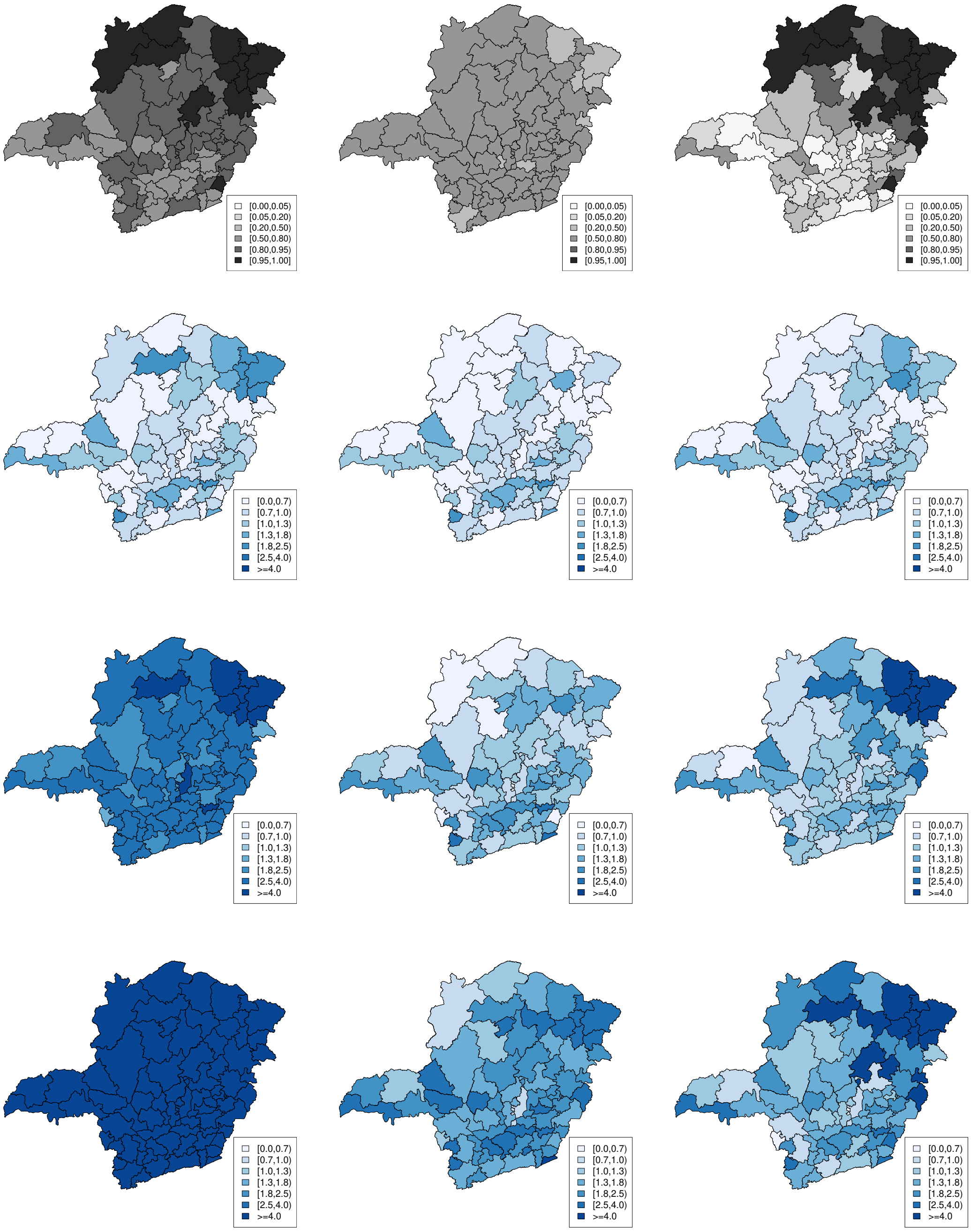


Figure 4.6: Case study results under RCPM1 (column 1), RCPM2 (column 2) and RCPM3 (column 3). Posterior mean of the censoring probabilities are presented in row 1. Relative risk estimates for the early neonatal mortality in Minas Gerais 1999-2001: the lower limit of the 95% HPD interval (row 2), the posterior mean (row 3) and the upper limit of the 95% HPD interval (row 4).

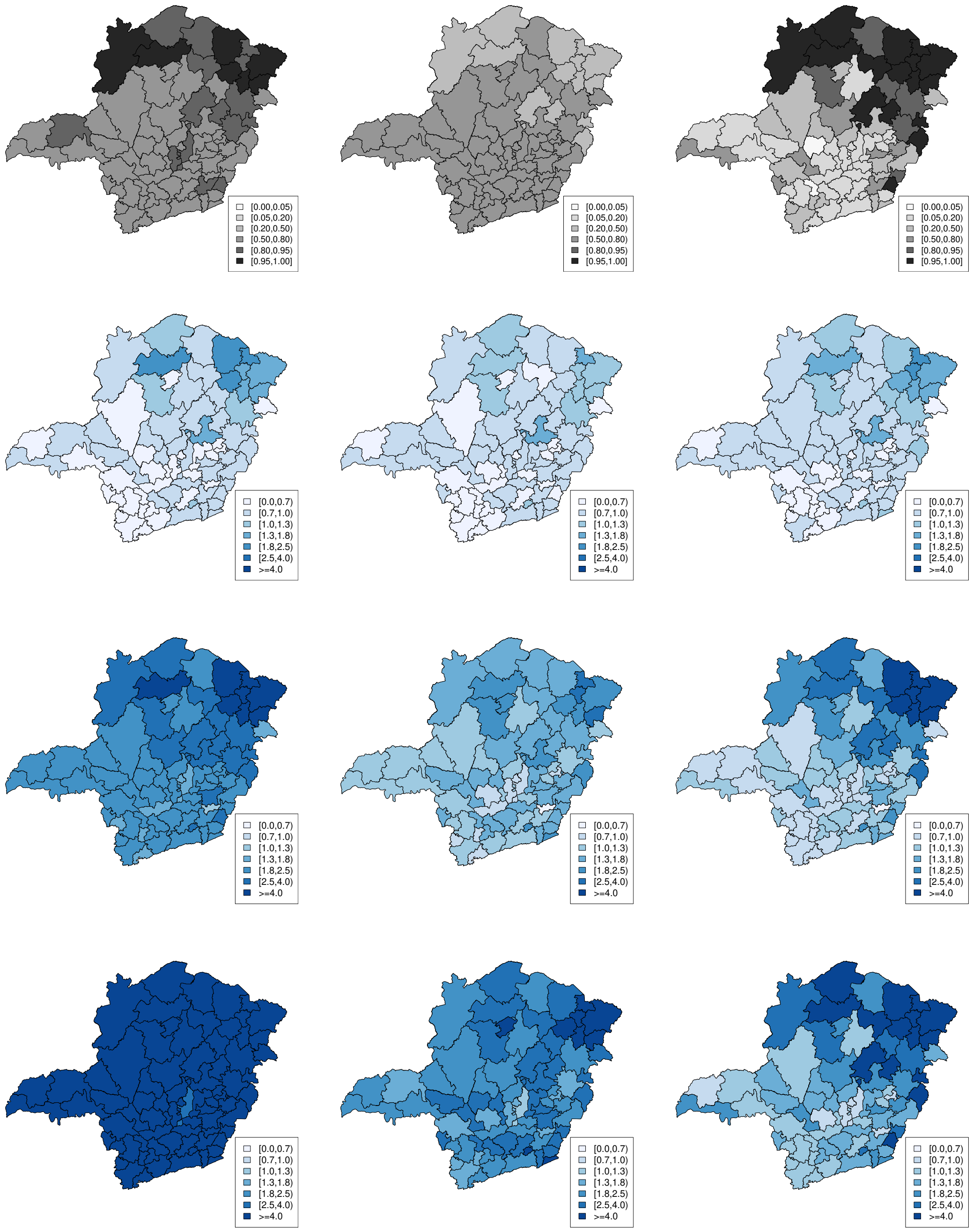


Figure 4.7: Case study results under RCPM1 (column 1), RCPM2 (column 2) and RCPM3 (column 3). Posterior mean of the censoring probabilities are presented in row 1. Relative risk estimates for the early neonatal mortality in Minas Gerais 2012-2014: the lower limit of the 95% HPD interval (row 2), the posterior mean (row 3) and the upper limit of the 95% HPD interval (row 4).

The relative risks estimates provided by RCPM (Figure 4.7) and CPM (Figure 4.5) based on data collected in the period 2012-2014 present the same pattern as the ones obtained for 1999-2001 data set. However, we observe an increase in the relative risk in some regions, typically in the North and Northeast area. Since the HDI indicates an improvement in the Brazilian socio-economic conditions in the last decade (see HDI maps for years 2000 and 2010 in Figure 4.1 and Figure 1 of the supplementary material, respectively), this increase in the estimates possibly indicates an improvement in data recording, and not an increase in the neonatal mortality rate. For some regions there is a change in the estimate for the underreporting probability π_i under the proposed RCPM.

4.5 Final comments and main conclusions

The precise mapping of risks related to vital statistics is an important tool to guide the definition of adequate health policies and to reduce the occurrence of events such as the infant mortality. Our main motivation was to obtain better estimations for the relative risk associated to early neonatal mortality (ENM) in Minas Gerais State, Brazil, using data registered in public hospitals between 1999 and 2001 and also between 2012 and 2014. The occurrence of underreporting in such data sets is quite likely [Campos *et al.*, 2007] and if it is not accounted for, estimates will be biased (underestimated).

In this work, we introduced the random censored Poisson model (RCPM) that, by jointly modeling the uncertainty about the counts and the data reporting process, accounts for the underreported information more appropriately. A challenge in this approach is the elicitation of the prior distribution for the censoring probability which discloses the prior knowledge about regions in which the mortality counts are underreported. We built some different prior distributions for that. We run a simulation study evaluating the effect of such different prior specifications in the relative risk estimates. We also compared the propose model with the so-called censored Poisson model (CPM) in different scenarios. The CPM arises as a particular case of our RCPM under degenerate prior distribution for the reporting process. We concluded that the proposed model tends to produce less biased estimates for the relative risk than the CPM, mainly in scenarios with poorer data information. More importantly, the effort of building informative prior distributions for the underreporting probability is compensated by a better estimation of the relative risk as well as it permits making good posterior inference about the data reporting process. The use of flat prior distributions for the underreporting probability must be avoided since it leads to truly poor estimates for the relative risk. For the case study, in both periods 1999-2001 and 2012-2014, we conclude that RCPM3 provides the best estimates for the relative risks because the ENM data set is very similar to the data considered in Scenario 3 of the simulation study and RCPM3 provided the best estimates in such scenario.

Moreno and Girón's model (MGM) is also an appropriate and competitive tool to account for underreported data. The main difference between the proposed RCPM ad MGM is the type of prior information required in their construction. The parameter $\boldsymbol{\theta}$ represents the relative risks in both models. Therefore, the major issue is related to parameters $\boldsymbol{\pi}$ and $\boldsymbol{\epsilon}$. To elicit a prior distribution for the censoring probability $\boldsymbol{\pi}$ under RCPM we only need the information about regions where data quality is not good or reliable. To build the prior for $\boldsymbol{\epsilon}$ under MGM, information about the proportion of underreporting in each area should be available. Therefore, the proposed RCPM is an important tool for cases in which one have information about the regions where the occurrence of underreporting is more likely but no information about the amount of censoring in each region.

The censoring mechanisms used in the proposed RCPM is applicable to other case studies and several other specifications for the logit function given in expression (4.3) can be thought. For example, the censoring probabilities can be modeled using other available socio-economic index, such as the human developing index. Moreover, the RCPM can easily be extended to account for both over-reporting or even more general misclassification on the data. Interesting topics for future research include the use of spatial random effects in the linear predictor and the assumption of some cluster structure between the regions in the map.

Acknowledgements

The authors would like to express their gratitude to the editors and two anonymous referees whose comments and suggestions have contributed to the improvement of the paper. The authors also thank CNPq, CAPES and FAPEMIG, three Brazilian funding agencies, for their partial support.

References

Alfonso, J., Lovseth, E., Samant, Y., and Jolm, J. (2015). Work-related skin diseases in Norway may be underreported: data from 2000 to 2013. *Contact Dermatitis*, **72**(6):409-412.

Assunção, R.M., Potter, J.E. and Cavenaghi, S.M. (2002). A Bayesian space varying parameter model applied to estimating fertility schedules. *Statistics in Medicine*, 21(14):2057–2075.

Assunção, R.M., Schmertmann, C. P., Potter, J.E. and Cavenaghi, S.M. (2005). Empirical Bayes estimation of demographic schedules for small areas. *Demography*, 42(3):537–558.

Bailey, T.C., Carvalho, M., Lapa, T., Souza, W. and Brewer, M. (2005). Modeling of underdetection of cases in disease surveillance. *Annals of Epidemiology*, 15(5):335–343.

Bernardinelli, L., Clayton, D., and Montomoli, C. (1995). Bayesian estimates of disease maps: How important are priors? *Statistics in Medicine*, 14(21–22):2411–2431.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of Royal Statistical Society, Series B*, 36(2):192–236.

Besag, J., York, J. and Molliè, A. (1991). Bayesian image restoration, with two applications in spatial statistics.*Annals of the Institute of Statistical Mathematics*, 43(1):1–59.

Brass, W., Coale, A.J., Demeny, P., Heisel, D.F., Lorimer, F., Romaniuk, A., and Van de Walle, E. (1968). Demography of Tropical Africa. *Princeton Univ. Press*, New Jersey,US.

Brass, W. (1996). Demographic Data Analysis in Less Developed Countries: 1946–1996. *Population Studies*, 50(3):451–467.

Campbell, L., Hills, S., Fischer, M., Jacobson, J., Hoke, C., Hombach, J., Marfin, A., Solomon, T., Tsai, T., Tsu, V. and Ginsburg, A. (2011). Estimated global incidence of Japanese encephalitis: a systematic review. *Bulletin of the World Health Organization*, 89:766–774E. DOI: 10.2471/BLT.10.085233.

Campos, D., Loschi, R. H., and França, E. (2007). Early neonatal hospital mortality in Minas Gerais: Association with healthcare variables and the issue of underreporting (available in portuguese). *Revista Brasileira de Epidemiologia*, **10**(2), 223–238.

Caudill, B. S. and Mixon Jr., F. G. (1995). Modeling Household Fertility Decisions: Estimation and Testing of Censored Regression Models for Count Data. *Empirical Economics*, **20**(2), 183–196.

Clayton, D. and Kaldor J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671–681.

Chib, S. (1992). Bayes inference in the Tobit Censored Regression model. *Journal of Econometrics*, 51(1–2):79–99.

Cordeiro, G.M. and Cribari-Neto, F. (2014). An introduction to Bartlett correction and bias reduction *Springer*, New York.

Dvorzak, M. and Wagner, H. (2016). Sparse Bayesian modelling of underreported count data. *Statistical Modelling*, **16**(1), 24–46.

França, E., Abreu, D., Campos, D. and Rausch, M. C. (2006). Avaliação da qualidade da informação sobre a mortalidade infantil em Minas Gerais: Utilização de uma metodologia simplificada (in portuguese). *Revista Médica de Minas Gerais*, **16**(1 supl 2), S28–S35.

Gould, J., Chavez, G., Marks, A. and Liu, H. (2002). Incomplete birth certificates: A risk marker for infant mortality. *American Journal of Public Health*, 92(1):79–81.

Heligman, L., Finch, G. and Kramer R. (1978). Measurement of Infant Mortality in Less Developed Countries. *Department of Commerce, Bureau of the Census.*

Hill, K., Choi, Y. and Timaeus, I.M. (2005). Unconventional approaches to mortality estimation. *Demographic Research*, 13(article 12), 281–300.

Moreno, E. and Girón J. (1998). Estimating with incomplete count data: A Bayesian approach. *Journal of Statistical Planning and Inference*, 66(1):147–159.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC, *R News*, 6(1):7–11.

Powers, S., Gerlach, R. and Stamey, J. (2010). Bayesian variable selection for Poisson regression with underreported responses. *Computational Statistics & Data Analysis*, 54(12):3289–3299.

R Core Team (2015). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, (2015). Available at https://www.R-project.org/

Schramm, J.M.A. and Szwarcwald, C.L. (2000). Sistema hospitalar como fonte de informações para estimar a mortalidade neonatal e a natimortalidade (in portuguese). *Revista de Saúde Pública*, 34(3):272–279.

Simões, C.C. (1999). Estimativas da mortalidade infantil por microrregiões e municípios (in portuguese). *Ministério da Saúde, Brasília.*

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Terza, J.P. (1985). A Tobit-type estimator for the Censored Poisson regression model. *Economics Letters*, 18(4):361–365.

Tibbetts, S.G. and Hemmens, C. (2010). Criminological Theory, A Text/Reader. *Sage Publication*, London, UK.

Viswanathan, K., Becker, S., Hansen, P., Kumar, D., Kumar, B., Niayesh, H., Peters, D. and Burnham, G. (2010). Infant and under-five mortality in Afghanistan: current estimates and limitations. *Bulletin of the World Health Organization*, 88:576–583.

Xu, Y., Zhang, W., Yang, R.Zou, B., and Zhao, Z. (2014). Infant mortality and life expectancy in China. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 20:379–385.

Whittemore, A.S. and Gong, G. (1991). Poisson Regression with Misclassified Counts: Application to Cervical Cancer Mortality Rates. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 40(1):81–93.

Winkelmann, R. (1996). Markov Chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics*, 21:575–587.

World Health Organization (WHO). (2006). Neonatal and perinatal mortality: Country, regional and global estimates. *World Health Organization (WHO) Library Cataloguing-in-Publication Data.*

Supplementary Material

S.1 More on simulations

We present more results obtained from the simulated studies in Section Section 4.3 and also other results for the case study using data from 2012 to 2014 that do not appear there. The analysis presented here are ancillary for the understanding of the main paper. In Tables 4.4 and 4.5 we present the results for two other scenarios that complements the simulation studies presented in Section 4.3 of the main paper. In Scenario 0, there is no censored region. In Scenario 5, the regions in which the $AI \leq 78.17$ (57 regions) were defined as being censored, where AI denotes the adequacy index discussed in Section 3 of the main paper. Such criteria establish that the proportion of censored regions are, respectively, 0% and 75%. In Table 4.4 the censoring level is $\delta = 0.80$ whereas in Table 4.5 we have $\delta = 0.60$.

For Scenario 0, since there is no truly censored region, results are the same in both tables and the regular Poisson model fits the data very well. Among the other models, RCP3 is the one with less biased estimates. RCP2 and MG3 also works well in this scenario. For Scenario 5 in Table 4.4 it can be noticed that RCP2, RPC3, MG2 and MG3 tends to subestimate the relative risk. This is an expected result because such models tend to correct for underreporting in a smaller number of regions than the number of truly censored regions, thus approximating to what is obtained with the regular Poisson model.

The percentage of coverage for the 95% HPD intervals become lower for RCP3 and MG3 when compared to scenarios in Table 4.1 of the main paper. For Scenario 5 in Table 4.5 we observe the same behavior for the evaluation metrics among the models as in Table 4.4 but, in censored regions, the estimates get worse. It occurs because, for $\delta = 0.6$ (Table 4.5), we have less data information (the greater the δ, the worse the estimates).

Table 4.4: Evaluation of the relative risk estimates with censoring level $\delta = 0.8$.

Model	censored regions			non-censored regions		
	MSE	BIAS	%HPD	MSE	BIAS	%HPD
	Scenario 0 (0% of censoring)					
RCP1	-	-	-	2.351	1.394	0.995
RCP2	-	-	-	0.039	0.140	0.994
RCP3	-	-	-	0.068	0.111	0.972
MG1	-	-	-	1.047	0.975	0.967
MG2	-	-	-	1.143	0.337	0.961
MG3	-	-	-	0.075	0.139	0.965
CP1	-	-	-	5.214	2.171	0.950
CP2	-	-	-	1.261	0.630	0.942
CP3	-	-	-	0.014	0.002	0.943
Poisson	-	-	-	0.014	0.002	0.944
	Scenario 5 ($AI \leq 78.17$, 75% of censoring)					
RCP1	1.734	1.144	1.000	3.318	1.625	0.990
RCP2	0.027	-0.057	0.999	0.023	0.114	0.991
RCP3	0.039	-0.058	0.885	0.006	0.005	0.953
MG1	0.574	0.718	1.000	0.949	0.920	0.963
MG2	0.099	-0.021	0.842	3.726	0.959	0.910
MG3	0.027	-0.040	0.933	0.008	0.022	0.959
CP1	0.063	-0.191	2.349	6.367	1.360	0.929
CP2	2.956	1.367	0.933	0.007	0.003	0.933
CP3	4.250	1.918	0.999	0.007	0.003	0.938
Poisson	0.070	-0.203	0.532	0.006	0.010	0.944

Table 4.5: Evaluation of the relative risk estimates with censoring level $\delta = 0.6$.

Model	censored regions			non-censored regions		
	MSE	BIAS	%HPD	MSE	BIAS	%HPD
	Scenario 0 (0% of censoring)					
RCP1	-	-	-	2.351	1.394	0.995
RCP2	-	-	-	0.039	0.140	0.994
RCP3	-	-	-	0.068	0.111	0.972
MG1	-	-	-	1.047	0.975	0.967
MG2	-	-	-	1.143	0.337	0.961
MG3	-	-	-	0.075	0.139	0.965
CP1	-	-	-	5.214	2.171	0.950
CP2	-	-	-	1.261	0.630	0.942
CP3	-	-	-	0.014	0.002	0.943
Poisson	-	-	-	0.014	0.002	0.944
	Scenario 5 ($AI \leq 78.17$, 75% of censoring)					
RCP1	1.275	0.915	1.000	3.422	1.654	0.988
RCP2	0.132	-0.269	0.900	0.023	0.114	0.994
RCP3	0.117	-0.269	0.458	0.006	0.006	0.952
MG1	0.225	0.420	1.000	0.982	0.926	0.964
MG2	0.176	-0.194	0.633	3.074	0.869	0.879
MG3	0.103	-0.267	0.486	0.008	0.022	0.958
CP1	0.226	-0.391	0.110	6.365	2.365	0.940
CP2	2.215	1.100	0.795	0.007	0.003	0.934
CP3	3.163	1.614	1.000	0.007	0.003	0.936
Poisson	0.226	-0.391	0.107	0.007	0.004	0.937

Note from Table 4.6 that in Scenario 0 we are only able to calculate the specificity for RCPM since there is no truly censored region. For this scenario, RCP3 presents specificity that is much lower than that observed in the other scenarios. This is quite natural since RCP3 induces censoring in some areas that are truly non-censored. The performance of RCP3 in Scenario 5 is worse than that observed in other scenarios. This result is also expected because there are much more truly censored regions than the model tends to indicate as being censored.

Table 4.6: Evaluating the inference about the censored areas

Model	$\delta = 0.8$			$\delta = 0.6$		
	Sensit.	Specif.	Accur.	Sensit.	Specif.	Accur.
	Scenario 0 (0% of censoring)					
RCP1	-	0.240	-	-	0.240	-
RCP2	-	0.495	-	-	0.495	-
RCP3	-	0.504	-	-	0.504	-
	Scenario 5 ($AI \leq 78.17$, 75% of censoring)					
RCP1	0.746	0.230	0.488	0.758	0.203	0.481
RCP2	0.501	0.498	0.499	0.504	0.493	0.498
RCP3	0.583	0.777	0.680	0.583	0.777	0.670

Finally, following the suggestion of a anonymous referee, we performed a simulation study considering that δ (the censoring level in the censored regions) is not the same for all censored regions. Instead, it can vary among them. In Table 4.7 we present the results under this approach for Scenario 2, in which regions with $AI \leq 20.0$ are considered as being censored (17 regions, 23%). We specify that the censoring level δ vary and equally spaced steps between 0.5 to 0.9 according to their adequacy index (AI). This is done in such a way that $\delta = 0.5$ is associated to the region with the lowest AI and $\delta = 0.9$ is associated to that one with the greatest AI. For non-censored regions, results for the evaluation metrics are quite similar to those observed in Scenario 2 when $\delta = 0.8$ (Table 4.1 of the main paper) and $\delta = 0.6$ (Table 4.2 of the main paper). In the censored regions, the regular Poisson model provided worse results than that observed for $\delta = 0.8$ but better results than that obtained for $\delta = 0.6$. The same behavior occur for the bias obtained under the three configurations of both RCPM and MGM. Note that for RCP3 and MG3 the relative risks $\boldsymbol{\theta}$ tends to be underestimated as in Table 4.2 of the main paper.

Table 4.7: Evaluation of the relative risk estimates with censoring level δ ranging from 0.5 to 0.9.

	Scenario 2 ($AI \leq 20.0$, 23% of censoring)					
RCP1	0.567	0.619	1.000	2.757	1.506	0.996
RCP2	0.215	-0.346	0.911	0.030	0.125	0.994
RCP3	0.077	-0.152	0.992	0.012	0.024	0.964
MG1	0.385	0.534	0.999	0.979	0.932	0.970
MG2	0.364	-0.259	0.623	1.081	0.424	0.947
MG3	0.073	-0.168	0.998	0.016	0.051	0.966
CP1	0.368	-0.512	0.266	1.496	0.628	0.934
CP2	0.922	0.726	0.918	0.364	0.170	0.939
CP3	1.956	1.298	0.999	0.007	0.000	0.941
Poisson	0.368	-0.512	0.268	0.001	0.000	0.941

S.2 More on Case Study

In this section we present in Figure 4.8 a comparison between the estimates for the relative risks $\boldsymbol{\theta}$ of early neonatal mortality (ENM) in the $n = 75$ regions of Minas Gerais state (MG) for data sets from 1999-2001 and 2012-2014 using the standartized mortality ratio (SMR). Also, the human development index (HDI) for those regions in 2010 is displayed. When comparing the estimates obtained in both periods, it can be noticed an increase in the SMR estimates in several regions of MG, mainly in the North, Northeast and West regions. Such a result may indicate an improvement on the quality of data registration or a real increase on the risk for these regions. We believe that the former is more plausible than the later, since the HDI map indicates that the socioeconomic condition of almost all regions also had an improvement in the decade considered. However, the estimates for the RRs in Northern regions still remain below that risks expected by the epidemiologists. Because of this, even today is quite important to account for underreporting when estimating the relative risks of ENM in Minas Gerais state.

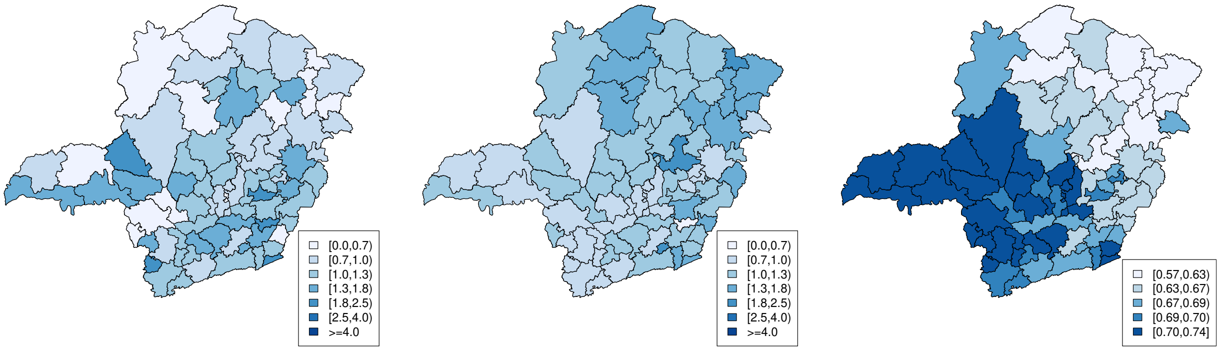


Figure 4.8: Early neonatal mortality RR estimates using SMR for data from 1999-2001 (left) and data from 2012-2014 (middle); and the HDI of MG in 2010 (right).

Chapter 5

Final Discussion and Future Work Directions

Mortality data are important in the measurement of population health and disease incidence. As such, providing reliable estimates for mortality rates helps in planning of public health policies and in the evaluation of the health system. It is also helpful to guide prevention, control and intervention by the responsible authorities, as well as to conduct the identification of regions that need special attention regarding the event under study.

Governments, international organizations and academic researchers use civil registration systems, population censuses, household surveys and demographic surveillance systems as the main data sources when assessing mortality status at the national and subnational levels [Silva, 2013]. The data are generally available as counts of deaths and population exposure to risk. The civil registration systems are universally recognized as the most ideal data source for the regular derivation of vital statistics, including mortality rates, because it entails the continuous, permanent, compulsory and universal recording of the occurrence and characteristics of vital events [United Nations, 2014]. Nevertheless, many countries lack basic vital registration systems that are critical for the accurate measurement of mortality rates.

When civil registration is lacking or incomplete, there is considerable uncertainty and limitations associated with mortality measurement. That is a special problem when analyzing data from small areas, specially in underdeveloped countries. As the demand for timely and accurate mortality estimates increases, there is a need to develop methods which incorporate both more flexible statistical models and traditional demographic procedures in order to overcome known data issues. The importance of proposing and developing new modeling frameworks to adequately incorporate and correct biases caused by data issues is evidenced by the negative impacts that such phenomena may cause in several systems and services essential for society.

This dissertation approached the problem of modeling count data with a defective reporting mechanism. More specifically, we consider the estimation of rates based on underreported counts as well as the problem of smoothing mortality schedules in subnational small populations where observed counts tend to be sparse and erratic. For the appropriate treatment of underreported count data, two Bayesian hierarchical methodologies were discussed (Chapters 2 and 4). For the estimation and smoothing of mortality curves, we propose a Bayesian relational dynamic Poisson model which have shown to be promising in many scenarios (Chapter 3).

The new methodology introduced in Chapter 2 has been accepted for publication at Bayesian Analysis Oliveira *et al.* [2020]. Such an approach allows the correction of underreporting bias provided that a hierarchical clustering structure for the areas of interest is available. Only prior information about the data quality in areas belonging to the best group is required to ensure model identifiability. In many situations, this approach might be less restrictive than some others proposed in the literature. We provide an analysis of infant mortality data in microregions of Minas Gerais, Brazil, and also an application to tuberculosis incidence in Brazilian subnational areas. The model has potential for application in many other epidemiological and environment problems.

Regarding to the modeling framework proposed in Chapter 2, interesting topics which deserve further investigation include the incorporation of traditional partition models within the proposed methodology as well as the exploration of spatial correlation in the reporting process. Teixeira *et al.* [2019] introduced a spatial clustering approach with basis on the well-known product partition model that can be explored to model the uncertainty about the clustering definition in our modeling framework. Another alternative for extension of the proposed model is the aggregation of counts measured in different periods of times. Authors such as Bracher and Held [2020] have recently proposed models for time series of underreported counts. By incorporating the time dependence into the modeling, it might be possible to required more feasible prior information to guarantee the model identifiability in general applications such as, for example, the underreporting level in more recent years where, in many fields of application, data are known as having better quality.

Still regarding the estimation of rates from underreported count data, the method present in Chapter 4 may provide valuable results in many practical situations in which the classical Poisson model is vulnerable to considerable bias. That method is based on the modeling framework introduced in Oliveira [2016], called random-censoring Poisson model (RCPM). Such an approach has been published at Statistics in Medicine [Oliveira *et al.* , 2017] after the substantial improvements made during the development of this dissertation. More specifically, the contributions for the RCPM involve the development and implementation of a correct and more efficient MCMC scheme to sample from the target posterior distributions (see the two last paragraphs of Section 4.2.1.3). A broader simulation study was also performed, including a comparison with the approach introduced by Moreno and Girón [1998]. A sensitivity analysis under both, the RCPM and the Moreno and Girón [1998]'s approaches, was accomplished. The application to infant mortality data in Minas Gerais, Brazil, was also improved by including a sensitivity analysis to the prior specifications for the censoring probabilities $\boldsymbol{\pi}$ and also by considering datasets from different periods of time. We found that, in a general context, the degree of underreporting bias correction is dependent on the level of information imposed by the prior distribution specified for the censoring probabilities.

The RCPM as presented in Chapter 4 is formulated to allow conjugate Gamma prior distributions for the parameter representing the event rates. A venue for further investigation is the inclusion of more flexibility into this model structure. For instance, the use of a regression model for the log-rates and also allowing for the inclusion of spatial random effects. The great difficult in doing that is the implementation of an efficient data-augmented MCMC scheme for sampling from the resulting posterior distributions. After doing that, an interesting strategy is to make the modeling and estimation scheme available for use by practitioners from different areas, for instance, through an R package or similar. The extension of the proposed random censoring Poisson model for a spatio-temporal context is also an interesting point for further investigation.

Besides the problems with imperfect vital registration systems, such as underreporting, another issue commonly faced when analyzing mortality data is the sparsity and high variability experienced by small subnational areas. Demographers and statistical epidemiologists have significantly improved estimates for small-area mortality schedules in recent years. We approach such a problem in Chapter 3 considering the estimation of mortality schedules by single-year age intervals and sexes. We propose a Bayesian regression model to relate the mortality rates to a standard mortality schedule obtained from the Human Mortality Database (HMD) [Wilmoth *et al.*, 2020]. Since there exists high sampling variability in observed mortality rates in small areas, the standard schedule is included for preventing the estimated mortality curve to departure from the pattern usually observed in human populations. The use of such relational models is a common strategy in demography (see references in Chapter 3). We present two different alternatives related to a Poisson and a Gaussian likelihood function, being the Poisson the most appropriated because it allows direct use of the observed null death counts. Attempting to provide smoothed estimates, we consider a dynamic dependence structure to ”borrow strength” across the age mortality rates thus giving the status of a time series for the data observed in sequential age intervals. Preliminary results are presented using both real and simulated datasets. Performance of the proposed Bayesian regression model is compared to a competing approach proposed in the literature. The model provided accurate measurement of age-specific mortality rates in many simulated scenarios, being less suitable for quite small and erratic populations. In the application to Brazilian municipality data, the models provided similar results except for some cases with around 1,000 exposures individuals. The limitations of our model in quite small areas must be further explored.

As discussed in Chapter 3), along with the realization of a broader simulation study, there are many open topics for future investigation under the proposed modeling framework. It includes the implementation of the model as a proper generalized dynamic linear model using, for instance, the well-known Kalman filtering and Kalman smoothing algorithms [West, Harrison and Migon, 1985; Campagnoli *et al.*, 2009] or appropriate Markov chain Monte Carlo techniques for dynamic generalized linear models [Gamerman, 1988; Schmidt and Pereira, 2011]. We suspect that such an strategy may increase flexibility and it may allow us to control the influence of the data across the dynamic structure. It may attenuate some drawbacks of our dynamic Poisson model, such as the strong influence of outliers observed in cases of very small populations. After doing that, it may worth investigating the combination of the proposed modeling framework with the Poisson model proposed in Chapter 2) to account for underreporting bias, possibly also allowing for zero inflation Piancastelli and Barreto-Souza [2019]; Gonçalves and Barreto-Souza [2020] and considering other models which naturally account for overdispersion. With this, we may have a model that effectively and reliably estimate the mortality schedule accounting for both sparsity and underreporting in the observed counts.

To conclude, we highlight that the methodological contributions of this dissertation are motivated by their intention to enable better estimation of mortality rates with basis on defective count data. The types of data issues discussed here are only a few among many others not covered in this work, e.g., delayed reporting, preferential sampling, overreporting or, more generally, misreported data. As in many other studies of flawed count data, the methods presented in this dissertation have limitations in terms of the type of extra information that is available. It is known that, when data faces problems such as underreporting and sparsity, it is not possible to perform bias correction in the observed rates if no extra information is incorporated into the modeling framework or if a more complex model structure is not considered. That said, we finish by highlighting the potential of the proposed models in several practical situation, since they figure as an alternative for statistical analyses of flawed count data in conditions where currently available methods might not be applicable.

References

Bracher, J. and Held, L. (2020). A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts. *arXiv:2003.05885 [stat.ME].*

Campagnoli, P., Petrone, S. and Petris, G. (2009). Dynamic Linear Models with R. Springer-Verlag New York.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalized linear models. *Biometrika*, **85**(1), 215–227.

Gonçalves, J.N., and Barreto-Souza, W. (2020). Flexible regression models for counts with high-inflation of zeros. *METRON*, **78**, 71–95.

Oliveira, G.L. (2016). Modeling Underreported Infant Mortality Data with a Random Censoring Poisson Model. Master's thesis, Statistics Department, Universidade Federal de Minas Gerais, Belo Horizonte. Available at http://hdl.handle.net/1843/BUBD-A89PEH.

Oliveira, G.L., Loschi, R.H. and Assunção, R.M. (2017). A random-censoring Poisson model for underreported data. *Statistics in Medicine*, **36**(30), 4873–4892.

Oliveira, G.L., Argiento, R., Loschi, R.H., Assunção, R.M., Ruggeri, F. and Branco, M.D. (2020). Bias correction in clustered underreported data. To appear at *Bayesian Analysis.*

Piancastelli, L.S.C., and Barreto-Souza, W. (2019). Inferential aspects of the zero-inflated Poisson INAR(1) process. *Applied Mathematical Modelling*, **74**, 457–468.

Schmidt, A.M. and Pereira, J.B.M. (2011) Modelling Time Series of Counts in Epidemiology. *International Statistical Review*, **79**(1), 58–69.

Teixeira, L.V. Assunção, R.M.and Loschi, R.H. (2019). Bayesian space-time partitioning by sampling and pruning spanning trees. *Journal of Machine Learning Research*, **20**(85), 1–35.

United Nations. (2014) . Principles and Recommendations for a Vital Statistics System. Department of Economic and Social Affairs Statistics Division of the United Nations, New York. Revision 3.

West, M., Harrison, J. and Migon, H. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, **80**, 73–83.

Wilmoth, J.R., Andreev, K., Jdanov, D., Glei, D.A. and Riffe, T. with the assistance of Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., Vachon, P., Winant, C. and Barbieri, M. (2020). Methods Protocol for the Human Mortality Database. University of California at Berkeley (United States) and the Max Planck Institute for Demographic Research (Rostock, Germany). Last Revised: August 8, 2020 (Version 6). Available at https://www.mortality.org/Public/Docs/MethodsProtocol.pdf.