Modeling Underreported Infant Mortality Data with a Random Censoring Poisson Model

Guilherme Lopes de Oliveira

Departamento de Estatística - ICEx - UFMG

February 26, 2016

Modelando Dados de Mortalidade Infantil Sub-Registrados usando um Modelo Poisson com Censura Aleatória

Guilherme Lopes de Oliveira

Orientador: Rosangela Helena Loschi
Co-orientador: Renato Martins Assunção

Defesa de dissertação para a obtenção do grau de Mestre em Estatística junto ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais.

Departamento de Estatística
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais

Belo Horizonte, MG - Brasil
26 de Fevereiro de 2016

To my parents and brothers.

“In the end, it’s not the years in your life that count.
It’s the life in your years.”

Abraham Lincoln

Acknowledgements

It was a grateful journey and I have to thank everyone who somehow contributed to this achievement.

First, I thank God for giving me health and strength to overcome the problems and to always keep me going ahead.

Second, I thank my parents and brothers for always believe in my potential and never let me give up.

I thank my friends that always motivated me and made me laugh on stress times, specially Ana Cláudia Silva, Ebert Lima, Rafael Aguiar and Vinícius Lacerda.

My friends and colleagues from the masters course also deserve my thanks for all the shared knowledge, for the constant help during the studies and for the *unstresseful conversations*, especially Edson Ferreira, Erick Amorim, Jéssica Almeida, Natália Araújo and Uriel Silva.

I thank my Professors Rosangela Loschi and Renato Assunção for their trust, support, patience, commitment and for always being encouraging me to continue studying.

To Professors Deise Campos, Leonardo Bastos and Wagner Souza I thank for attending the examination board and contribute to the improvement of the final results of my dissertation.

Also, I thank the professors and staff of the Departamento de Estatística da UFMG who somehow had helped me along this journey, in especial to Rogéria Figueiredo and Professor Ela Mercedes de Toscano.

Finally, I thank to CAPES for the scholarship during the Master's course, to CNPq for the financial support during the undergraduate course and FAPEMIG for the support to participate of conferences throughout my journey at UFMG.

Thank you all, Guilherme Oliveira.

Abstract

In poor and socially deprived areas, economic, social and health data are typically underreported. As a consequence, inference using the observed counts for the event of interest will be biased and risks will be underestimated. To overcome this problem, Bailey et al. (2005) propose to consider data from suspected areas as censored information and develop a spatial Bayesian approach for the so-called Censored Poisson model (CPM). However, the CPM assumes that all censored areas are precisely known *a priori*, which is not a simple task in many practical situations. To account for potential underreporting in an infant mortality dataset, we propose an extension on the CPM by jointly modeling the data generating and the data reporting processes. We assume that observed counts have a Poisson distribution and the underreporting probabilities are associated to an appropriate logistic model. By doing that, we introduce the Random Censoring Poisson model (RCPM) in which the censoring mechanism is treated as random instead of requiring a previous specification of the censored (underreported) areas. Informative priors on the data reporting process are considered. We also propose a MCMC sampling scheme based on the data augmentation technique. By artificially augmenting the data through latent variables, we facilitate the posterior sampling process. To evaluate the proposed model, we run a simulation study in which such a model is compared with the CPM using different fixed censoring criteria. Also, we apply the proposed model to map the early neonatal mortality rates in Minas Gerais State, Brazil, where data quality is truly poor in many regions.

Keywords: underreporting, infant mortality, Censored Poisson model, data augmentation.

Resumo

Em áreas pobres e socialmente mais desfavorecidas, dados econômicos, sociais e de saúde são tipicamente subnotificados. Consequentemente, inferência utilizando as contagens observadas para o evento de interesse será tendenciosa e os riscos inerentes serão subestimados. Para contornar este problema, Bailey et al. (2005) propõem considerar os dados provenientes de áreas suspeitas como informações censuradas e desenvolvem uma abordagem Bayesiana espacial para o chamado modelo Poisson Censurado (MPC). Este modelo assume que todas as áreas censuradas são precisamente conhecidas *a priori*, o que não é uma tarefa simples em muitas situações práticas. Então, para levar em conta uma potencial subnotificação em um conjunto de dados de mortalidade infantil, nós propomos o modelo Poisson Censurado Aleatoriamente (MPCA) como uma extensão do MPC através da modelagem conjunta dos processos de geração e de reportação/registro dos dados em vez de requerer uma pré-especificação das áreas censuradas. Assume-se que as contagens observadas têm uma distribuição Poisson e as probabilidades de subnotificação são associados a um modelo logístico apropriado. Distribuições *a priori* informativas são consideradas para o processo de reportação dos dados. Propomos também um esquema de amostragem MCMC baseado na técnica de aumento de dados. Aumentando artificialmente os dados através de variáveis latentes, nós facilitamos o processo de amostragem *a posteriori*. Para avaliar o modelo proposto, apresentamos um estudo de simulação em que tal modelo é comparado com o MPC usando diferentes critérios de censura fixos. Por fim, o modelo proposto é aplicado no mapeamento do risco relativo de mortalidade neonatal precoce no Estado de Minas Gerais, Brasil, onde a qualidade dos dados é verdadeiramente precária em muitas regiões.

Palavras-chave: subnotificação, mortalidade infantil, modelo Poisson Censurado, aumento de dados.

List of Figures

List of Tables

List of Nomenclatures

Nomenclature	**Description**
AI	Adequacy Index
BHM	Bayesian Hierarquical model
CAR	Conditional Autoregressive model
cpf	cumulative probability function
CPM	Censored Poisson model
ENM	Early Neonatal Mortality
fcd	full conditional distribution
FI	Functional Illiteracy
HDI	Human Development Index
HPD	Highest Posterior Density interval
IBGE	*Instituto Brasileiro de Geografia e Estatística*
IdC	Infant Deaths with Ill-defined Cause
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCMC	Markov Chain Monte Carlo
MG	Minas Gerais State
MSE	Mean Squared Error
MSPE	Mean Squared Percentage Error
pf	probability function
RCPM	Random Censoring Poisson model
RR	Relative Risk
SIH	*Sistema de Informações Hospitalares*
SIM	*Sistema de Informações sobre Mortalidade*
SINASC	*Sistema de Informações sobre Nascidos Vivos*
SMR	Standardized Mortality/Morbidity Ratio
SUS	*Sistema Único de Saúde*

Contents

Chapter 1

Introduction

Studies related to geographical distribution of disease or mortality incidence and its relationship with potential risk factors have been a rich field for the development of new statistical methods and models. The mapping of the relative risks inherent to this events in small geographical areas is commonly called *disease mapping* and it plays an important role, for instance, in suggesting etiological hypotheses as well as to guide epidemiological control and government intervention. In recent years, powerful and flexible statistical tools have been proposed for disease mapping. A general overview in this topic can be found in Besag (1974), Breslow and Day (1987) and Lawson et al. (1999), to cite a few.

Many epidemiological studies focus on *early deaths* rates because they are indicators of the population health condition. Among them, the *infant mortality* rate is considered of major interest. The mortality rates commonly have their level associated with socioeconomic factors that determine the living conditions and the performance of health services, such as the access and the quality of medical care. To assess the magnitude of the infant mortality rate it is common to use the *early neonatal mortality* (ENM) rate. The ENM is understood as the infant deaths that occur in the first seven days of life and it is obtained by dividing the number of children that died in the first seven days of life by the total of live births in the period of interest.

In recent decades, the early neonatal mortality has risen its participation in the Brazilian infant mortality (Schramm and Szwarcwald, 2000a). Social and economic inequality are the main factors to explain the increase in the death risk of newborns, since these inequalities are usually associated with maternal health problems and difficulties in accessing neonatal care.

For appropriately planning interventions in the ENM rate, the Brazilian health authorities must receive correct information about infant deaths and live births. In Brazil, data related to mortality are continuously collected since 1976 through the *Sistema de*

Informações sobre Mortalidade (SIM) and data related to live births are collected by the *Sistema de Informações sobre Nascidos Vivos* (SINASC) since 1990, both systems were implemented by the Brazilian Ministry of Health. Due to the data continued collection by the SIM and SINASC, these systems are the best available data sources for monitoring infant mortality in Brazilian municipalities.

However, even with the advances achieved in recent years with relation to data collection systems, in developing countries, such as Brazil, several information are not correctly recorded as they should be. In fact, Schramm and Szwarcwald (2000b), MS-Brasil (2004), Machado et al. (2006), Campos et al. (2007), Frias et al. (2008), Lima et al. (2009) and Guimarães et al. (2013) indicate that information on infant mortality and live births are not correctly recorded in the Brazilian SIM and SINASC systems, mainly in socially deprived areas where the educational level is also precarious. Therefore, the *underreporting* problem may be present in several statistical analysis, especially when one is using Brazilian public health data.

The dataset that motivated this work corresponds to the number of live births and early infant deaths that took place in public hospitals of the 853 municipalities of Minas Gerais State (MG) between 1999 and 2001. Note that, given the serious problems related to underreporting of deaths and births in the SIM and SINASC, we consider data available at the *Sistema de Informações Hospitalares* (SIH) of the Brazilian *Sistema Único de Saúde* (SUS) because Schramm and Szwarcwald (2000b) and Campos et al. (2007) indicate that SIH provides more reliable information than SIM and SINASC.

According to Campos et al. (2007) and references therein, in MG the early neonatal mortality rate is very high compared to those observed in the other States of Brazilian Southeast and South regions, mainly in more socially deprived areas in the North of MG. Although most of the ENM occurs in hospitals, in MG hospitals are heterogeneously distributed around the State, making the access to health care in socially deprived areas quite difficult. Also because of this, data on ENM are usually underreported and the quality of information produced in the State is quite poor. In fact, MG is the only state in the Brazilian Southeast region where the official infant mortality rates are estimated by the *Instituto Brasileiro de Geografia e Estatística* (IBGE) using indirect methods (MS-Brasil (2004) and Ortiz (2000)).

Figure 1.1 presents a first study involving our dataset of interest. The 853 municipalities of Minas Gerais State are grouped into 75 regions in order to avoid regions with very small or zero counts, which leads to unstable estimates for the mortality rates as discussed in Assunção et al. (1998).

Figure 1.1 (left) displays the maximum likelihood estimates for the relative risk (RR) of ENM in Minas Gerais State obtained by fitting the standard Poisson model (see Section 2.1). The results indicate that northern regions of MG experience the lowest ENM rates, which are close to those observed for highly developed countries. This estimates are inconsistent with the expected by epidemiologists for those regions, because North and Northeast of MG are poorly developed regions and they present the worst social indicators in the State, as can be seen in Figure 1.1 (right) that displays the Human Development Index (HDI) in 2000 for the $n = 75$ regions of MG.

The standard Poisson model does not account for the spatial correlation among neighbouring areas. Considering such type of correlation is a strategy commonly used to smooth and to overcome inconsistencies in the mortality rates estimation. Araújo and Loschi (2013) mapped the relative risks of ENM in Minas Gerais State using a Bayesian approach for the Spatial Poisson model that includes covariates and spatially structured random effects (see Section 2.1.1). The posterior means for the RR obtained under such a model are displayed in Figure 1.1 (middle).

Despite this more sophisticated model considers the spatial correlation between neighboring areas, it does not overcome the underestimation of the relative risk in the poorest regions in North and Northeast of MG, since the estimates in these regions remained below than the expected and similar to those seen in highly developed countries. It is noticeable from Figure 1.1 that the maximum likelihood (left)and the Bayesian (middle) estimates for the ENM in MG are quite similar. Such results raise some doubts regarding the quality of early infanty mortality and live births data collected from the SIH, as had already been observed with relation to data from SIM and SINASC.

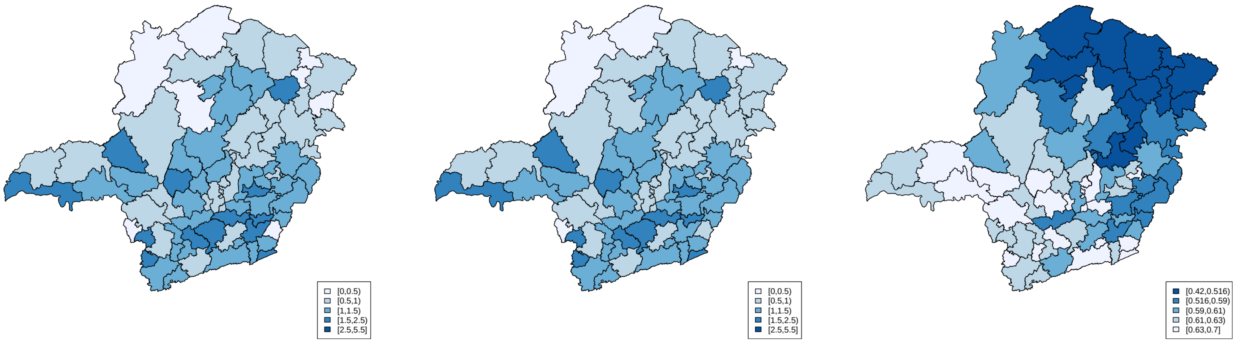

Figure 1.1: RR's estimates of ENM in MG using the SMR (left) and the posterior mean under the SPM in Araújo and Loschi (2013) (middle); and the Human Development Index in MG (right) [Source: IBGE 2000].

In fact, a possible explanation for those inconsistencies in the ENM's relative risk estimation for the northern regions of MG is the occurrence of underreporting on the SIH's data, as discussed in Campos et al. (2007). As pointed out by those authors, the underreporting of live births and early infant deaths discloses the socioeconomic, geographical and cultural inequalities as well as the marginalization of population groups in MG. Consequently, official statistics for the State do not measure the death risk of newborns in all its magnitude, making difficult an adequate epidemiological analysis.

In addition to Campos et al. (2007), Andrade et al. (2006) also indicate a possible relationship between socioeconomic status and quality of mortality information when analyzing data from Paraná State, concluding that in areas facing socioeconomic problems the occurrence of underreporting is quite plausible. Also analyzing Brazilian public health data, Bailey et al. (2005) discuss about underreporting in leprosy incidence in Olinda city, Pernambuco State.

The important point is that, if underreporting occurs and it is not accounted for, inference using the observed counts will be biased and, consequently, the inherent relative risks will be underestimated. Whatever the source, underreporting will invalidate the assumptions of the standard Poisson model which is conventionally used in count data problems. Therefore, when using data with suspected underreporting, it is quite important to use models that account for it. Though information about unreported events is missing, underreporting is different from the usual concept of missing data. In usual missing data problems, information that data are missing is available and hence can be incorporated in the analysis, whereas for an unreported event no information at all is generated.

To overcome the underreporting problem in their leprosy incidence dataset, Bailey et al. (2005) propose to consider data with suspected underreporting as censored information and use a Censored Poisson model (CPM) that takes into account the spatial association among neighboring areas. The big challenge in considering the CPM is the prior definition of all censored (underreported) areas that is needed in its construction. Usually, information about the censored areas are obtained indirectly and *ad-hoc* procedures are considered to determine them. For their particular application, Bailey et al. (2005) use a social deprivation indicator as a criterion for considering data from certain regions as unreliable data (underreported data).

However, to account for potential underreporting, it would be more appropriate the specification of a joint model for the data generating and the data reporting processes. In this context, with an application to workers absenteeism data from the German Socio-Economic Panel, Winkelmann (1996) derives a modified latent Poisson regression model that allows for underreporting in the counts. It is assumed that the true (but unobserved) absent counts are generated by a Poisson process and the reporting/non-reporting is independent for each event. As far as we know, that is the only approach proposed in the Literature for jointly model the data generating and the data reporting processes in the context of underreported count data.

Dvorzak and Wagner (2015) extend the model from Winkelmann (1996) by incorporating both cluster analysis and Bayesian variable selection to estimate risk of cervical cancer death using underreported data. In their extension of Winkelmann (1996)'s model, the intensity of the Poisson process and the reporting probability are both related to a set of potential covariates through the specification of regression models. Identification of the proposed model requires additional information, which can be provided either by additional data on the reporting process (validation data), parameter restrictions or informative prior on parameters, e.g., provided by experts.

In this work, we introduce a different approach for jointly model the data generating and the data reporting processes in the context of underreported data. We do that by extending the CPM presented in Bailey et al. (2005). Basically, we introduce on the CPM a random mechanism to specify the censored (underreported) areas, instead of using a fixed vector to previously indicate those censored ones. Therefore, as opposed to what is found in Bailey et al. (2005), we can now estimate the probability of the information in each area being censored at the same time that the relative risk for the event of interest is being estimated. We call the proposed model by Random Censoring Poisson model (RCPM). Therefore, the RCPM arises as an alternative model to that one proposed in Dvorzak and Wagner (2015) to handle underreported count data.

We develop an algorithm to sample from the posterior distribution that relies on the data augmentation strategy (Tanner and Wong (1987) and Chib (1992)), which simplify substantially the posterior sampling process. We run a Monte Carlo simulation study for comparing the RCPM that is been proposed with the CPM from Bailey et al. (2005), in which the censored areas must be previously specified. We consider different scenarios for generating the datasets in such a simulation study. Moreover, we consider the proposed model to analyze the ENM's data in Minas Gerais State. Results are compared with those ones obtained by using the CPM under three different fixed censoring criteria proposed in Oliveira and Loschi (2013).

This work is organized as follows. In Chapter 2, we review the standard and Spatial Poisson models that are widely considered for disease mapping and we present the Censored Poisson model proposed in Bailey et al. (2005), including some studies with simulated data. In Chapter 3, we propose the Random Censoring Poisson model (RCPM).

Simulation studies on that model are performed and results are compared with those ones provided by the CPM. An application of the proposed model to the ENM data from MG are presented in Chapter 4. Chapter 5 closes this work by presenting the main conclusions and some discussions about the results. We also present some topics for future research on extending the proposed model.

Chapter 2

Background and Theoretical Framework

The goal of this chapter is to introduce the basic elements needed to understand the problem we have at hand, a solution already presented in the Literature and the solution proposed in Chapter 3. Firstly, some methods widely used in the context of disease mapping will be discussed, including the Spatial Poisson model (Besag et al., 1991). We then present the Censored Poisson model (CPM) proposed in the Literature to model underreported count data. Some simulation studies involving the CPM will be performed in order to better understand its advantages and disadvantages.

We start highlighting that the use of maps to display the geographical variability of the relative risk (RR) inherent to certain events as disease or mortality is quite popular nowadays. The statistical problem of searching for efficient models to appropriately mapping those quantities has received considerable attention recently, particularly because this kind of maps can help us to detect areas where the event is especially prevalent and also for detecting previously unknown risk factors. The risk may reflect actual deaths due to a disease (mortality) or, if it is not fatal, the number of people who suffer from a disease (morbidity) for the population at risk in a certain period of time. Hence, for doing such analysis, basic data must include information about the population at risk and the number of cases in each area.

The term *disease mapping* is concerned with the estimation of the true underlying distribution of disease or mortality rates for a given event, disclosing the spatial patterns among the associated rates. In general, the goal is to map the spatially smoothed rates for each area by borrowing information from neighbouring areas. Most of the existing methods for disease mapping are more suitable for detecting gradual regional changes rather than detecting abrupt changes associated with clustering. This text review some methods related to the former. Readers interested in clustering analysis in the disease mapping context may see, for instance, Besag et al. (1991), Holmes et al. (1999),Knorr-Held and Best (2001), Denison and Holmes (2001), Hegarty and Barry (2008) and Teixeira et al. (2015).

The type of data more commonly encountered in disease mapping is the count by area or *areal data*, since in most cases the event exact locations are unknown due to medical confidentiality, for instance. Throughout this work only areal data will be considered.

To establish notation, along this work Y_i and E_i will denote, respectively, the observed and the expected number of cases in area $i = 1, ..., n$. The Y_i is a random variable that assumes the value y_i after observation. The quantity E_i is fixed and a known function of the number n_i of individuals at risk in area i given by

$$E_i = n_i \left(\frac{\sum\limits_i y_i}{\sum\limits_i n_i} \right) = n_i \overline{r},$$

where $\overline{r} = \sum_i y_i \left[\sum_i n_i\right]^{-1}$ denotes the overall disease (mortality) rate in the whole region.

This chapter is organized as follows. In Section 2.1 we present the standard and the Spatial Poisson models. In Section 2.2 the Censored Poisson model (CPM) is discussed and presented as an alternative approach to handle underreported count data. Section 2.3 presents some simulation studies involving the CPM.

2.1 Poisson Model

To assess the status of an area with respect to the incidence of an event, it is convenient to firstly obtain the expected incidence given the population at risk in the area and then compare it with the observed incidence. This approach has been traditionally used in the analysis of counts within areas or sub-regions. The ratio of observed to expected counts in each area is called *Standardized Mortality/Morbidity Ratio* (SMR) and it is given by

$$SMR_i = \frac{y_i}{E_i}, \tag{2.1}$$

for $i = 1, ..., n$. The ratio in expression (2.1) gives a naive estimator of the relative risk (RR) in area i (Breslow and Day, 1987).

Maps built using the SMR are often a starting point in disease mapping. However, many events of interest are uncommon or rare and sometimes area i is relatively small. In both cases, the SMR tends to present high variability with extreme values tending to occur in areas with the small populations. Therefore, SMR can be useless for mapping the desired relative risks. Mapping based on that quantity can also fail when dealing with underreported data, since the regions of greatest potential interest are usually small and often associated with the less reliable data. Actually, the SMR is not an appropriate choice in a great number of situations.

Alternatively, to estimate the relative risks we can assume that the observed count of cases Y_i in each area i has a Poisson distribution with mean $\mu_i = E_i\theta_i$, where θ_i denotes the true relative risk associated to this area. In that case, if $\boldsymbol{Y} = (Y_1, ..., Y_n)$ are independent, given $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, so that

$$Y_i|\theta_i \overset{ind}{\sim} \text{Poisson}(E_i\theta_i); \tag{2.2}$$

and, if θ_i is taken as a fixed effect associated to area i, then the maximum likelihood estimator for θ_i is the SMR_i given in expression (2.1). Consequently, this approach is not a good choice for mapping rare events data. Despite this, some inference results considering parameter θ_i as a fixed effect under the standard Poisson model presented in (2.2) can be found in Breslow and Day (1987); such as classical confidence intervals and hypothesis tests. In Section 2.1.1, we briefly discuss about an efficient strategy to better estimate the relative risks $\boldsymbol{\theta}$ that consists in considering the spatial correlation among the neighbouring areas.

2.1.1 Spatial Poisson Model

As previously discussed, the main goal in disease mapping is to research for methods capable to produce more reliable maps of the underlying geographical variation in disease or mortality risks. According to Bailey (2001), a good method must be capable to reduce the excess of local variability as well as to correct the variations produced by risk factors or population differences such as age, sex and so on.

To deal with this issue, some smoothing of the realtive risks is incorporated into the standard Poisson model in (2.2) by taking θ_i as a random effect (or a function of random effects). This strategy allows for overdispersion in the standard Poisson model caused, for instance, by unobserved covariates or confounding factors (Mollié (1995) and Clayton and Bernardinelli (1996)).

Random effects can be directly associated to regions or covariates. A simple general specification of random effect models for count data was suggested by Besag et al. (1991) and the most simple or natural way to handle these models is to adopt a Bayesian framework. In the Bayesian context, a random effect model is called Bayesian hierarchical model (BHM), where a prior distribution is assigned to each θ_i (e.g., see Breslow and Day (1987)).

Therefore, in Bayesian disease mapping, a BHM combines two types of information: the one provided by the observed counts in each region, usually summarized by the Poisson likelihood $\pi(\boldsymbol{Y} \mid \boldsymbol{\theta})$, and the prior information about the relative risk behavior in the overall map, summarized by its prior distribution $\pi(\boldsymbol{\theta})$. Therefore, $\pi(\boldsymbol{\theta})$ should reflect the prior knowledge about the variation in relative risk over the map Bernardinelli et al. (1995).

As in model (2.2), assume that the counts $\boldsymbol{Y}$ in the n different areas are independent given $\boldsymbol{\theta}$, so that

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

The Spatial Poisson model suggested by Besag et al. (1991) considers the relative risk θ_i as a function of random effects to allow for overdispersion produced by unobserved covariates or confounding factors as well as to reflect the explicit spatial dependence among $\boldsymbol{Y}$. It is assumed that

$$\log \mu_i = \log E_i + \log \theta_i = \log E_i + \upsilon_i + s_i, \tag{2.3}$$

where υ_i and s_i are, respectively, a non-spatially structured and a spatially structured random effect. The quantity υ_i accounts for a dependence among the counts $\boldsymbol{Y}$ induced by unmeasured covariates or confounding factors. Typically, υ_i does not account for an explicit spatial dependence among the counts which may arise, for instance, through lesser variability of rates on neighbouring densely populated urban areas as opposed to sparsely populated rural areas or through an infectious etiology of the disease. Thus, the spatially structured random effect s_i is introduced into the model to describe such spatial association. Details on this model can be found, e.g., in Besag et al. (1991), Mollié (1995) and Clayton and Bernardinelli (1996).

The typical prior assumption for υ_i is the Normal distribution $N(\mu_\upsilon, \sigma^2_\upsilon)$, with the hyperpriors being a Normal distribution for the hyperparameter μ_υ and a Gamma distribution for the precision hyperparameter $\tau^2_\upsilon = 1/\sigma^2_\upsilon$. The prior specification for $\boldsymbol{s} = (s_1, ..., s_n)$ must disclose the spatial dependence among the areas. Usually, it is assigned a Conditional Autoregressive model (CAR) as prior distribution for $\boldsymbol{s}$, in which the mean value for the marginal distribution of s_i is a weighted average of the neighboring random effects and the variance σ^2_s controls the strength of this local spatial dependence. A vague Gamma distribution is commonly assumed for the precision hyperparameter $\tau^2_s = 1/\sigma^2_s$. A formal definition of the CAR model and a detailed explanation on this topic can be found, for instance, in Besag and Kooperberg (1995) and Banerjee et al. (2004).

The basic Bayesian hierarchical model in (2.3) can be extended by including k covariates, $(x_{i1},...,x_{ik})$, related to suspected risk factors and so that

$$\log\ \mu_i = \log\ E_i + \sum_{j=1}^{k} \beta_j x_{ij} + \upsilon_i + s_i, \tag{2.4}$$

where μ_i, E_i, υ_i and s_i are defined as before. The parameter β_j is a fixed effect associated to the j-th covariate, $j = 1,...,k$, and reflects the influence of this covariate on the log relative risk given by $\log\ \theta_i = \sum_{j=1}^{k} \beta_j x_{ij} + \upsilon_i + s_i$. A constant term β_0 can also be considered, so that $\log\ \theta_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \upsilon_i + s_i$. Usually it is assumed that the fixed effects $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ are independent and identically distributed (iid) according to a non-informative (vague) prior distribution, e.g., a zero centered multivariate Normal distribution having a diagonal covariance matrix with large variances. MCMC methods are used to sample from the joint posterior distribution $\pi(\beta_0, \boldsymbol{\beta}, \boldsymbol{s}|\boldsymbol{Y})$. Further details and variations on this basic modeling framework may be found in many published examples of ecological and epidemiological studies (e.g. see Lawson et al. (1999) and references therein).

2.2 Censored Poisson Model

The Censored Poisson model (CPM) was firstly proposed by Terza (1985). Despite being mainly considered for modeling count data that exhibit either over or under-dispersion (Cameron and Trivedi, 1998) it has also been considered for modeling censored data as discussed in Famoye and Wang (2004), where a generalization of the CPM is developed to handle general types of censoring. By simplicity and for our purpose, we only consider right censored data.

As in the previous models, let Y_i be the count for the event of interest occurred in area i, $i = 1,...,n$, and assume that

$$Y_i|\mu_i \overset{ind}{\sim} \text{Poisson}(\mu_i).$$

The assumption that all Y_i are completely observed is unrealistic in many count data applications. Rather, it is possible that the reported number of events y_i constitutes only a fraction of all events and, thefore, data are underreported.

To built the CPM, consider that some observable variables Y_i are not completely observed and thus they are considered as being censored. If no censoring occurs for the

i-th observation y_i, the complete information is considered, that is, $Y_i = y_i$. However, if a censoring occurs for the i-th observation, the true number of cases associated to this observation is considered at least equal to the observed value, i.e., $Y_i \geq y_i$. Denote by γ_i the censoring indicator variable such that

$$\gamma_i \; = \; \begin{cases} 1, & \text{if information at area } i \text{ is censored,} \\ 0, & \text{otherwise.} \end{cases}$$

In Famoye and Wang (2004)'s approach, the censoring vector $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_n)$ must be known and fixed *a priori*.

Assuming independence between the counts Y_i, given γ_i and μ_i, for $i = 1, ..., n$; and also assuming that the censoring mechanism is independent of the number of events in each area, Famoye and Wang (2004) built the likelihood function that characterizes the CPM as being

$$\begin{aligned} L(\boldsymbol{\mu}; \boldsymbol{y}, \boldsymbol{\gamma}) \; &= \; \prod_{i=1}^{n} \left\{ \left[f_{Y_i|\mu_i}(y_i) \right]^{1-\gamma_i} \; \left[1 - F_{Y_i|\mu_i}(y_i - 1) \right]^{\gamma_i} \right\} \\ &= \; \prod_{i=1}^{n} \left\{ \left(\frac{e^{\mu_i} \mu_i^{y_i}}{y_i!} \right)^{1-\gamma_i} \; \left(\sum_{y \geq y_i} \frac{e^{\mu_i} \mu_i^{y}}{y!} \right)^{\gamma_i} \right\}. \qquad (2.5) \end{aligned}$$

where $f_{Y_i|\mu_i}$ and $F_{Y_i|\mu_i}$ denote, respectively, the probability function (pf) and cumulative probability function (cpf) of a random variable with distribution Poisson(μ_i).

Bailey et al. (2005) propose to consider the CPM to handle underreported leprosy data from Olinda city, Pernambuco State, Brazil. Basically, the underreported counts are treated as censures considered to be lower bounds to the real (but unobserved) counts. In such approach, the censoring indicator vector $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_n)$ is fixed *a priori*, that is, we must precisely know all the areas whose counts are underreported. However, censored observations are not obviously identified as in Survival Analysis studies. In a general context, several criteria may be available to partitioning the observations as censored or non-censored. In their particular application, Bailey et al. (2005) make such a partition of the areas using values of a social deprivation indicator, based on the fact that the observations considered most unreliable are those from the poorer areas. The occurrence rate of leprosy is taken as being

$$\log \; \mu_i \; = \; \log \; E_i + \beta_0 + \beta \mathrm{x}_i + \upsilon_i + s_i,$$

where E_i denotes the expected number of leprosy cases in each area, x_i is the mean centered proportion of the population at risk in the i-th area with monthly income below one minimum legal wage and, respectively, s_i and υ_i are the spatially and non-spatially structured random effects in each census tract.

For posterior inference the MCMC methodology is considered. For the non-spatially structured random effects, υ_i, are assumed independent Normal distributions with zero mean and variance σ^2_υ. The joint prior distribution for the spatially structured random effects, $\boldsymbol{s} = (s_1, ..., s_n)$, is taken as a Gaussian intrinsic CAR model, where the mean value for $s_i|s_j$, for all $j \neq i$, is a weighted average of the neighbouring random effects and the variance, σ^2_s, controls the strength of this local spatial dependence. The neighbourhoods are defined by simple binary adjacency weights, that is, trough a proximity matrix W, in which the entries $\omega_{ij} = 1$ if area i shares a common boundary with area j and $\omega_{ij} = 0$ otherwise. An improper flat Uniform prior distribution is assigned for the intercept β_0 and a flat Gamma hyperprior distribution is assumed for the precisions $\tau = 1/\sigma^2_\theta$ and $\phi = 1/\sigma^2_\upsilon$. For parameters β it is used a vague zero mean Normal distribution.

Bailey et al. (2005) concluded that CPM provided reasonable estimates for the leprosy rates in Olinda. However, it is not trivial how to define the censored areas in real problems like that discussed in Bailey et al. (2005). An ideal model should also infer about the censored areas. For that purpose, in Chapter 3 we introduce a Poisson model that allows us to make inference about the censored areas at the same time that we infer about the inherent relative risks. Before doing that, we will perform some simulation studies involving the CPM in order to better evaluate the quality of its estimates. Such studies for the CPM in the context of underreported data is not presented in Bailey et al. (2005) neither in other papers. The details and results for the simulation studies are presented in Section 2.3.

2.3 Simulation Studies on the Censored Poisson Model

In this section we will present several studies considering artificial data generated in different scenarios. The goal is to investigate the quality of the estimates provided by the Censored Poisson model (CPM) presented in Section 2.2. Specifically, we will evaluate the effect of the *data censoring level*, i.e., effect of the proportion of information that is censored (underreported) in certain areas; the effect of the proportion of censored areas in the hole map when including spatial random effects in the relative risk modelling and, at last, the effect of clustering the areas. In some scenarios, we also evaluate the effect of the prior distribution assigned to the relative risks. We consider the $n = 75$ regions of the Minas Gerais State map. Computational programs to obtain the estimates were implemented in R 3.1.3 (www.r-project.org) and OpenBUGS (www.openbugs.net) languages. In all case, we run the MCMC for 85000 iterations, discarding the first 20000 draws as a burn-in period and considering a lag 13 to avoid correlation. Thus, we consider a sample of size 5000 for making posterior inference. Also, in all figures presented throughout this section the symbols + and ∘ will represent the censored and non-censored areas, respectively.

2.3.1 Simulation 1: The *Data Censoring Level* Effect

Our interest here is to investigate the behavior of the relative risk estimates for different censoring levels in the observed counts under the CPM. Assume that the event count y_i^* in area $i = 1, ..., 75$ is generated from a Poisson distribution such that

$$Y_i^*|\mu_i \overset{ind}{\sim} Poisson(\mu_i),$$

where μ_i is so that

$$\mu_i = \begin{cases} 70, & i = 1, ..., 25 \\ 30, & i = 26, ..., 50 \cdot \\ 10, & i = 51, ..., 75, \end{cases}$$

and denote by $\boldsymbol{\mu}^T$ this true values of $\boldsymbol{\mu}$ used to generate the data.

We assume 21 censored areas according to the following censoring indicator vector

$$\gamma_i = \begin{cases} 1, & i = 1, ..., 7 \text{ and } i = 26, ..., 32 \text{ and } i = 51, ..., 57 \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

Let δ be the censoring level desired, i.e., the proportion of the generated value y_i^* that will be correctly reported in each censored area. In this case, $(1 - \delta) \times 100\%$ of the information will be missed (underreported) if the area is a censored one. Thus, the counts y_i in censored areas are built by multiplying the generate value y_i^* by δ. Since the count must belong to the set of positive integer numbers, $\mathbb{Z}^+$, in all censored area the observation y_i is taken as the smallest integer greater or equal than the value obtained from the multiplication $y_i^* \times \delta$, that is,

$$y_i = \begin{cases} \lceil y_i^* \times \delta \rceil, & \text{if } \gamma_i = 1 \\ y_i^*, & \text{if } \gamma_i = 0. \end{cases}$$

For example, if $\delta = 0.8$ and $y^* = 100$ for a censored area, the reported value for this area will be $y = 80$. We assume $\delta = 0.9, 0.8, 0.7$ and 0.4. Thus, we have four simulated scenarios.

For modeling the simulated datasets, we fit the Censored Poisson model given by

$$\begin{aligned} Y_i|\gamma_i, \mu_i &\overset{ind}{\sim} CP(\mu_i) \\ \mu_i &\overset{ind}{\sim} Gamma(\alpha, \phi), \end{aligned}$$

where by notation $CP(\mu_i)$ we mean that observation Y_i has a Poisson distribution with rate μ_i and, given μ_i and γ_i, the contribution of this observation for the likelihood function corresponds to the i-th term of the function in (2.5).

The censored areas are pre-fixed as required in Bailey et al. (2005) and assumed to be those areas indicated in (2.6). We assume two different prior specifications for μ_i. For results presented in Figure 2.1 we choose a $Gamma(\alpha, \phi)$ so that *a priori* $E[\mu_i] = \mu_i^T$ and $Var[\mu_i] = 100$, whereas in Figure 2.2 we choose α and ϕ such that $E[\mu_i] = 40.0$ and $Var[\mu_i] = 100$, for $i = 1, ..., 75$. By doing this, despite of the large variance, we have one case in which the prior distribution represents well the generated data ($E[\mu_i] = \mu_i^T$, for all i) and another case in which the prior distribution has the same shape for all regions ($E[\mu_i] = 40.0$, for all i).

Figures 2.1 and 2.2 compare the posterior means of $\boldsymbol{\mu}$ obtained for each censoring level. The estimates of $\boldsymbol{\mu}$ in non-censored areas ($\circ$) are not influenced by the censoring level being quite close in all cases. It is also noticeable that better estimates are achieved in non-censored areas when the prior distribution is centered on $\boldsymbol{\mu}^T$ (Figure 2.1) than when the prior distribution is centered on an arbitrary value for all areas (Figure 2.2). In censored regions ($+$), the estimates of $\boldsymbol{\mu}$ are more similar when the associated censoring levels in the data are closer. From Figure 2.1, we noticed that the posterior means in censored areas tends to overestimate the true relative risk, mainly in datasets in which a lower censoring level is assumed. Figure 2.2 discloses that the posterior mean in censored areas tends to approximate of the prior mean. We also noted that the influence of prior information depends on the censoring level: as the censoring level increases, the influence of the prior mean also increases.

Those results bring evidences about the importance of choosing an adequate prior distribution for $\boldsymbol{\mu}$, mainly in censored regions. Since the prior information is truly important for posterior inference, in practice, information provided by experts on the area of interest it is of great importance. Moreover, non-informative prior must be avoided, unless we really do not have any prior information.

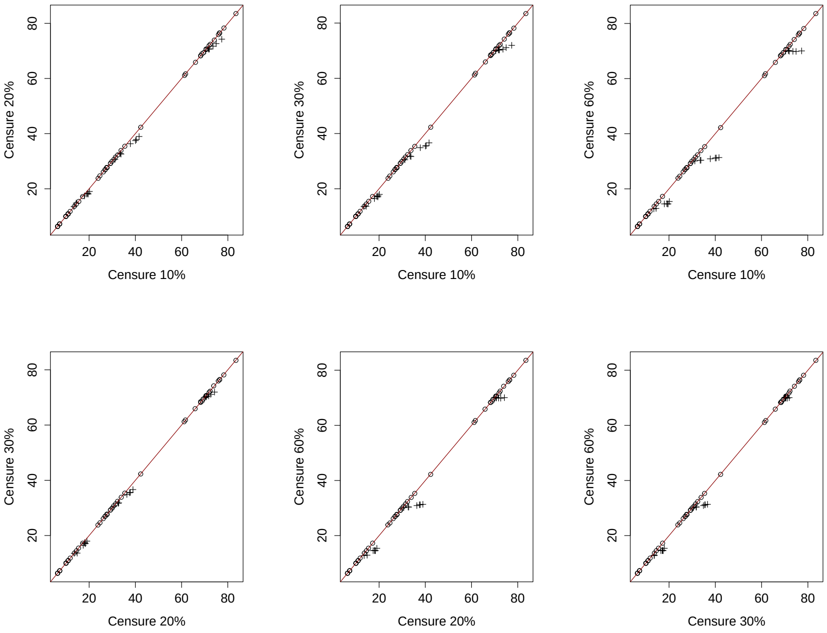

Figure 2.1: Comparing the posterior means of $\boldsymbol{\mu}$ for different censoring levels δ assuming prior mean $E[\mu_i] = \mu_i^T$ and prior variance $Var[\mu_i] = 100$.

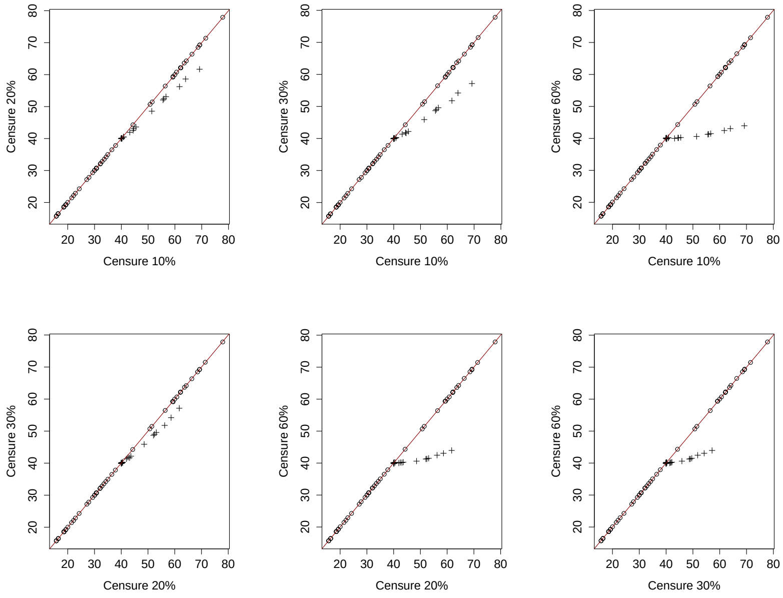

Figure 2.2: Comparing the posterior means of $\boldsymbol{\mu}$ for different censoring levels δ assuming prior mean $E[\mu_i] = 40.0$ and prior variance $Var[\mu_i] = 100$.

2.3.2 Simulation 2: The Proportion of Censored Areas Effect

In this section we consider five censoring criteria that induces different proportions of censored areas in the data. We fit three different Censored Poisson models such that, in Case 1, the relative risk (RR) is considered to be a spatially structured random effect, in Case 2, the RR is considered to be a spatially non-structured random effect and, in Case 3, the RR is a function of spatially and non-spatially structured random effects. Therefore, at the same time, we are evaluating the effect of the number of censored areas and the effect of including or not including a spatial random effect for modeling the relative risks $\boldsymbol{\theta}$.

Considering the known latitudes (denoted by Lat) inherent to the $n = 75$ regions of MG map, data are generated assuming an increasing relative risk from the South to the North, so that $\theta_i = \exp\{a + bLat_i\}$, for $i = 1, ...75$. To determine the values of a and b, we fixed that the region with the smallest latitude has $\theta = 0.3$ and the region with the greatest latitude has $\theta = 3.0$ and solve the following equation system

$$\begin{cases} \exp\{a + b\min(Lat_i)\} = 0.3 \\ \exp\{a + b\max(Lat_i)\} = 3.0, \end{cases} \tag{2.7}$$

which provides $a = 5.71$ and $b = 0.31$.

Assuming we have access to the expected number of cases in area $i = 1, ..., 75$, denoted E_i, the count in each area is generated from a Poisson distribution such that

$$Y_i^*|\theta_i \overset{ind}{\sim} Poisson(E_i\theta_i). \tag{2.8}$$

Five fixed censoring criteria are considered. They differ from each other due to the proportion of censored areas. We consider a scenario without censored areas (Criterion 1), the Criterion 2 that consider almost 50% of the areas as being censored and the three censoring criteria proposed by Oliveira and Loschi (2013) (denoted by Criterion 3, 4 and 5). These censoring criteria are summarized below with their respective proportion of censored areas given in parentheses.

Criterion 1 : no area is censored (0%)

Criterion 2 : $\gamma_i = 1$ for $i = 1, ..., 7, 16, ..., 22, 31, ..., 37, 46, ..., 52, 61, ..., 67$ (47%)

Criterion 3 : $\gamma_i = 1$ if $HDI_i \leq HDI_{15\%}$ (16%)

Criterion 4 : $\gamma_i = 1$ if $AI_i \leq 20.0$ (23%)

Criterion 5 : $\gamma_i = 1$ if $FI_i \leq FI_{50\%}$ and $IdC_i \leq IdC_{50\%}$ (36%), where HDI_i represents the Human Development Index in 2000 for the regions of MG, AI_i is the Adequacy Index proposed by França et al. (2006) which measures the quality of mortality information in MG, FI_i denotes the proportion of Functional Illiteracy in each region of the State, IdC_i is the proportion of Infant Deaths with Ill-defined Cause in these regions and $X_{\alpha\%}$ denotes the α-th percentile of the quantity X.

In this study we only consider the censoring level $\delta = 0.7$, that is, the counts y_i in censored areas are given by $\lceil y_i^* \times 0.7 \rceil$, where y_i^* is the value generated from (2.8). To model the simulated datasets we fit the Censored Poisson model as described in the following Cases 1, 2 and 3. In all cases. The censoring indicator vectors used to generate the data are considered in the modeling, that is, the correct censoring criteria used to generate the datasets are considered in the fitted models.

Case 1: Spatially Structured Random Effect Model

The CPM considered here assumes that $\theta_i = \exp\{s_i\}$, where s_i represents a spatially structured random effect, such that

$$\begin{aligned} Y_i|\gamma_i, \mu_i &\overset{ind}{\sim} CP(\mu_i) \\ \mu_i &= E_i \exp\{s_i\} \\ \boldsymbol{s}|W, \boldsymbol{\xi} &\sim \text{CAR}(W, \boldsymbol{\xi}), \end{aligned}$$

where W and $\boldsymbol{\xi}$ are, respectively, the proximity matrix inherent to the map and the hyperparameters associated to the CAR model (Banerjee et al., 2004).

Figure 2.3 shows the comparison between the posterior means of $\boldsymbol{\theta}$ and its true values for each censoring criteria. The estimates in non-censored areas ($\circ$) tend to be quite close to their true values, for all criteria considered in the model construction. Thus, the proportion of censored neighbouring areas does not affect the estimates in areas where data have good quality. In censored regions ($+$), the posterior mean tends to overestimate the relative risk presenting some extreme values in a few regions. The estimates tend to have the same behavior independent of the proportion of censored areas. The extreme values generally occur in censored areas in which the most neighbouring areas are also censored.

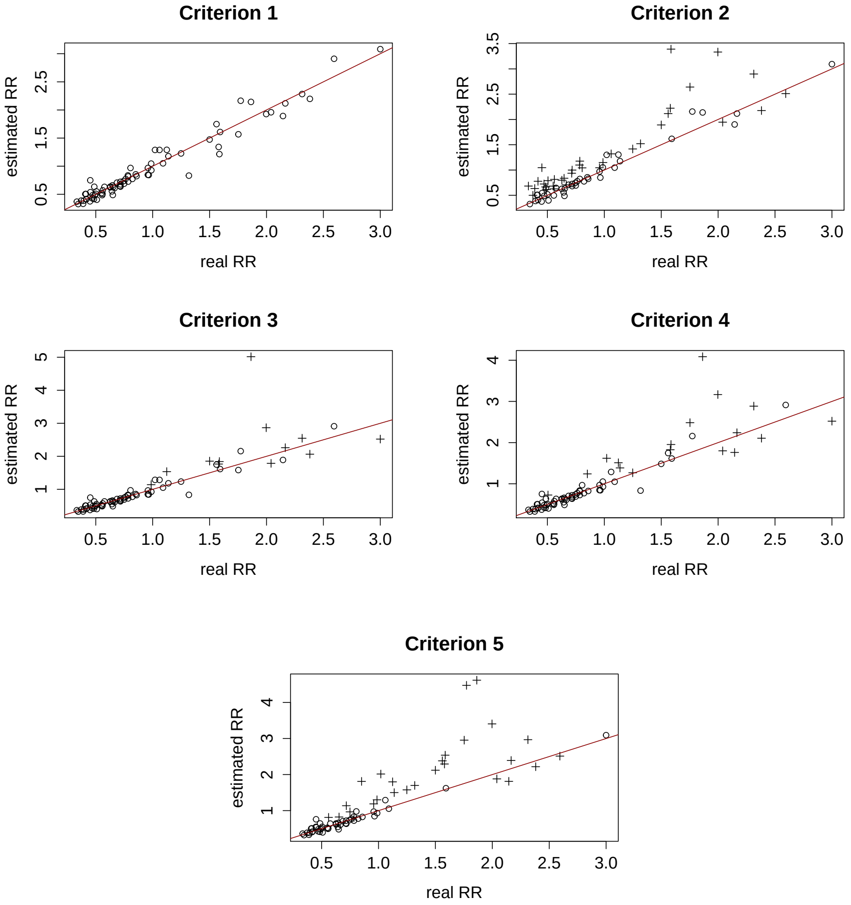

Figure 2.3: Comparison between the posterior mean of $\boldsymbol{\theta}$ and its true value in Case 1.

Case 2: Spatially non-Structured Random Effect Model

Assume now that in the CPM we model the relative risk in each area as being $\theta_i = \exp\{\upsilon_i\}$, where υ_i represents a spatially non-structured random effect, so that

$$\begin{array}{rcl} Y_i|\gamma_i, \mu_i & \overset{ind}{\sim} & CP(\mu_i) \\ \mu_i & = & E_i \exp\{\upsilon_i\} \\ \upsilon_i & \overset{iid}{\sim} & Normal(0.0, 2.0), \end{array}$$

Figure 2.4 compares the posterior mean of $\boldsymbol{\theta}$ with its true value for each censoring criteria. As in Case 1, the estimates in non-censored areas ($\circ$) are quite close to their true values for all criteria. In all censored areas ($+$) the relative risk is overestimated by the posterior mean, except for one specific area under Criterion 4. However, the model considered here seems to provide better estimates for the $\boldsymbol{\theta}$ than that model considered in

Case 1 - we notice that, in general, the overestimation is lesser than that observed in Case 1 in relation to the extreme estimates for the relative risks. This apparent superiority of the model with a non-spatial random effect (Case 2) in relation to the model with a spatial random effect (Case 1) is an unexpected result, because the data were generated assuming an increasing relative risk from the South to the North, which establishes a kind of spatial structure in the map. However, we must consider that the spatial structure assumed in Case 1 takes into account the information in neighbouring areas to estimate the risk in each area, which seems to affect the estimates, mainly in censored areas in which the most neighbouring areas are also censored - in general, a greater overestimation (extreme values) is noted for such areas.

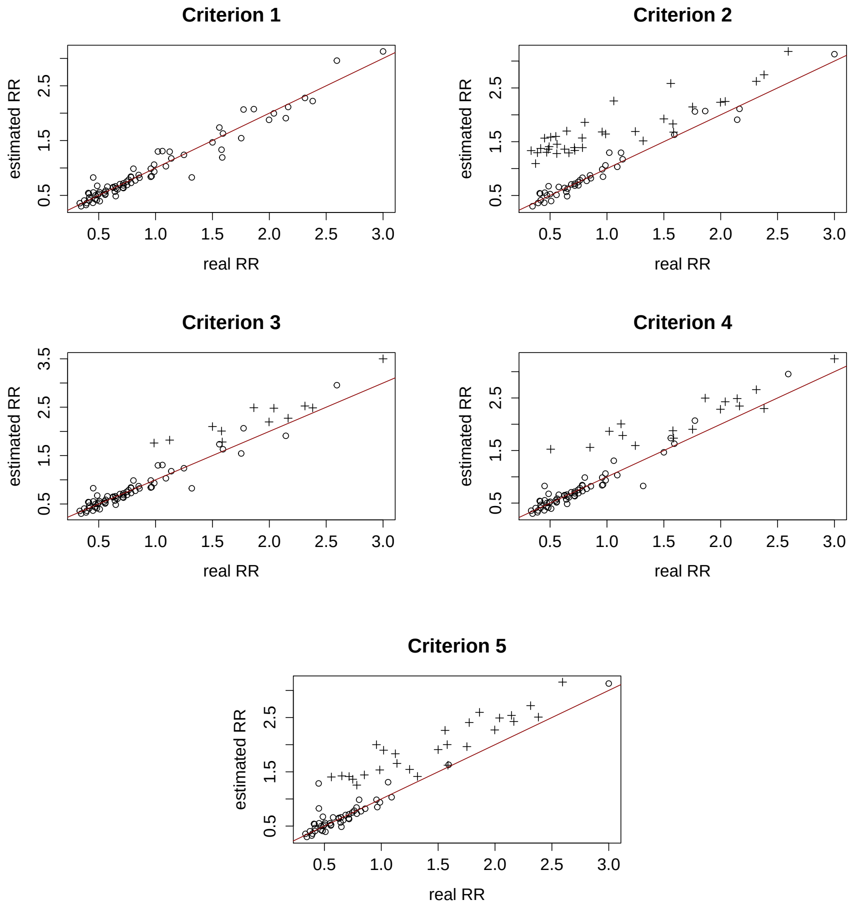

Figure 2.4: Comparison between the posterior mean of $\boldsymbol{\theta}$ and its true value in Case 2.

Case 3: Spatially Structured and non-Structured Random Effects Model

In this last case, the CPM fitted to analyze the datasets considers that $\theta_i = \exp\{\upsilon_i + s_i\}$, where υ_i and s_i represents the spatially non-structured and structured random effect, respectively, and now

$$\begin{array}{rcl} Y_i|\gamma_i, \mu_i & \overset{ind}{\sim} & CP(\mu_i) \\ \mu_i & = & E_i \exp\{\upsilon_i + s_i\} \\ \upsilon_i & \overset{iid}{\sim} & Normal(0.0, 2.0) \\ \boldsymbol{s}|W, \boldsymbol{\xi} & \sim & \text{CAR}(W, \boldsymbol{\xi}). \end{array}$$

Figure 2.5 shows the comparison between the posterior mean of $\boldsymbol{\theta}$ and its true value for each censoring criteria. The posterior estimates of RR are quite similar to those obtained in Case 2, showing that the spatially structured random affect does not play an important role in the relative risks estimation for the generated scenarios.

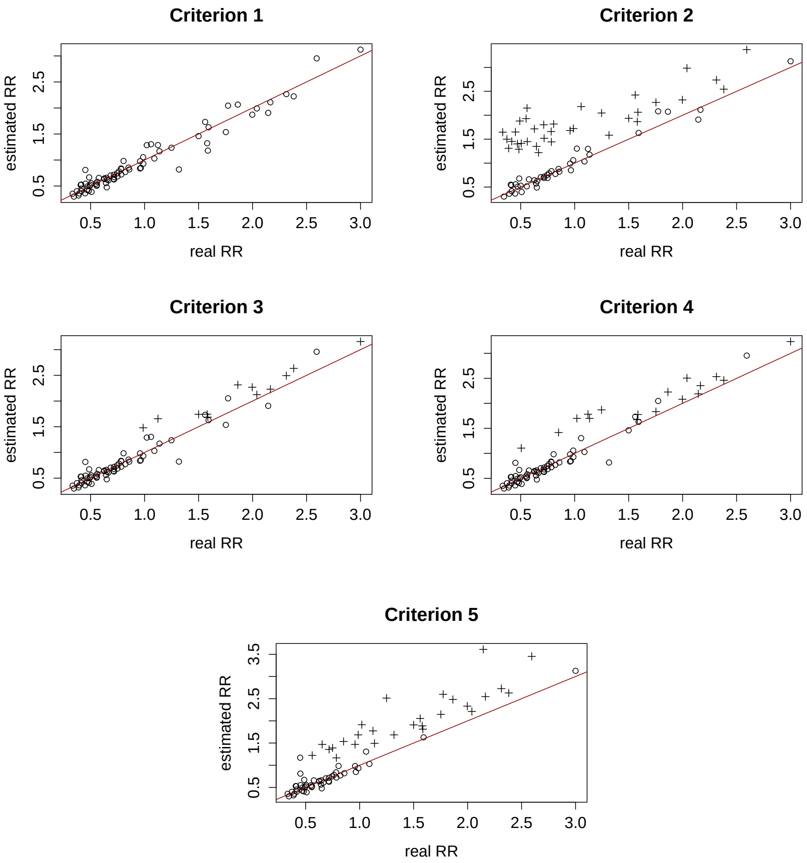

Figure 2.5: Comparison between the posterior mean of $\boldsymbol{\theta}$ and its true value in Case 3.

2.3.3 Simulation 3: The Clusterization Effect

In this section we evaluate the behavior of the relative risk estimates considering the existence of clusters in the $n = 75$ regions of MG map. Those clusters induce a partition of the map, denoted by ρ. Assume the existence of five independent clusters $\mathcal{C}_j,\ j = 1, ..., 5$, such that $\rho = \{\mathcal{C}_1 = (1, ..., 15), \mathcal{C}_2 = (16, ..., 30), \mathcal{C}_3 = (31, ..., 45), \mathcal{C}_4 = (46, ..., 60), \mathcal{C}_5 = (61, ..., 75)\}$. Also assume that such a partition induces the existence of $\boldsymbol{\theta}_\rho = (\theta_{\mathcal{C}_1}, \theta_{\mathcal{C}_2}, \theta_{\mathcal{C}_3}, \theta_{\mathcal{C}_4}, \theta_{\mathcal{C}_5})$ such that the counts Y_i in areas belonging to a given cluster $\mathcal{C}_j$ are independent with $\theta_i \overset{d}{=} \theta_k$ for all $(i, k) \in \mathcal{C}_j$. That is, the counts Y_i in areas belonging to cluster $\mathcal{C}_j$ are independent and identically distributed.

As before, assume that the expected number of cases in area i, E_i, is available. Thus, the count y_i^* in each area is generated from a Poisson distribution given by

$$Y_i^*|\theta_i \overset{ind}{\sim} Poisson(E_i\theta_i),$$

where

$$\theta_i = \begin{cases} 7.4 \text{ for all } i \in \mathcal{C}_1 \\ 2.7 \text{ for all } i \in \mathcal{C}_2 \\ 1.0 \text{ for all } i \in \mathcal{C}_3 \\ 0.6 \text{ for all } i \in \mathcal{C}_4 \\ 0.4 \text{ for all } i \in \mathcal{C}_5 \end{cases}$$

Denote the true value of θ_i used to generate the data by θ_i^T. We only consider the censoring level $\delta = 0.7$, which means that the counts y_i in censored areas are given by $\lceil y_i^* \times 0.7 \rceil$ (see Section 2.3.1). The censoring mechanism used to generate the datasets is fixed and corresponds to the five censoring criteria presented in Section 2.3.2.

To analyze the simulated datasets, we fit the following model

$$\begin{aligned} Y_i|\gamma_i, \theta_i &\overset{ind}{\sim} CP(E_i\theta_i) \\ \theta_i &\overset{ind}{\sim} Gamma(\alpha_i, \phi_i), \end{aligned}$$

The study is divided into four schemes. The difference among the schemes is due to the Gamma prior distribution assumed to model the uncertainty about the relative risks $\boldsymbol{\theta}$ as well as by the fact we are or not informing the correct partition ρ to the model. In all schemes, the correct censoring criteria used to generate the datasets are informed to the model. These four schemes and the associated results are presented in the following.

Scheme 1: Using Informative Priors and Informing the Correct Partition

In this first scheme, to built the prior distribution of θ_i we choose α_i and ϕ_i so that *a priori* $E[\theta_i] = \theta_i^T$ and $Var[\theta_i] = 100$ for all i. Moreover, the exact partition $\rho = \{\mathcal{C}_1 = (1, ..., 15), \mathcal{C}_2 = (16, ..., 30), \mathcal{C}_3 = (31, ..., 45), \mathcal{C}_4 = (46, ..., 60), \mathcal{C}_5 = (61, ..., 75)\}$ is reported to the model.

Comparisons between the posterior mean of relative risks $\boldsymbol{\theta}$ and their true values for each censoring criteria are shown in Figure 2.6. In all criteria, for both non-censored ($\circ$) and censored ($+$) areas, the estimates are exactly equal to their true values or quite close to them. Such a result indicates that we can obtain optimal posterior estimates for the relative risks $\boldsymbol{\theta}$ if we provide a good prior information for it as well as a good information about the partition structure inherent to the map.

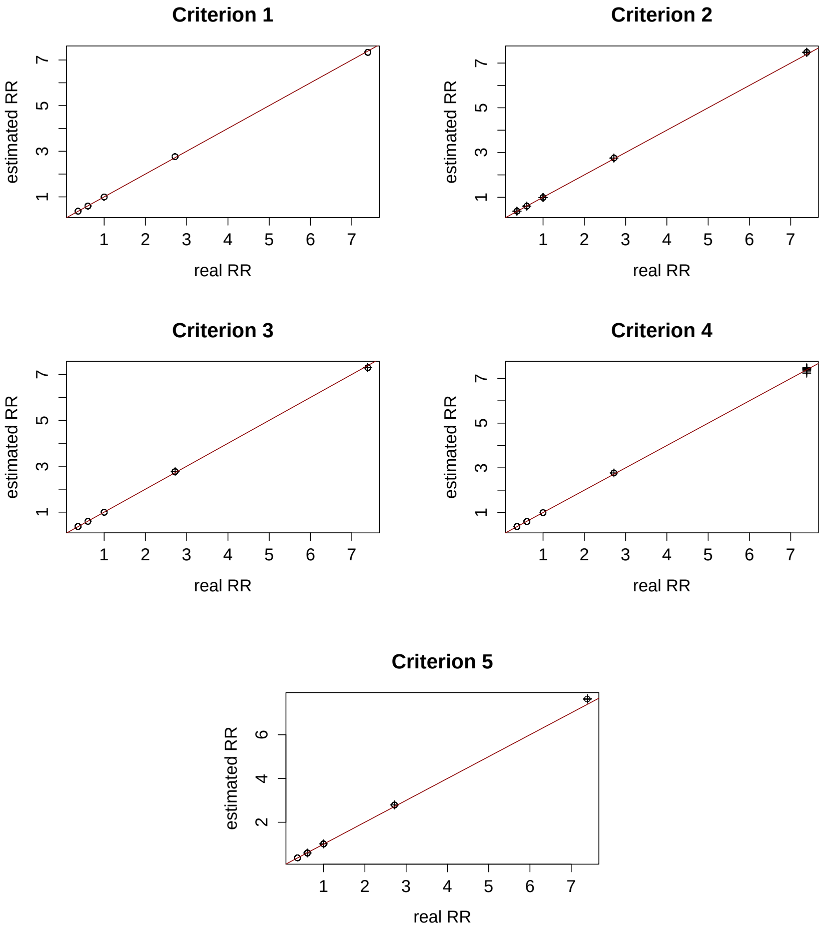

Figure 2.6: Comparing the posterior mean of the $\boldsymbol{\theta}$ and its true value in Scheme 1.

Scheme 2: Using Informative Priors and not-Informing the Correct Partition

As in Scheme 1, here we built the prior distribution of θ_i choosing α_i and ϕ_i so that *a priori* $E[\theta_i] = \theta_i^T$ and $Var[\theta_i] = 100$ for all i. However, instead of reporting the correct partition to the model, we report $\rho = \{\mathcal{C}_1 = (1), \mathcal{C}_2 = (2), \mathcal{C}_3 = (3), ..., \mathcal{C}_{75} = (75)\}$, i.e., the model will estimate the parameters θ_i by treating each area as a single cluster.

The posterior mean of the relative risks $\boldsymbol{\theta}$ are compared to their true values in Figure 2.7 for each censoring criteria. In general, for both non-censored ($\circ$) and censored ($+$) areas, the posterior estimates are close to their true values although not so close as seen in Scheme 1. Anyway, we have some evidence that, if we provide a good prior information for all θ_i, the relative risks will be well estimated even when the correct partition of the map is not identified.

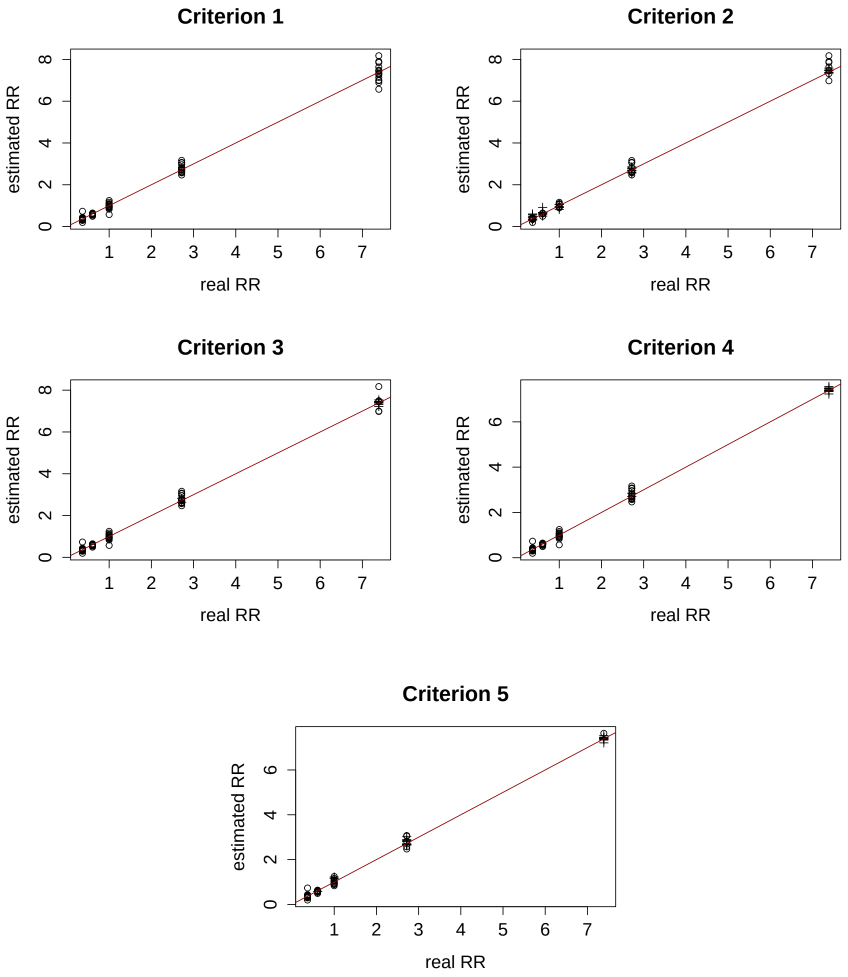

Figure 2.7: Comparing the posterior mean of the $\boldsymbol{\theta}$ and its true value in Scheme 2.

Scheme 3: Using non-Informative Priors and Informing the Correct Partition

To built the prior distribution of θ_i in this scheme, we choose α_i and ϕ_i so that *a priori* $E[\theta_i] = 5.0$ and $Var[\theta_i] = 100$ for all i. Moreover, assume that the correct partition $\rho = \{\mathcal{C}_1 = (1, ..., 15), \mathcal{C}_2 = (16, ..., 30), \mathcal{C}_3 = (31, ..., 45), \mathcal{C}_4 = (46, ..., 60), \mathcal{C}_5 = (61, ..., 75)\}$ is reported to the model.

Figure 2.8 displays the comparison between the posterior mean of $\boldsymbol{\theta}$ and its true value for each censoring criteria. Except under Criterion 4, for both non-censored ($\circ$) and censored ($+$) areas, in most areas the estimates are equal to the true values or quite close to them, as observed in Scheme 1.

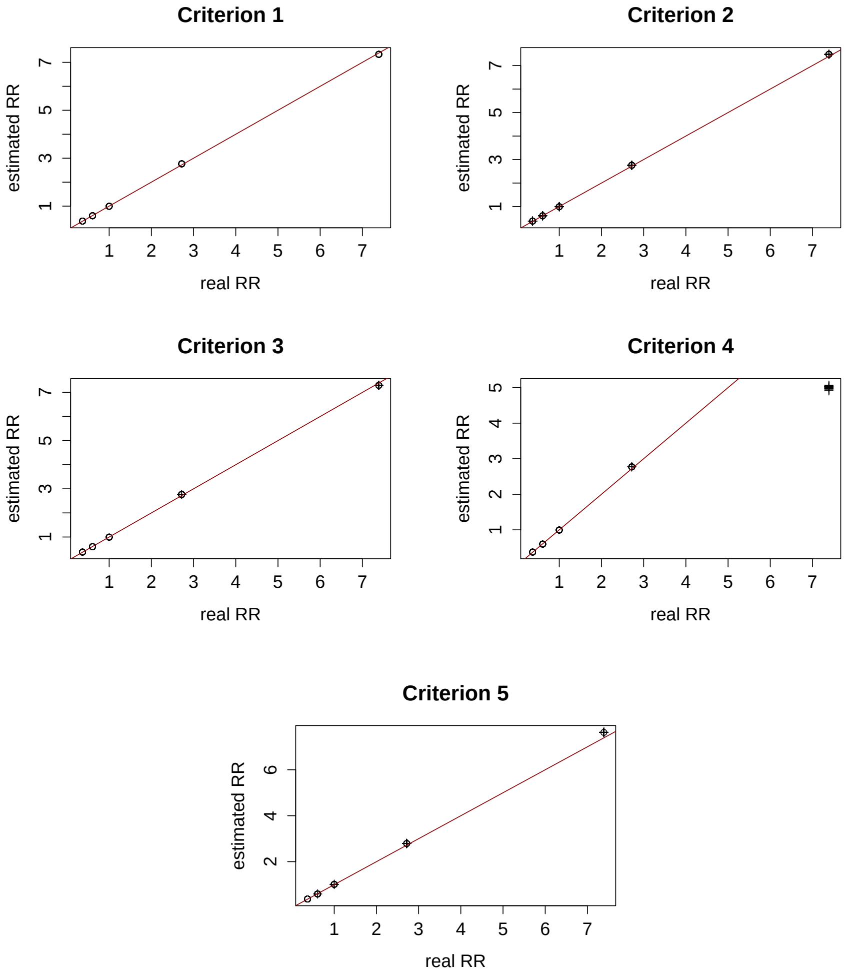

Figure 2.8: Comparing the posterior mean of the $\boldsymbol{\theta}$and its true value in Scheme 3.

Specifically under Criterion 4, all areas belonging to cluster $\mathcal{C}_1$ are censored. The true relative risk for the areas belonging to cluster $\mathcal{C}_1$ is 7.39, but their posterior mean tends to 5.0, which is the prior mean. All other clusters have at least one non-censored area under Criterion 4 and the posterior mean for the RR in areas belonging to these clusters remaining quite close to its true value, as observed in all other censoring criteria.

Therefore, there is an evidence that the prior distribution has a strong influence on the posterior estimates of the relative risk in clusters where 100% of the areas are censored, even if the partition inherent to the map is correctly identified. The influence of the prior choice is minimized if the cluster contains at least one non-censored area and the correct partition is reported to the model.

Scheme 4: Using non-Informative Priors and not-Informing the Correct Partition

In this last Scheme, we perform a simulation study in which α_i and ϕ_i are chosen so that *a priori* $E[\theta_i] = 5.0$ and $Var[\theta_i] = 100$ for all i and, at the same time, we report to the model that $\rho = \{\mathcal{C}_1 = (1), \mathcal{C}_2 = (2), \mathcal{C}_3 = (3), ..., \mathcal{C}_{75} = (75)\}$, i.e., the model will estimate $\boldsymbol{\theta}$ by treating each area as a single cluster. Thus, we do not provide to the model a good prior information for the relative risks neither the correct partition inherent to the map.

Results are shown in Figure 2.9 for each censoring criteria. The estimates in non-censored areas ($\circ$) are close to their true values, as observed in Scheme 2. However, in censored areas ($+$) the posterior estimates for the risks are dominated by the prior distribution tending to the prior mean.

Since the censored areas are treated as single clusters, there is no surprise in that conclusion because the prior information tends to be dominant on the posterior inference if few data information is available. We, therefore, have some evidence that, if we do not have a good prior information about the relative risks neither a good information about the partition structure inherent to the map, the relative risks $\boldsymbol{\theta}$ will tend to not be well estimated, mainly in censored areas.

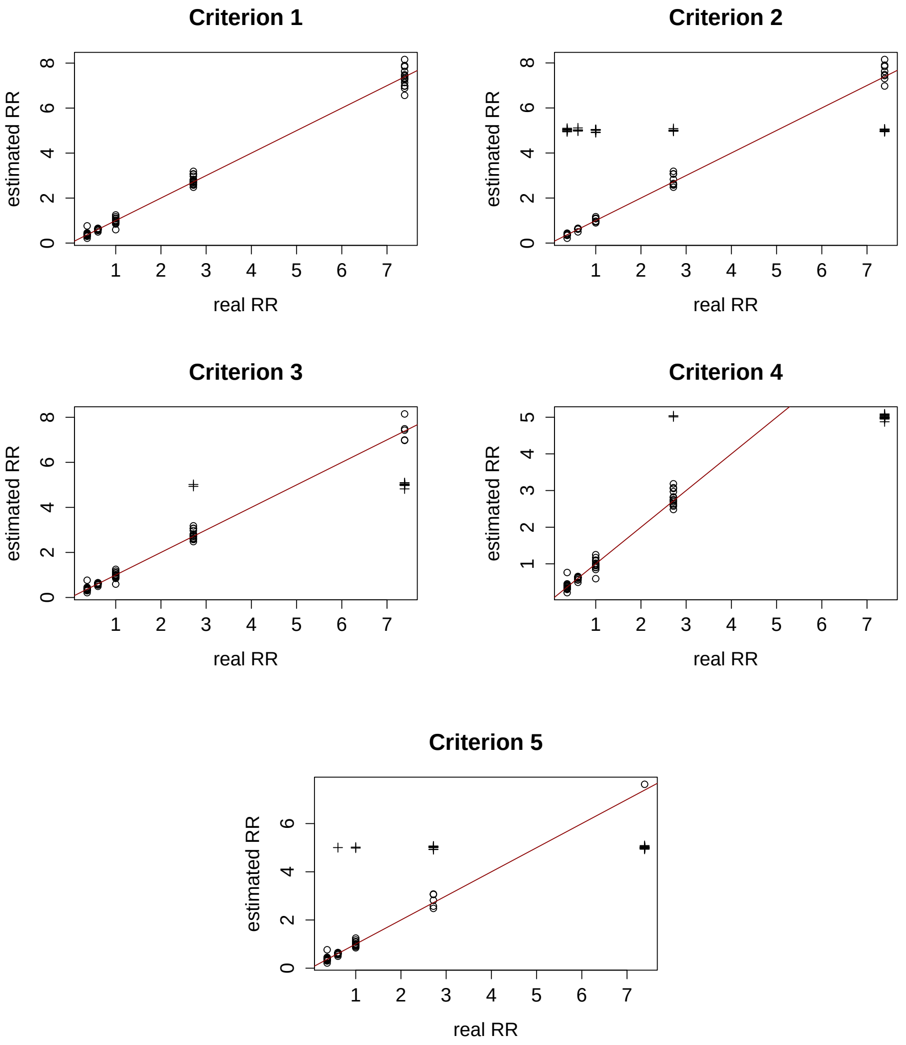

Figure 2.9: Comparing the posterior mean of the $\boldsymbol{\theta}$ and its true value in Scheme 4.

2.3.4 Conclusions on the Simulation Studies

In summary, considering the simulation studies in which artificial data were fitted by using different specifications of the Censored Poisson model (CPM) proposed in Bailey et al. (2005), we conclude that choosing an adequate prior distribution for the relative risks is truly important for obtaining good posterior inference, mainly in censored regions. In this sense, information provided by experts on the area of interest it is of great importance and non-informative prior must be avoided, unless we really do not have any prior information. We also noted that the influence of the prior distribution chosen for the relative risks is greater in datasets with greater censoring levels.

For datasets generated in Simulation 2, we notice that the spatially structured random effect does not play an important role in the estimation of the relative risks when compared with a spatially non-structured random affect, independently of the proportion of censored areas.

In general, if there exist a clustering of areas in the map, optimal posterior estimates for the risks are achieved if a good prior information is provided for them and the correct information about the partition structure inherent to the map is informed to the model. If we provide a good prior information for the relative risks, they are well estimated even when the correct partition of the map is not identified. There is an evidence that the prior distribution has a strong influence on the posterior estimates in clusters where 100% of the areas are censored, even if the partition inherent to the map is correctly identified. The influence of the prior choice is minimized if the clusters contain at least one non-censored area and the correct partition is reported to the model. At last, we notice that, if we do not have a good prior information about the relative risks neither a good information about the partition structure inherent to the map, the relative risks will tend to not be well estimated, mainly in censored areas.

The CPM used in these simulation studies requires the pre-establishment of the censored areas, but it is not trivial how to precisely define these areas in problems involving real data, i.e., data that are not simulated. Therefore, to account for potential underreporting in real data problems, it would be more appropriate the specification of a joint model for the data generating and the data reporting processes. Next chapter, we introduce an extension of the Censored Poisson model by treating the censoring mechanism as random.

Chapter 3

Random Censoring Poisson model

In poor and socially deprived areas, economic, social and health data are typically underreported. As a consequence, inference using the observed counts will be biased and the risks will be underestimated. As mentioned before (Section 2.2), to overcome this problem, Bailey et al. (2005) proposes to consider data from the suspected areas as censored information and develop a Bayesian spatial approach for the called Censored Poisson model (CPM) in Famoye and Wang (2004). A limitation of the CPM is that the censored areas must be precisely known a priori, which is not a simple task in many practical situations. Therefore, to account for potential underreporting, it would be more appropriated to jointly model the behavior of the observed data and the data reporting process.

This chapter presents the main contributions of this work. We propose an extension on the CPM by introducing a random censoring mechanism on it, as opposed to requiring a prior specification of the censored areas. We call the proposed model by Random Censoring Poisson model (RCPM). In such a model, the relative risks and the probabilities of underreporting are both estimated. Basically, we introduce a latent random variable in the modeling for such a purpose, which receives the same status of a parameter in the Bayesian approach that is being considered. That is, we introduce a CPM in which the censoring mechanism is treated as random. The joint distribution of all quantities involved in the proposed model as well as their full conditional distributions are provided in this chapter. To efficiently sample from the posterior, we also introduce a posterior sampling scheme which relies on the data augmentation technique.

This chapter is organized as follows. In Section 3.1 we provide the details on the theoretical specification of the proposed model. In Section 3.2 we discuss about the MCMC scheme needed to perform posterior inference on the proposed model. Section 3.3 presents the data augmentation technique proposed to facilitate the posterior sampling process. Performance of the proposed model is illustrated considering simulated scenarios in Section 3.4 as well as the ENM dataset in Chapter 4.

3.1 Model Specification

Suppose a map formed by n regions. Assume that Y_i is the count for the event of interest in region i, which occurs with rate μ_i for $i = 1, ..., n$, so that

$$Y_i|\mu_i \overset{ind}{\sim} \text{Poisson}(\mu_i). \tag{3.1}$$

As in Famoye and Wang (2004), we assume that some observable variables Y_i are not completely observed and thus they are considered as being censored. For our purpose, let γ_i be a latent random variable given by

$$\gamma_i \;=\; \begin{cases} 1, & \text{if area } i \text{ is censored,} \\ 0, & \text{otherwise,} \end{cases} \tag{3.2}$$

and assume that the probability of the region i be a censored one (underreported) is $P(\gamma_i = 1) = p_i$, $p_i \in (0, 1)$.

Assuming independence between the counts Y_i, given γ_i and μ_i, for $i = 1, ..., n$; and also assuming that the censoring mechanism is independent of the number of events in each area, the likelihood function associated to observed counts $\boldsymbol{y} = (y_1, ..., y_n)$ is

$$\begin{aligned} L(\boldsymbol{\mu}; \boldsymbol{Y}, \boldsymbol{\gamma}) &= \prod_{i=1}^{n} \Big\{ \big[f_{Y_i|\mu_i}(y_i)\big]^{1-\gamma_i} \; \big[1 - F_{Y_i|\mu_i}(y_i - 1)\big]^{\gamma_i} \Big\} \\ &= \prod_{i=1}^{n} \left\{ \left(\frac{e^{\mu_i} \mu_i^{y_i}}{y_i!} \right)^{1-\gamma_i} \left(\sum_{y \geq y_i} \frac{e^{\mu_i} \mu_i^{y}}{y!} \right)^{\gamma_i} \right\}. \end{aligned} \tag{3.3}$$

where $f_{Y_i|\mu_i}$ and $F_{Y_i|\mu_i}$ denote, respectively, the probability function (pf) and cumulative probability function (cpf) of a random variable with distribution Poisson(μ_i).

Suppose that for each area is available a set of covariates $\boldsymbol{X}_i = (X_{i1}, ..., X_{ik})$ related to, for instance, the socioeconomic/educational level or access to health services, which provide information on suspect regions of underreporting. Such covariates might be used to appropriately model the uncertainty about the underreporting probabilities $\boldsymbol{p} = (p_1, ..., p_n)$. *A priori*, it is expected that regions with the worst social deprivation indicators have p_i close to 1.0 whereas regions with the best ones have p_i close to 0.0. To describe such a behavior, an appropriate scaling/ordering can be chosen for each

$\boldsymbol{X}_j,\ j = 1, ..., k$, such that the underreporting probability p_i can be modeled using a logit regression model given by

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{logit}(p_i) = \lambda_0 + \boldsymbol{\beta}\boldsymbol{X}_i, \tag{3.4}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)$ and the prior distributions of each $\beta_j,\ j = 1, ..., k$ must put positive probability mass in the appropriate part of the real axis $\mathbb{R}$ (positive or negative part) depending on the ordering of the associated covariate $\boldsymbol{X}_j$. If the highest values for the social deprivation indicator $\boldsymbol{X}_j$ are associated to the worst regions, then β_j must have domain on the positive real numbers, $\mathbb{R}^+$. Similarly, if the highest values for the social deprivation indicator $\boldsymbol{X}_j$ are associated to the best regions, then β_j must have domain on the negative real numbers, $\mathbb{R}^-$. By doing that, we ensure the desired relationship between $\boldsymbol{p}$ and $\boldsymbol{X}_j$, which should disclose that the highest values for the underreporting probabilities are associated to the regions with the worst social deprivation indicators.

In this context, we propose the following Bayesian hierarchical model

$$\begin{aligned}
Y_i|\gamma_i, \mu_i &\overset{ind}{\sim} CP(\mu_i) \\
\log\ \mu_i &= \log\ E_i + \log\ \theta_i \\
\theta_i &\overset{ind}{\sim} \pi_{\theta_i} \\
\gamma_i|p_i &\overset{ind}{\sim} Ber(p_i) \\
\text{logit}(p_i) &= \lambda_0 + \boldsymbol{\beta}\boldsymbol{X}_i \\
\lambda_0 &\sim \pi_{\lambda_0} \\
\boldsymbol{\beta} &\sim \pi_{\boldsymbol{\beta}},
\end{aligned}$$

where θ_i represents the true relative risk associated to area i and E_i denotes the expected number of cases in such area. As before, by notation $CP(\mu_i)$ we mean that observation Y_i has a Poisson distribution with rate μ_i and, given μ_i and γ_i, the contribution of this observation for the likelihood function corresponds to the i-th term of the function in (3.3).

Obviously, several structures can be chosen for modeling the relative risks θ_i. For example, in a simple context an appropriate Gamma prior distribution can be assigned for each θ_i. In other case, random effects may be introduced in the modeling of μ_i to account for extra-Poisson variations, so that

$$\log \mu_i = \log E_i + \upsilon_i + s_i,$$

where υ_i is a non-spatially structured random effect which usually account for the dependence among the counts $\boldsymbol{Y}$ induced by unmeasured covariates and s_i represents a spatially structured random effect which account for an explicit spatial dependence among the counts $\boldsymbol{Y}$. Details on this modeling strategy were discussed in Section 2.1.1.

Also, the relative risk θ_i may be modeled including a suitable linear combination of available covariates $\boldsymbol{W} = (W_1, ..., W_l)$ related to suspected risk factors measured in each area i, so that

$$\log \mu_i = \log E_i + \sum_{j=1}^{l} \omega_j \boldsymbol{W}_{ij} + \upsilon_i + s_i.$$

From the modeling point of view, both sets of covariates $\boldsymbol{X}$ and $\boldsymbol{W}$ might be equal, overlapping or one might be a subset of the other (Dvorzak and Wagner, 2015). However, when all available variables are included as regressors in both parts of the model (in the Poisson and logit parts), model identification may require additional information and it must be investigated.

The joint distribution of the complete model is given by

$$\begin{aligned}
\pi(\boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \lambda_0, \boldsymbol{\beta}) &= L(\boldsymbol{\mu}, \boldsymbol{\gamma}; \boldsymbol{Y}) \pi(\boldsymbol{\gamma}|\lambda_0, \boldsymbol{\beta}) \pi(\boldsymbol{\theta}) \pi(\lambda_0) \pi(\boldsymbol{\beta}) \\
&= \prod_{i=1}^{n} \left\{ \left[g_i(\lambda_0, \boldsymbol{\beta}) \left(1 - F_{Y_i|\mu_i}(y_i - 1)\right)\right]^{\gamma_i} \right. \qquad (3.5) \\
&\times \left. \left[(1 - g_i(\lambda_0, \boldsymbol{\beta}))\, f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i} \pi_{\theta_i} \right\} \pi_{\lambda_0} \pi_{\boldsymbol{\beta}},
\end{aligned}$$

where $f_{Y_i|\mu_i}$ and $F_{Y_i|\mu_i}$ are as defined in 3.3 and $g_i(\lambda_0, \boldsymbol{\beta}) = p_i = (1+\exp\{-(\lambda_0 + \boldsymbol{\beta} \boldsymbol{X}_i)\})^{-1}$.

3.2 Posterior Sampling Scheme

A convenient MCMC sampling scheme must be implemented for posterior inference. Such scheme corresponds to the Gibbs Sampler, which is based on the full conditional distribution (fcd) of all parameters and latent random variables involved in the proposed model. To establish notation, let $\boldsymbol{V}$ be a vector with m components and denote by $\boldsymbol{V}_{-i}$ the vector $\boldsymbol{V}$ without the i-th component, that is, $\boldsymbol{V}_{-i} = (V_1, ..., V_{i-1}, V_{i+1}, ..., V_m)$. Define $\boldsymbol{\psi} = (\boldsymbol{\theta}, \lambda_0, \boldsymbol{\beta})$. Assuming the joint distribution in (3.5), we obtain the fcd of all quantities involved in the proposed model.

For $i = 1, ..., n$, the full conditional distribution of θ_i is given by

$$\pi(\theta_i|\boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{\psi}_{-\theta_i}) \;\propto\; \left[f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i} \left[1 - F_{Y_i|\mu_i}(y_i - 1)\right]^{\gamma_i} \pi_{\theta_i}.$$

Therefore, if a Gamma(α_i, ϕ_i) prior distribution is assigned to the relative risk θ_i, its fcd assumes the following different expressions depending if area i is censored or non-censored

$$\pi(\theta_i|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{\psi}_{-\theta_i}) \;\propto\; \begin{cases} \text{Gamma}\left(y_i + \alpha_i, \frac{\phi_i}{E_i\phi_i + 1}\right), \text{ if } \gamma_i = 0, \\ \text{Gamma}\left(\alpha_i, \phi_i\right) \times \left[1 - F_{Y_i|\mu_i}(y_i - 1)\right], \text{ if } \gamma_i = 1. \end{cases} \tag{3.6}$$

For the i-th component of the latent censoring vector $\boldsymbol{\gamma}$ the fcd has closed form and it is given by

$$\begin{aligned} \pi(\gamma_i|\boldsymbol{Y}, \boldsymbol{\gamma}_{-i}, \boldsymbol{\psi}) &\propto L(Y_i|\theta_i, \gamma_i)\pi(\gamma_i|\lambda_0, \boldsymbol{\beta}) \qquad (3.7) \\ &\propto \left\{g_i(\lambda_0, \boldsymbol{\beta})\left[1 - F_{Y_i|\mu_i}(y_i - 1)\right]\right\}^{\gamma_i} \left\{\left[1 - g_i(\lambda_0, \boldsymbol{\beta})\right] f_{Y_i|\mu_i}(y_i)\right\}^{1-\gamma_i} \end{aligned}$$

therefore, $\gamma_i|\boldsymbol{Y}, \boldsymbol{\gamma}_{-i}, \boldsymbol{\psi} \sim Ber\left(\frac{A_i}{A_i+B_i}\right)$, where $A_i = g_i(\lambda_0, \boldsymbol{\beta})\left[1 - F_{Y_i|\mu_i}(y_i - 1)\right]$ and $B_i = \left[1 - g_i(\lambda_0, \boldsymbol{\beta})\right] f_{Y_i|\mu_i}(y_i)$.

The full conditional distribution for $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{\psi}_{-\boldsymbol{\beta}}) \propto \prod_{i=1}^{n} \pi(\gamma_i|\lambda_0, \boldsymbol{\beta})\pi_{\boldsymbol{\beta}}, \tag{3.8}$$

and the fcd of the parameter λ_0 it is similar to (3.8), replacing $\pi_{\boldsymbol{\beta}}$ for π_{λ_0}.

Note from (3.5) that the joint distribution of the proposed model involves a cumulative probability function for the censored (underreported) areas and this makes it difficult the posterior sampling process of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. Particularly, for the posterior sampling of $\boldsymbol{\theta}$ one Metropolis-Hastings step will be needed. However, a worse problem arises in the posterior sampling process of $\boldsymbol{\gamma}$. Note that the fcd of γ_i in (3.7) involves a direct comparison between the terms A_i and B_i. The fact is that term A_i is always much higher than the term B_i, providing a ratio $\frac{A_i}{A_i+B_i}$ with value always next to 1.0 and, thereby, we generate more censored areas than we should. That problem happens because the term A_i involves a cumulative probability function, whereas the term B_i involves the probability in a single point.

In order to overcome that issue, we consider the data augmentation technique presented in Section 3.3. Basically, to simplify the structure of the likelihood function and the posterior inference, the observed data in censored areas are replaced for augmented values, which are generated from an appropriated truncated distribution. As an advantage, this data augmentation step provides a complete joint distribution for the augmented model that do not involves a cumulative probability function and, therefore, the posterior sampling is facilitated.

3.3 Data Augmentation for Posterior Sampling

The basic idea of the data augmentation technique proposed by Tanner and Wong (1987) is simple. According to those authors, suppose the observed data $\boldsymbol{Y}$ is augmented by the latent quantity $\boldsymbol{Z}$. If $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are both known, then the problem is straightforward to analyze, i.e., the augmented posterior distribution $\pi(\phi|\boldsymbol{Y},\boldsymbol{Z})$ can be calculated in a simplest form. However, the posterior distribution of interest is $\pi(\phi|\boldsymbol{Y})$, which may be difficult to calculate directly. If, however, we can generate multiple values of $\boldsymbol{Z}$ from the predictive distribution $\pi(\boldsymbol{Z}|\boldsymbol{Y})$ (that is, multiple imputations of $\boldsymbol{Z}$), then $\pi(\phi|\boldsymbol{Y})$ can be approximately obtained as the average of $\pi(\phi|\boldsymbol{Y},\boldsymbol{Z})$ over the imputed $\boldsymbol{Z}$'s. Nevertheless, $\pi(\boldsymbol{Z}|\boldsymbol{Y})$ depends, in turn, on $\pi(\phi|\boldsymbol{Y})$. Hence, if $\pi(\phi|\boldsymbol{Y})$ was known, it could be used to calculate $\pi(\boldsymbol{Z}|\boldsymbol{Y})$. That mutual dependence between $\pi(\phi|\boldsymbol{Y})$ and $\pi(\boldsymbol{Z}|\boldsymbol{Y})$ leads to an iterative algorithm to calculate the desired posterior distribution $\pi(\phi|\boldsymbol{Y})$. In practice, to implement the algorithm, we must be able to sample from two distributions: $\pi(\phi|\boldsymbol{Y},\boldsymbol{Z})$ and $\pi(\boldsymbol{Z}|\boldsymbol{Y},\phi)$.

Chib (1992) combines the data augmentation idea (Tanner and Wong, 1987) and the Gibbs sampler (Gelfand and Smith, 1990) to built an elegant solution for the censored data problem in the context of the well-known Tobit model, in which the censures are fixed. The essential idea is simple and we now extend it for our underreported data problem, in which the censoring mechanism is treat as random.

Consider that a sample $\boldsymbol{Y}=(y_1,...,y_n)$ of size n is available in which n_c observations are censored (underreported) and $n_o=(n-c_c)$ observations are non-censored (correctly observed). Denote by $\boldsymbol{y}^c$ and $\boldsymbol{y}^o$ the set of censored and non-censored observations, respectively. Suppose that along with the censored observations, $\boldsymbol{y}^c$, we have available the corresponding latent data $\boldsymbol{Z}$, which it is a vector of dimension $n_c\times 1$. Although $\boldsymbol{Z}$ it is not observed, a method that is based on simulating $\boldsymbol{Z}$ is available.

Let $\mathcal{C}$ denote the set of indexes 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 Truncated Poisson distribution with rate μ_i and support $[y_i,y_i+1,y_i+2,...)$, whose probability function is given by

$$\pi(Z_i=z_i|\boldsymbol{Y},\boldsymbol{\psi},\gamma_i=1)=\frac{f_{Z_i|\mu_i}(z_i)}{1-F_{Z_i|\mu_i}(y_i-1)},\ z_i=y_i,y_i+1,..., \tag{3.9}$$

where $f_{Z_i|\mu_i}$ and $F_{Z_i|\mu_i}$ are, respectively, the probability function and cumulative probability function of a random variable Z_i with distribution Poisson with rate is μ_i.

Therefore, 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 $\boldsymbol{y}_i^c$ replaced by z_i, for all $i \in \mathcal{C}$, that is

$$y_i^z|\gamma_i \;=\; \begin{cases} y_i, & \text{if} \quad \gamma_i = 0, \\ z_i, & \text{if} \quad \gamma_i = 1, \end{cases} \tag{3.10}$$

where $z_i \geq y_i$ is generated from (3.9). Consequently, $L(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\psi}, \boldsymbol{\gamma}) = L(\boldsymbol{y}^o, \boldsymbol{y}^c, \boldsymbol{Z}|\boldsymbol{\psi}, \boldsymbol{\gamma}) = L(\boldsymbol{y}^o, \boldsymbol{Z}|\boldsymbol{\psi}, \boldsymbol{\gamma}) = L(\boldsymbol{Y}^z|\boldsymbol{\psi}, \boldsymbol{\gamma})$, i.e., the data-augmented likelihood do not depends on a cumulative probability function.

The most important point is that the conditional probability function of the latent data $\pi(z_i|\boldsymbol{Y}, \boldsymbol{\psi}, \gamma_i)$ is available in a tractable form and the data-augmented posterior distribution $\pi(\boldsymbol{\psi}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{Z})$ has a more simple form than $\pi(\boldsymbol{\psi}, \boldsymbol{\gamma}|\boldsymbol{Y})$. Both $\pi(z_i|\boldsymbol{Y}, \boldsymbol{\psi}, \gamma_i)$ and $\pi(\boldsymbol{\psi}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{Z})$ are the inputs for the Gibbs Sampler algorithm and enable us to recursively simulate the desired posterior distribution of $\boldsymbol{\psi}$ and $\boldsymbol{\gamma}$, $\pi(\boldsymbol{\psi}, \boldsymbol{\gamma}|\boldsymbol{Y})$. Chib (1992) proves that, when the data augmentation technique is used, the posterior inference for parameters of interest remains the same as in the initial model.

The complete joint distribution under the data-augmented model is given by

$$\begin{aligned} \pi(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \lambda_0, \boldsymbol{\beta}) \;&=\; L(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\gamma})\pi(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\gamma})\pi(\boldsymbol{\gamma}|\lambda_0, \boldsymbol{\beta})\pi(\boldsymbol{\theta})\pi(\lambda_0)\pi(\boldsymbol{\beta}) \\ &=\; \prod_{i=1}^{n} \Big\{ \left[1 - F_{Y_i|\mu_i}(y_i - 1)\right]^{\gamma_i} \left[f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i} \left[\frac{f_{Z_i|\mu_i}(z_i)}{1 - F_{Z_i|\mu_i}(y_i - 1)}\right]^{\gamma_i} \\ &\times\; \left[g_i(\lambda_0, \boldsymbol{\beta})\right]^{\gamma_i} \left[1 - g_i(\lambda_0, \boldsymbol{\beta})\right]^{1-\gamma_i} \pi_{\theta_i} \Big\} \, \pi_{\lambda_0} \pi_{\boldsymbol{\beta}} \\ &=\; \prod_{i=1}^{n} \Big\{ \left[g_i(\lambda_0, \boldsymbol{\beta}) f_{Z_i|\mu_i}(z_i)\right]^{\gamma_i} \left[(1 - g_i(\lambda_0, \boldsymbol{\beta}))\, f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i} \pi_{\theta_i} \Big\} \\ &\times\; \pi_{\lambda_0} \pi_{\boldsymbol{\beta}}, \end{aligned} \tag{3.11}$$

where $f_{Y_i|\mu_i}$, $f_{Z_i|\mu_i}$, μ_i and $g_i(\lambda_0, \boldsymbol{\beta})$ are as defined previously in (3.5). Note that the joint distribution in (3.11) does not depend on a cumulative probability function.

Posterior inference depends on the full conditional distribution (fcd) of $\boldsymbol{\theta}$, $\boldsymbol{Z}$, $\boldsymbol{\gamma}$, λ_0 and $\boldsymbol{\beta}$. In the following, we provide the fcd for all these quantities. The posterior fcd of $\boldsymbol{Z}$, $\pi(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\psi}, \boldsymbol{\gamma})$, is already given in (3.9) with $\boldsymbol{\psi} = (\boldsymbol{\theta}, \lambda_0, \boldsymbol{\beta})$.

For $i = 1, ..., n$, the full conditional distribution of θ_i is given by

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

Therefore, in this case if a Gamma(α_i, ϕ_i) prior distribution is assigned to the relative risk θ_i, its fcd has closed form even in censored areas, so that

$$\pi(\theta_i|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\psi}_{-\theta_i}) \;\propto\; \begin{cases} \text{Gamma}\left(y_i+\alpha_i, \frac{\phi_i}{E_i\phi_i+1}\right), & \text{if } \;\gamma_i=0, \\ \text{Gamma}\left(z_i+\alpha_i, \frac{\phi_i}{E_i\phi_i+1}\right), & \text{if } \;\gamma_i=1. \end{cases} \tag{3.12}$$

For the i-th component of the latent vector $\boldsymbol{\gamma}$, under the data-augmented model the fcd becomes

$$\begin{aligned} \pi(\gamma_i|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\gamma}_{-i},\boldsymbol{\psi}) \;&\propto\; \left[f_{Z_i|\mu_i}(z_i)\right]^{\gamma_i}\left[f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i}\pi(\gamma_i|\lambda_0,\boldsymbol{\beta}) \\ &\propto\; \left[g_i(\lambda_0,\boldsymbol{\beta})f_{Z_i|\mu_i}(z_i)\right]^{\gamma_i}\left\{\left[1-g_i(\lambda_0,\boldsymbol{\beta})\right]f_{Y_i|\mu_i}(y_i)\right\}^{1-\gamma_i}, \end{aligned} \tag{3.13}$$

and, therefore, $\gamma_i|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\gamma}_{-i},\boldsymbol{\psi} \sim Ber\left(\frac{A_i^*}{A_i^*+B_i^*}\right)$, where $A_i^*=g_i(\lambda_0,\boldsymbol{\beta})f_{Z_i|\mu_i}(z_i)$ and $B_i^*=\left[1-g_i(\lambda_0,\boldsymbol{\beta})\right]f_{Y_i|\mu_i}(y_i)$. Note that now the ratio $\frac{A_i^*}{A_i^*+B_i^*}$ does not involves a cumulative probability function as in (3.7). Because of this, a more efficient scheme is obtained to sampling from the posterior distribution of the censoring indicator parameter.

The full conditional distributions for parameters $\boldsymbol{\beta}$ and λ_0 under the data-augmented model remains the same ones presented in Section 3.2.

A subsequent sampling strategy that can be considered is to use a Gibbs Sampler algorithm to sequentially sample from the full conditional distributions given in (3.8), (3.9), (3.12) and (3.13) and, then, obtaining an approximate sample of the desired posterior distribution $\pi(\boldsymbol{\psi},\boldsymbol{\gamma}|\boldsymbol{Y})$.

However, we note that convergence it is not achieved if $\boldsymbol{\gamma}$, $\boldsymbol{Z}$, λ_0 and $\boldsymbol{\beta}$ are sampled individually because of the strong dependence between them, especially the dependence between $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Therefore, it is needed to jointly sample those quantities. For such a purpose, we need to specify the joint full conditional distribution of (γ, $\boldsymbol{Z}$, λ_0, $\boldsymbol{\beta}$), which is given by

$$\pi(\boldsymbol{\gamma},\boldsymbol{Z},\lambda_0,\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{\theta}) \propto \prod_{i=1}^{n}\left\{\left[f_{Z_i|\mu_i}(z_i)\right]^{\gamma_i}\left[f_{Y_i|\mu_i}(y_i)\right]^{1-\gamma_i}\pi(\gamma_i|\lambda_0,\boldsymbol{\beta})\right\}\pi_{\lambda_0}\pi_{\boldsymbol{\beta}}, \tag{3.14}$$

and a Metropolis-Hastings (M-H) step is thus needed to sample from (3.14).

Given the data augmentation strategy presented in this section, we suggest the following MCMC scheme for posterior sampling:

1. Sample $(\gamma,\boldsymbol{Z},\lambda_0,\boldsymbol{\beta})$ from $\pi(\boldsymbol{\gamma},\boldsymbol{Z},\lambda_0,\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{\theta})$ in (3.14) using a M-H step;

2. Sample θ_i from $\pi(\theta_i|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\psi}_{-\theta_i})$ in (3.12), for all $i=1,...,n$.

Next section presents a simulation study involving the proposed Random Censoring

Poisson model and considering the data augmentation strategy presented here. A Monte Carlo study is performed in order to compare the estimates provided by the RCPM with the ones obtained using the CPM, in which the censoring mechanism must be previously determined.

3.4 Simulation Study

This section present a simulation study that enables us to evaluate the performance of the model proposed in Section 3.1. The goal is to compare the estimates for the relative risks provided by the proposed RCPM and three different specifications for the CPM from Bailey et al. (2005). The study is based on the $n = 75$ regions that comprise the map of Minas Gerais State (MG), which is the region considered in the early neonatal mortality mapping presented in Chapter 4.

We consider four different scenarios to generate the datasets. In all scenarios we assume that the expected number of cases E_i is known, for $i = 1, ..., 75$. In order to perform a Monte Carlo study, for each scenario we consider $R = 100$ datasets (replications) generated from the associated Poisson distribution and, to introduce underreporting in the data, we consider the censoring level $\delta = 0.7$ (see Section 2.3.1).

Each dataset is analyzed considering the RCPM and the CPM with three different specifications. To obtain the posterior estimates, for each dataset we use the MCMC method running a chain of 85000 iterations and discarding the first 20000 iterations as the burn-in period. To avoid a strong correlation among the generated samples, we consider a lag of length 13 obtaining a sample of size 5000 in all cases. The computational programs to obtain the estimates are implemented in R 3.1.3 language.

The different scenarios used for generating the datasets, the models considered to fit all these datasets and the evaluation metrics used to compared such models are presented in the following.

3.4.1 Data Generation

The four scenarios considered to generate the datasets used in this simulation study are presented in this section.

1. **Scenario I:** Similarly to what is considered in Section 2.3.2, in this first scenario we assume an increasing relative risk from the South to the North of Minas Gerais State, i.e., we assume that the risk increases as the latitude increases. Consider the latitudes inherent to the $n = 75$ regions of the map, denoted by Lat. Denote by θ_i

the relative risk in area i, for $i = 1, ..., 75$, such that

$$\theta_i = \exp\{5.71 + 0.31 Lat_i\}. \tag{3.15}$$

By doing this, we assume that the region with the smallest latitude has $\theta = 0.3$ and the region with the greatest latitude has $\theta = 3.0$.

Each dataset is generated from a Poisson distribution, so that

$$Y_i^* | \theta_i \overset{ind}{\sim} \text{Poisson}(E_i \theta_i), \tag{3.16}$$

and censoring is introduced in such datasets considering the following criterion proposed in Oliveira and Loschi (2013)

$$\gamma_i = \begin{cases} 1, & \text{if } AI_i \leq 20.0, \\ 0, & \text{otherwise}, \end{cases} \tag{3.17}$$

where AI is the Adequacy Index proposed by França et al. (2006). The AI measures the quality of mortality information in each region of Minas Gerais State. The criterion in (3.17) defines 23% of the regions as being censored.

2. **Scenario II:** This second scenario differs from **Scenario I** due to the criterion chosen for censoring the generated data. Here, we consider another censoring criteria proposed in Oliveira and Loschi (2013) so that

$$\gamma_i = \begin{cases} 1, & \text{if } HDI_i \leq HDI_{15\%}, \\ 0, & \text{otherwise}. \end{cases} \tag{3.18}$$

where HDI represents the Human Development Index for the regions of MG map in 2000 and $HDI_{15\%}$ denotes the 15-th percentile of the quantity HDI. Assuming this criterion, 16% of the areas are censored.

3. **Scenario III:** In this case, we also assume that the relative risks $\boldsymbol{\theta}$ increase as the latitude increases but, instead of using the expression in (3.15), we now consider that regions having similar latitudes will receive the same relative risk. Denote by $Lat_{\alpha\%}$ the α-th percentile of the latitudes inherent to the $n = 75$ regions of the map. Assume there exist five groups of regions in the map, denoted by $\boldsymbol{G} = (\mathcal{G}_1, ..., \mathcal{G}_5)$,

so that

$$\theta_i = \begin{cases} 0.3 \text{ for all } i \in \mathcal{G}_1 = \{i : Lat_i \leq Lat_{13\%}\}, \\ 0.6 \text{ for all } i \in \mathcal{G}_2 = \{i : Lat_{13\%} < Lat_i \leq Lat_{33\%}\}, \\ 1.0 \text{ for all } i \in \mathcal{G}_3 = \{i : Lat_{33\%} < Lat_i \leq Lat_{67\%}\}, \\ 2.0 \text{ for all } i \in \mathcal{G}_4 = \{i : Lat_{67\%} < Lat_i \leq Lat_{87\%}\}, \\ 3.0 \text{ for all } i \in \mathcal{G}_5 = \{i : Lat_i > Lat_{87\%}\}. \end{cases}$$

By doing this, we are considering that the group of regions with the smallest latitudes have $RR = 0.3$ and the group of regions with the greatest latitudes have $RR = 3.0$.

Using that values for the relative risks $\boldsymbol{\theta}$, datasets are generated assuming that the count in area $i = 1, ..., 75$ has a Poisson distribution as given in (3.16) and the criterion in (3.17) is considered for censoring the generated data.

4. **Scenario IV:** Datasets in this last scenario differ from those ones in **Scenario III** due to the censoring criterion considered to introduce underreporting. Here, we consider the censoring criterion defined in (3.18) instead of that one presented in (3.17).

3.4.2 Data Modelling

To analyze the datasets generated as discussed in Section 3.4.1, we consider the four different models that are presented in the following.

1. **Model I:** Corresponds to the proposed Random Censoring Poisson model (Section 3.1) such that

$$\begin{aligned} Y_i|\gamma_i, \theta_i &\overset{ind}{\sim} CP(E_i\theta_i) \\ \theta_i|\alpha_{\theta_i}, \phi_{\theta_i} &\overset{ind}{\sim} Gamma(\alpha_{\theta_i}, \phi_{\theta_i}) \\ \gamma_i|p_i &\overset{ind}{\sim} Ber(p_i) \\ \text{logit}(p_i) &= \lambda_0 - \beta AI_i \\ \lambda_0 &\sim LN(-0.873, 0.6) \\ \beta &\sim LN(-2.994, 0.6), \end{aligned} \tag{3.19}$$

where θ_i represents the relative risk associated to area i, E_i and AI denote, respectively, the expected number of cases and the Adequacy Index (França et al., 2006) in area i and LN is the notation for the log-Normal distribution. The hyperparameters α_{θ_i} and ϕ_{θ_i} are chosen so that, *a priori*, $Var[\theta_i] = 1.0$ and $E[\theta_i] = \theta_i^T$, where θ_i^T is the true relative risk used for generating the datasets and $i = 1, ..., 75$.

As discussed in Section 3.1, the prior distributions for the hyperparameters λ_0 and β must be appropriately chosen depending on the covariate used in the logit function for modeling the underreporting probabilities. In our case, the Adequacy Index proposed in França et al. (2006) is used as such covariate.

As said before, the AI measures the quality of infant mortality information in each region of Minas Gerais state and it has its values ranging from -83.84 to 100.0. The highest values for the AI are associated with the best regions: the greater the AI, the better the information on infant mortality. It is expected that highest values for the underreporting probabilities $\boldsymbol{p}$ are associated to the regions with the poorest quality on infant mortality information.

To ensure such a behavior for $\boldsymbol{p}$, we include the AI term in a negative way in the logit function given in (3.4) and, at the same time, we specify that $\beta \in \mathbb{R}^+$, where $\mathbb{R}^+$ represents the set of positive real numbers. We assigned a $LN(-2.993, 0.6)$ prior distribution for the hyperparameter β so that $E[\beta] = 0.06$ and $Var[\beta] = 0.002$. For the hyperparameter λ_0 we choose a $LN(-0.873, 0.6)$ prior distribution with $E[\lambda_0] = 0.5$ and $Var[\lambda_0] = 0.11$.

Actually, the hyperparameter λ_0 may have domain in the set real numbers, , but we choose such log-Normal prior distribution for this term in order to ensure the prior desired relationship between the underreporting probabilities $\boldsymbol{p}$ and the Adequacy Index. The choice of those prior expected values and prior variances for the hyperparameters λ_0 and β ensures that the underreporting probability p for the eight regions with the worst quality on infant mortality information varies around a mean value greater than 0.87 and, at the same time, p varies around a mean value smaller than 0.01 for the regions with the ten best values of AI.

The use of informative prior distributions on the data reporting process is needed to obtain posterior inference consistent with the problem we have at hands and, moreover, the assignment of such informative prior distributions in the context of epidemiological studies is encouraged in Bernardinelli et al. (1995). Figure 3.1 show the densities of the prior distributions for λ_0 and β and also the prior relationship between AI and the underreporting probabilities $\boldsymbol{p}$ based on the prior mean of λ_0 and β.

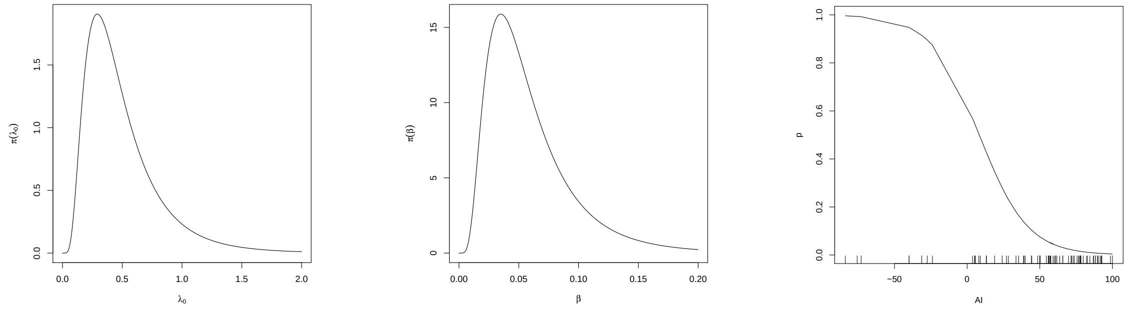

Figure 3.1: Densities of the prior distributions for λ_0 (left) and β (middle); and AI versus $\text{logit}(p_i) = E[\lambda_0] - E[\beta]AI_i$ (right).

2. **Model II:** Corresponds to the Censored Poisson model (Bailey et al., 2005) presented in Section 2.2 such that

$$\begin{aligned} Y_i|\gamma_i, \theta_i &\overset{ind}{\sim} CP(E_i\theta_i) \\ \theta_i|\alpha_{\theta_i}, \phi_{\theta_i} &\overset{ind}{\sim} Gamma(\alpha_{\theta_i}, \phi_{\theta_i}), \end{aligned}$$

   where θ_i and E_i represent, respectively, the relative risk and the expected number of cases in area i and here γ_i is a known censoring indicator. As in **Model I**, the hyperparameters α_{θ_i} and ϕ_{θ_i} are chosen so that, *a priori*, $Var[\theta_i] = 1.0$ and $E[\theta_i] = \theta_i^T$, where θ_i^T is the true relative risk used for generating the datasets and $i = 1, ..., 75$.

   By using such a model, we must previously specify the censored areas. In this case, we consider $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{75})$ as being the true censoring vector used for generating the data. By doing this, we are providing to the model a 100% correct information about the true censoring inherent to the generated data, that is, the estimation using **Model II** will be extremely favored.

3. **Model III:** In this case, we also consider the CPM described in **Model II**, but instead of providing a completely correct information about censures in the generated data, we correctly report to the model almost 50% of the true censoring indicator vector $\gamma = (\gamma_1, ..., \gamma_{75})$ used to generate these datasets.

4. **Model IV:** Also here, the CPM described in **Model II** is used for modeling the datasets, but in this case we report to the model a censoring indicator vector $\gamma = (\gamma_1, ..., \gamma_{75})$ completely different of that one used in the data generation.

3.4.3 Evaluation Metrics

We use some metrics to evaluate quality of the posterior estimates for the true relative risks $\boldsymbol{\theta}$. Denote by $\widehat{\theta}_{ij}$ a posterior estimate of the relative risk in area i for the j-th dataset (replication), where $i = 1, ..., n$ and $j = 1, ..., R$. In our study, $n = 75$ and $R = 100$. The evaluation metrics considered here are the traditional ones: Mean Squared Error (MSE), Mean Squared Percentage Error (MSPE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Bias, where

$$\begin{aligned}
MSE_i &= \frac{1}{nR}\sum_{i=1}^{n}\sum_{j=1}^{R}\left(\widehat{\theta}_{ij} - \theta_i\right)^2, \\
MSPE_i &= \frac{1}{nR}\sum_{i=1}^{n}\sum_{j=1}^{R}\left(\frac{\widehat{\theta}_{ij} - \theta_i}{\theta_i}\right)^2, \\
MAE_i &= \frac{1}{nR}\sum_{i=1}^{n}\sum_{j=1}^{R}\left|\theta_i - \widehat{\theta}_{ij}\right|, \\
MAPE_i &= \frac{1}{nR}\sum_{i=1}^{n}\sum_{j=1}^{R}\left|\frac{\theta_i - \widehat{\theta}_i}{\theta_i}\right|, \\
Bias_i &= \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{1}{R}\sum_{j=1}^{R}\widehat{\theta}_{ij}\right) - \theta_i\right].
\end{aligned}$$

All these this metrics reflect the error on estimating the parameter of interest. Thus, the closest these metrics are to zero, the better the model.

3.4.4 Results

The datasets generated for each scenario presented in Section 3.4.1 are modeled using the models described in Section 3.4.2. Results for the evaluation metrics obtained for each combination of scenario and model are shown in Table 3.1.

We note from Table 3.1 that **Model II** had the best performance in all scenarios, that is, it produces less biased estimates with the smallest variance in all cases. Such a result it is not a surprise since **Model II** receives a 100% correct information about the true censoring in the data and it is highly favored by this. On other hands, **Model IV** receives a completely wrong information about the true censoring of the data and it provides the greatest errors on the estimation in all scenarios.

When Adequacy Index is considered as a criterion for censoring the data (23% of the regions are censored), **Scenarios I** and **III**, we note that **Models I** and **III** present similar behavior with evaluation metrics very close to each other, with a little smaller values for these metrics tending to occurs for **Model III**. Therefore, for a dataset with 23% of censored areas, we have evidences that the Random Censoring Poisson model (RCPM), **Model I**, provides as good estimates as the Censored Poisson model (CPM), **Model III**, in which almost 50% of truly the censored areas are correctly identified. When considering **Scenarios II** and **IV**, in which the Human Development Index is considered for censoring the data (16% of the regions are censored), we note that RCPM, **Model I**, provides a smaller bias than CPM in which almost 50% of the true censored regions are correctly informed, **Model III** .

In relation to the structure of the relative risks $\boldsymbol{\theta}$ in the map, we note that, in general, the models provide worst estimates when the risk assumes an structure of clustering, **Scenarios III** and **IV**, if compared with the metrics obtained for scenarios without a clustering structure, **Scenarios I** and **II**.

Scenario	Model	MSE	MRSE	MAE	MAPE	Bias
I	**I**	0.040	0.032	0.129	0.137	-0.049
	II	0.008	0.020	0.059	0.094	0.001
	III	0.039	0.031	0.112	0.130	-0.049
	IV	0.077	0.055	0.175	0.171	-0.101
II	**I**	0.043	0.032	0.127	0.135	-0.026
	II	0.009	0.022	0.065	0.100	0.001
	III	0.047	0.031	0.124	0.131	-0.049
	IV	0.061	0.037	0.151	0.148	-0.068
III	**I**	0.077	0.029	0.152	0.128	-0.063
	II	0.010	0.017	0.067	0.082	0.003
	III	0.071	0.028	0.144	0.118	-0.062
	IV	0.132	0.038	0.222	0.149	-0.134
IV	**I**	0.072	0.028	0.164	0.124	-0.063
	II	0.013	0.018	0.076	0.089	0.001
	III	0.081	0.028	0.158	0.121	-0.069
	IV	0.111	0.033	0.199	0.138	-0.101

Table 3.1: Evaluation metrics for the Monte Carlo study.

Figure 3.2 present the box-plots for the relative risk estimates obtained under **Models I-IV** for the $R = 100$ datasets generate in **Scenario I**. The truly censored areas in **Scenario I** are the seventeen first regions that appear in the horizontal axis. In **Model II**, all of them are correctly identified as censored. In **Model III**, the seven first and the $17-th$ region are correctly identified as being censored and, in order to maintain the proportion of censured areas in the dataset, regions 41-49 are wrongly identified as being censored. For **Model IV**, no region is correctly identified as censored but regions 21-26, 38-47 and 75 are wrongly identified as being censored ones.

We note that **Models II, III** and **IV** provide optimal estimates for the relative risk in truly censored regions that are correctly identified as being censored. In this areas, the box-plots are centered around the true RR and they have a very small amplitude, i.e., there is a small variability in the estimation around the true relative risks.

However, if a truly censored area is identified as being a non-censored one, its RR tends to be badly estimated and with a greater variability. In general, for the truly non-censored areas, the models are comparable providing very similar estimates for the RR, even if one of these areas is wrongly identified as being a censored one - except for region 75 in **Model IV**. Box-plots for the posterior estimates of the relative risks in **Scenarios II-IV** exhibit a similar behavior to those presented in Figure 3.2 and will be omitted.

The posterior estimates of the underreporting probabilities $\boldsymbol{p}$ when considering **Model I** in **Scenario I** are presented in Figure 3.3, jointly with the prior expected probabilities (solid line). The posterior estimates are obtained using the plug-in method (red dotted line) and considering the posterior proportion of samples where the censoring indicator variable $\gamma_i = 1$ (black dotted line), for $i = 1, ..., 75$. In the plug-in method, we substitute λ_0 and β in the logit function for their posterior means. In general, we obtain a posterior estimate for $\boldsymbol{p}$ that it is close to its prior expected mean, with high deviations for the regions having the greatest values for the Adequacy Index (horizontal axis). The posterior estimate for $\boldsymbol{p}$ using the plug-in method is closer to its prior expectation than the estimate using the posterior proportion of $\gamma_i = 1$ for $i = 1, ..., 75$.

Figure 3.4 presents the box-plots for the posterior proportions of $\gamma_i = 1$ in **Scenario I**, $i = 1, ..., 75$. There is a small variability in the estimation and some extreme values can be noted in almost regions that have a positive AI.

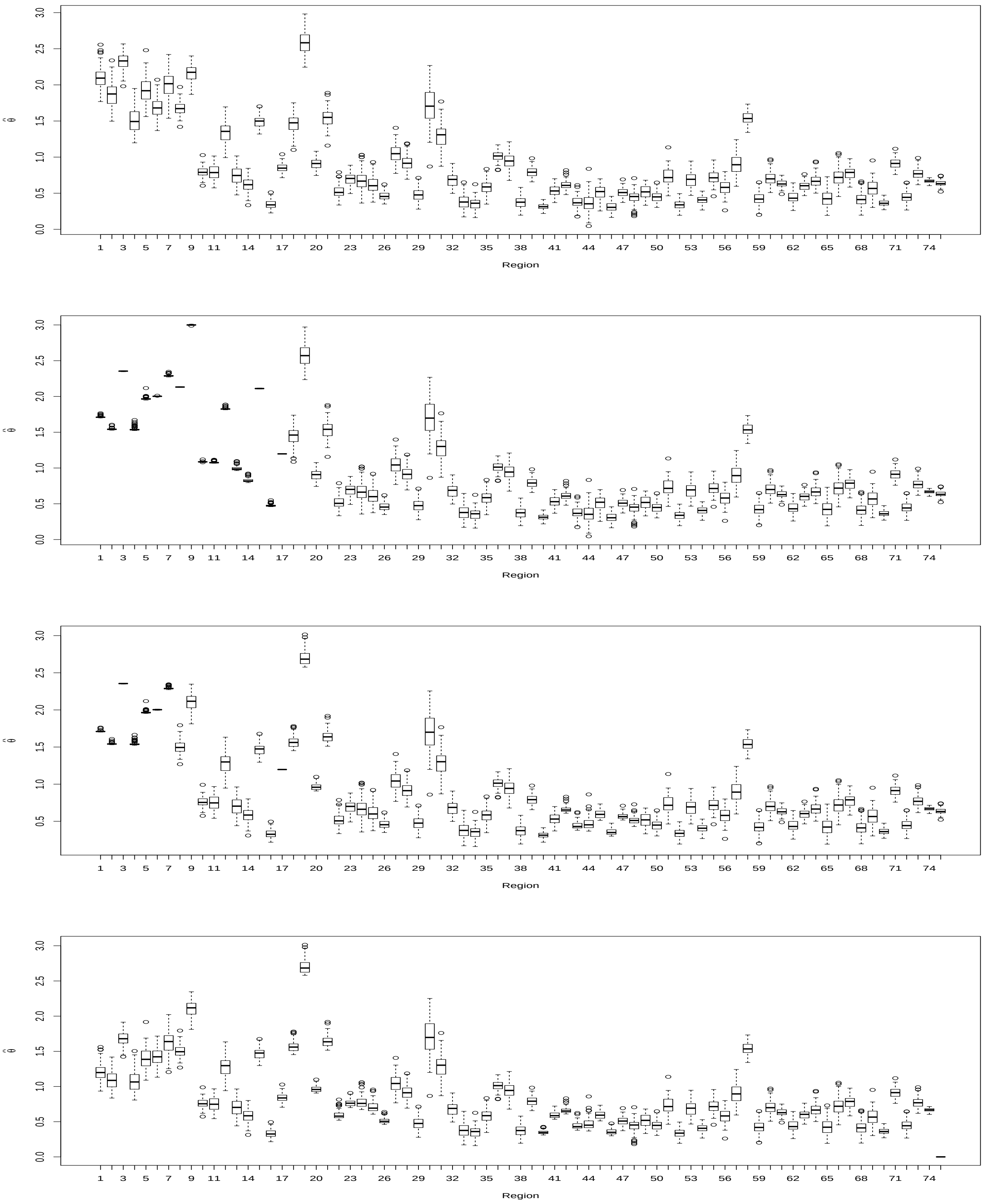

Figure 3.2: Box-plots for the posterior means of $\boldsymbol{\theta}$ in **Scenario I** using **Model I**, **II**, **III** and **IV** from the top to the bottom, respectively.

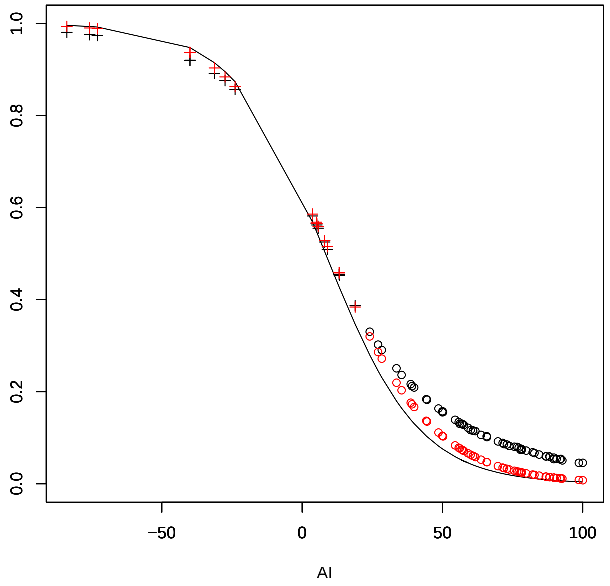

Figure 3.3: Prior expected mean for the underreporting probabilities $\boldsymbol{p}$ (solid line) and their posterior estimates in **Scenario I** using **Model I**: plug-in method (red dotted line) and posterior proportion of $\gamma_i = 1,\ i = 1, ..., 75$ (black dotted line).

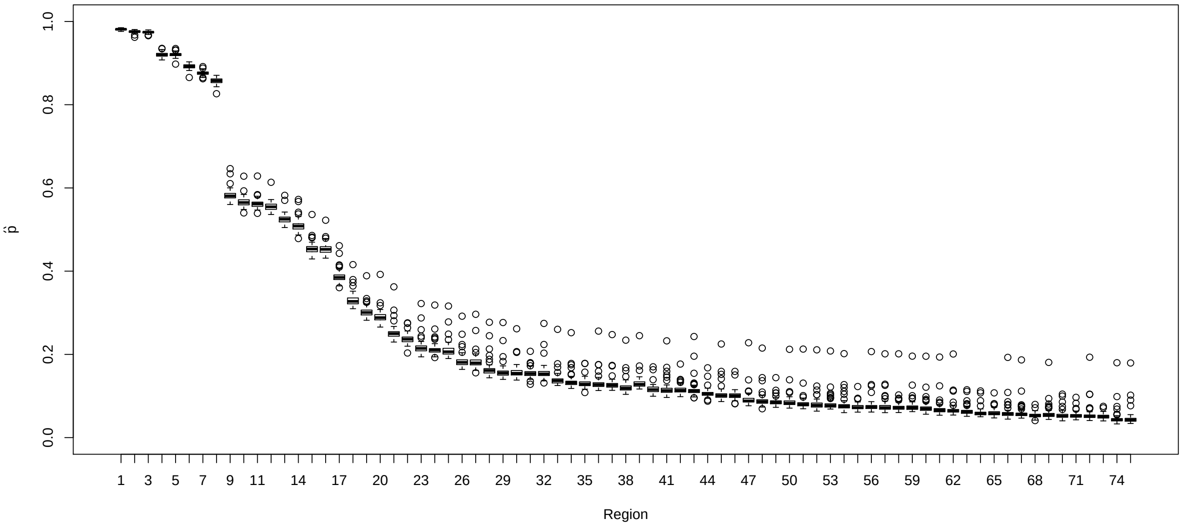

Figure 3.4: Box-plots for the posterior proportions of $\gamma_i = 1,\ i = 1, ..., 75$ in **Scenario I** using **Model I**.

3.4.5 Conclusions on the Simulation Study

In summary, we conclude that, independently of the scenario considered to generate the data, **Model II** produces less biased posterior estimates for the relative risks, which is an expected result since this model is extremely favored for receiving a completely correct information about the true censoring inherent to the data. In turn, **Model IV** receives no information about the true censoring used to generate the data and, as expected, this model produces the worst posterior estimates for the relative risks $\boldsymbol{\theta}$ in all scenarios.

For the datasets in which the proportion of censored areas is 23%, **Scenarios I** and **III**, **Models I** and **II** present a quite similar performance in the estimation of $\boldsymbol{\theta}$ because, according Table 3.1, such models produces bias and error metrics that are quite close in all cases.

When the proportion of censored data is 16%, the proposed RCPM (**Model I**) tends to provide less biased estimates for the RR than the CPM that receives an almost 50% correct information about the true censoring vectors used to generate the data (**Model III**).

In relation to the three different specifications of the CPM considered in **Scenario I** (**Models II**, **III** and **IV**), we notice from Figure 3.2 that the variability of the posterior estimate for θ_i in a truly censored area increases significantly if this area is identified as being a non-censored one. On other side, there is no significant difference in the estimate for θ_i in truly non-censored areas that are reported to the model as being censored ones.

The underreporting probabilities $\boldsymbol{p}$ are estimated with a very small variability according to the box-plots in Figure 3.4 and, in general, the posterior estimates of $\boldsymbol{p}$ are quite close to its prior expected estimates (Figure 3.3).

In next chapter, we analyze the ENM data from Minas Gerais State using the proposed model.

Chapter 4

Case Study: Mapping the ENM rate in Minas Gerais State

In this chapter we will map the relative risk of early neonatal mortality (ENM) in Minas Gerais State (MG), Brazil. The ENM is understood as the infant deaths that occurs in the first seven day of life in the period of interest. Our dataset corresponds to the number of early neonatal deaths that took place in public hospitals of the 853 municipalities of MG from 1999 to 2001. The 853 municipalities are grouped into $n = 75$ regions in order to avoid regions with very small or zero counts, which leads to unstable estimates for the mortality rates (Assunção et al., 1998).

As discussed in Chapter 1, in some regions the information about ENM in Minas Gerais State are not correctly recorded in the Hospital Information System (SIH) and the occurrence of underreporting in this dataset is quite likely.

To estimate the relative risks associated to the ENM in MG, we consider the Random Censoring Poisson model (RCPM) proposed in Chapter 3 and also the Censored Poisson model (CPM) proposed by Bailey et al. (2005). Both models account for suspect underreporting in the data. Three different fixed censoring criteria are considered for the CPM and the results are compared with that one provided by the RCPM, in which the censoring mechanism is treated as random. Section 4.1 presents some details about the models considered for fitting the ENM data and in Section 4.2 results and conclusions are provided.

4.1 Data Modelling

The Censored Poisson model (Bailey et al., 2005) treats data with suspected underreporting as censored information and the observed counts in suspect areas are considered to be lower bounds to the real number of cases. In the case study presented in this chapter, we consider the following specification for the CPM

$$\begin{aligned} Y_i|\gamma_i, \theta_i &\overset{ind}{\sim} CP(E_i\theta_i) \\ \theta_i|\alpha_{\theta_i}, \phi_{\theta_i} &\overset{ind}{\sim} Gamma(\alpha_{\theta_i}, \phi_{\theta_i}), \end{aligned}$$

where θ_i and E_i represent, respectively, the relative risk and the expected number of cases in area i, $i = 1, ..., 75$. For remembering, by notation $CP(\mu_i)$ we mean that observation Y_i has a Poisson distribution with rate μ_i and, given μ_i and γ_i, the contribution of this observation for the likelihood function corresponds to the i-th term of (3.3).

As discussed in Section 2.2, the CPM is dependent on the previous identification of all censored areas - the censoring indicator vector $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{75})$ must be known *a priori*. However, censored observations are not obviously identified in practical problems. Based on the socioeconomic level, educational variables and access to health services, Oliveira and Loschi (2013) proposed three different fixed criteria to previously determine the censored regions in their application to ENM data in Minas Gerais State and these criteria are considered here. In addition to those criteria given in (3.17) and (3.18), we also consider the criterion defined as

$$\gamma_i = \begin{cases} 1, & \text{if } FI_i \leq FI_{50\%} \text{ and } IdC_i \leq IdC_{50\%}, \\ 0, & \text{otherwise}, \end{cases} \tag{4.1}$$

where FI and IdC denote the proportions of Functional Illiteracy and Infant Deaths with Ill-defined Cause, respectively, and $X_{\alpha\%}$ denotes the α-th percentile of the quantity X.

The Random Censoring Poisson model proposed in Chapter 3 will be also considered for mapping the ENM in MG. In this novel approach, a random censoring mechanism is incorporated into the CPM as opposed to requiring a prior specification of the censored areas. We consider the specification of the RCPM given in (3.19), that is,

$$\begin{aligned} Y_i|\gamma_i, \theta_i &\overset{ind}{\sim} CP(E_i\theta_i) \\ \theta_i|\alpha_{\theta_i}, \phi_{\theta_i} &\overset{ind}{\sim} Gamma(\alpha_{\theta_i}, \phi_{\theta_i}) \\ \gamma_i|p_i &\overset{ind}{\sim} Ber(p_i) \\ \text{logit}(p_i) &= \lambda_0 - \beta AI_i \\ \lambda_0 &\sim LN(-0.873, 0.6) \\ \beta &\sim LN(-2.994, 0.6), \end{aligned}$$

where θ_i and E_i are defined as in the previous CPM model. The quantity AI denote the Adequacy Index (França et al., 2006) that it is been considered for modeling the underreporting probabilities $\boldsymbol{p} = (p_1, ..., p_{75})$.

A discussion about the choice of those log-Normal prior distributions for parameters λ_0 and β are presented in Section 3.4.2. Basically, that choice ensures the desired prior relationship between the AI and the underreporting probabilities $\boldsymbol{p}$. Figure 3.1 show the densities of the prior distributions for λ_0 and β and also the prior relationship between AI and $\boldsymbol{p}$ based on the prior mean of λ_0 and β.

For the Gamma prior distribution of the relative risk θ_i under both CPM and RCPM models, the hyperparameters α_{θ_i} and ϕ_{θ_i} are chosen so that, a priori, $Var[\theta_i] = 3.0$, for $i = 1, ..., 75$, and

$$E[\theta_i] \;=\; \begin{cases} 5.0, & \text{if } i = 1, ..., 8 \; (AI_i \leq 0.0), \\ 3.0, & \text{if } i = 9, ..., 17 \; (0.0 < AI_i \leq 20.0), \\ 1.5, & \text{if } i = 18, ..., 34 \; (20.0 < AI_i \leq 56.0), \\ 1.0, & \text{if } i = 34, ..., 75 \; (AI_i > 56.0). \end{cases}$$

The choice of those values for the prior expectation of the relative risks $\boldsymbol{\theta}$ are based on information provided by experts in the study of the ENM and it represents their knowledge about the ENM's relative risk behavior over the $n = 75$ regions of Minas Gerais State. The use of informative prior distributions for the relative risks based on information provided by experts in the area of interest in the context of epidemiological studies is discussed and encouraged in Bernardinelli et al. (1995), for instance.

Therefore, in summary, we are considering four different models for estimating the ENM's relative risk in MG, which are summarized in Table 4.1.

Table 4.1: Summary of the models used in the ENM mapping

Model Label	**Model Specification**
RCPM	RCPM as specified in (3.19)
CPM1	CPM using the criterion in (3.17)
CPM2	CPM using the criterion in (3.18)
CPM3	CPM using the criterion in (4.1)

For all models in Table 4.1 we run the MCMC for 155000 iterations, discarding the first 50000 draws as a burn-in period and considering a lag 21 to avoid correlation. Thus, in all case we consider a sample of size 5000 for making posterior inference.

4.2 Results

Figure 4.1 displays the relative risk estimates considering the posterior medians under the models CPM1 (row 1), CPM2 (row 2) and CPM3 (row 3). In this Figure, the columns present the censored regions (left) and the posterior medians of the relative risk of ENM (right).

By comparing the maps in Figure 4.1 with those ones shown in Figure 1.1, we notice that the RR estimates in non-censored regions remains essentially the same, but when underreporting is considered, Figure 4.1, the most regions in North and Northeast of MG start to present a high estimate for the risks, as it is expected by the epidemiologists. Therefore, independently on adopted censoring criterion, the CPM seems to provide better estimates for the ENM1s relative risk in those regions of MG with suspect underreporting if compared with the maximum likelihood estimates under the traditional Poisson model, which corresponds to the standardized mortality ratio (SMR) presented in (2.1).

However, the modeling using the CPM seems to be highly affected by the pre-established censoring criterion, because the estimate for the RR is always dramatically increased if a region is censored by any criterion whereas its estimate for the RR is quite similar to its SMR if this region is not censored. 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 MG map. That region is censored under CPM1 and CPM2 but it is not censored under CPM3. Under models CPM1 and CPM2, its RR is estimated between 2.5 and 5.5 while it is estimated between 0.0 and 0.5 under the model CPM3. Moreover, comparing the results obtained using different censoring criteria is quite complicated, because the censoring mechanism used by the CPM establishes which regions are censored with probability 1.0; and it is neither a simple task to determine which criterion is more likely than the others.

In order to overcome that problem of choosing a specific censoring criterion and, therefore, deciding with probability 1.0 which are the censored regions, we propose in Chapter 3 the Random Censoring Poisson model (RCPM) in which the censoring mechanism is incorporated into the modeling. Figure 4.2 displays the relative risk estimates considering the posterior medians under RCPM (right) and the posterior estimates of the underreporting probabilities $\boldsymbol{p}$ based on the posterior proportion of $\gamma_i = 1, \; i = 1, ..., 75$ (left).

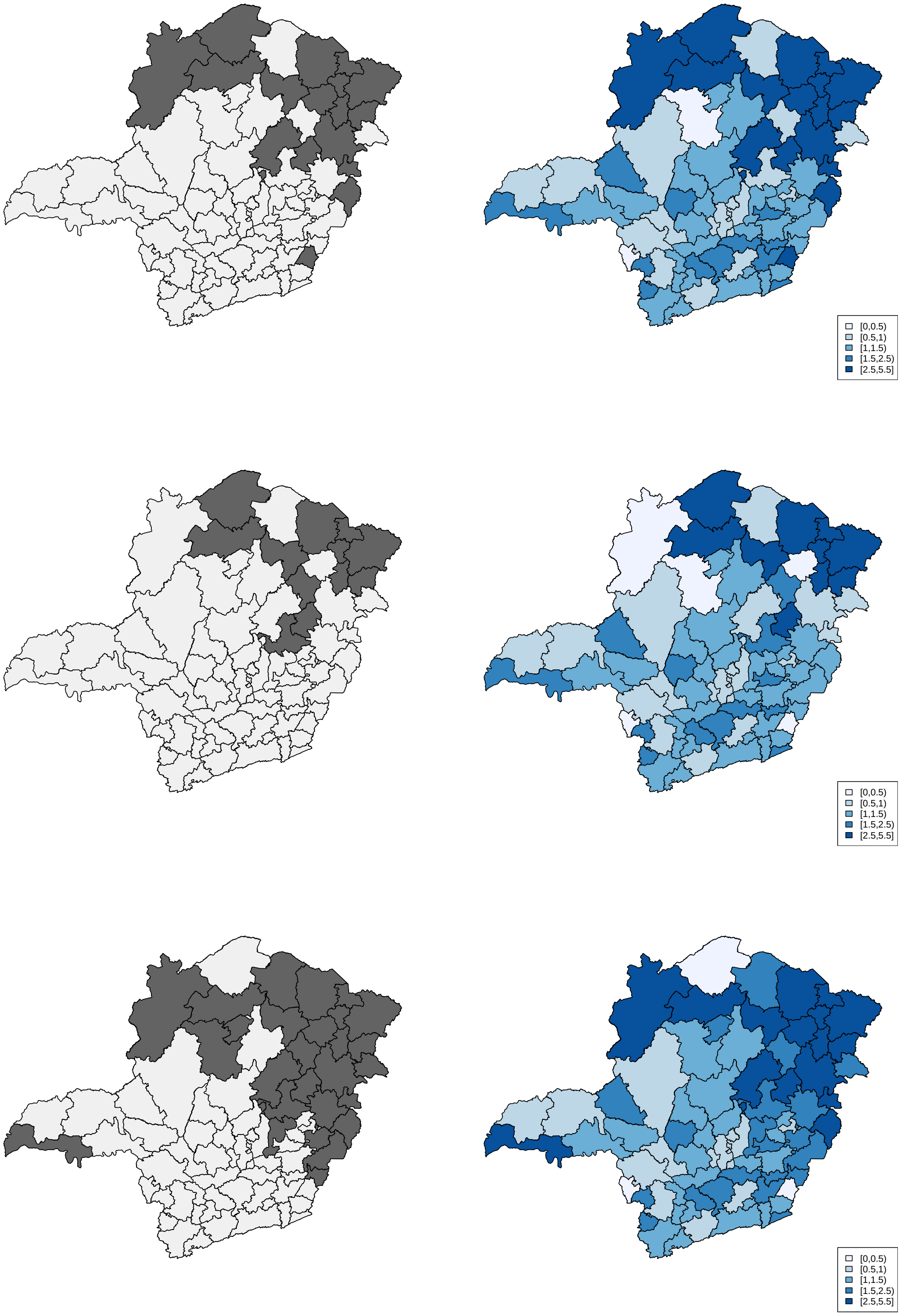

Figure 4.1: Mapping the relative risk of ENM in MG using CPM1 (row 1), CPM2 (row 2) and CPM3 (row 3). In each row: censored regions (left) and posterior medians of the relative risks (right).

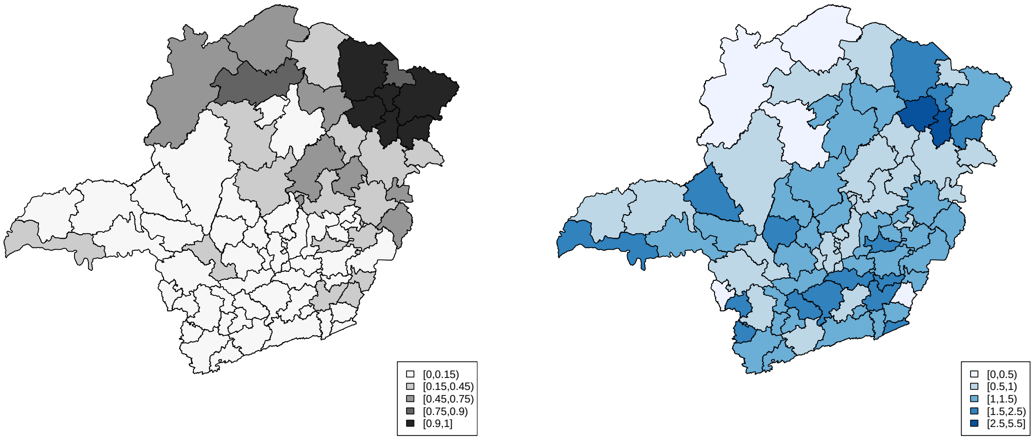

Figure 4.2: Posterior mean for the underreporting probabilities (left) and posterior medians for the relative risk of ENM in MG using RCPM (right).

By comparing the maps in Figure 4.2 with those in Figure 1.1, we notice that the RR estimates in regions having an underreporting probability smaller than 0.75 remains essentially the same. However, for the regions having an underreporting probability greater than 0.75, those regions mainly concentrated in the Northeast of MG, the relative risk estimate is considerably increased and, therefore, it gets closer to what is expected by the epidemiologists.

Comparing Figures 4.1 and 4.2 we note that in regions where the posterior estimate for the underreporting probability is low (smaller than 0.45), estimates provided by the proposed RCPM is quite similar to that obtained using the CPM in all cases. It is also noticeable that RCPM and CPM2 produced more similar maps for the relative risk of ENM in the State. However, the left map in Figure 4.2 points out that regions with the highest underreporting probabilities are concentrated in the Northeast of MG, which is more similar to the map of censored areas in CPM3.

Moreover, regarding to Figure 4.2, it can be noted that only two regions present a posterior median for the RR greater than 2.5 (in the Northeast of MG). A total of 8 regions (approximately 11%) present an estimate for the probability of being censored greater than 0.75 and 6 regions (approximately 8%) have this estimate between 0.75 and 0.45, whereas 31 regions (approximately 41%) present an estimate for the underreporting probability lower than 0.10.

Figure 4.3 show the posterior means of the ENM's relative risk in MG (middle) and the lower (left) and upper (right) limits of the 95% Highest Posterior Density interval (HPD) under all models given in Table 4.1.

In general, the posterior means of the relative risks $\boldsymbol{\theta}$ are quite close to its posterior median shown in Figures 4.1 and 4.2. Based on the HPD for models CPM1 (row 2), CPM2 (row 3) and CPM3 (row 4), we note there is more uncertainty about the RR in censored regions, since the HPD associated to these areas disclose high posterior variance. The same is observed in regions having the highest posterior estimates for the underreporting probability in the RCPM (row 1). In non-censored regions for models CPM1, CPM2 and CPM3 or in regions where estimate the posterior estimate for the underreporting probability is smaller than 0.75, the HPD discloses there is a small variability in the estimation of the ENM relative risk.

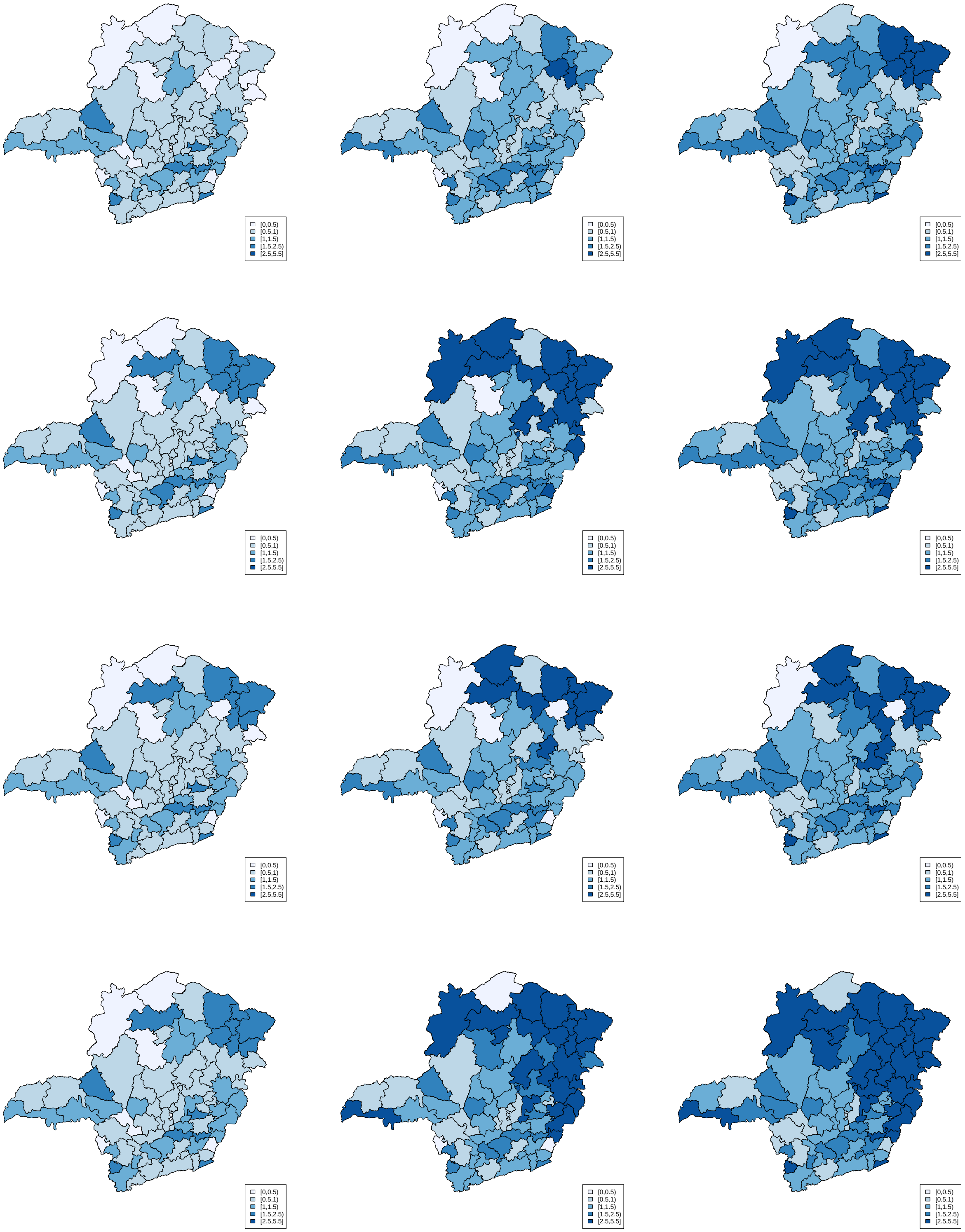

Figure 4.3: Mapping the relative risk of ENM in MG using RCPM (row 1), CPM1 (row 2), CPM2 (row 3) and CPM3 (row 4): the lower limit of the 95% HPD (column 1), the posterior mean (column 2) and the upper limit of the 95% HPD (column 3).

Chapter 5

Concluding Remarks

In this work, we were concerned with estimating and mapping the relative risk associated to a given event when available data is underreported. Particularly, we approach the problem of underreporting in the counts of early neonatal mortality (ENM) registered in public hospitals of Minas Gerais State, Brazil, between 1999 and 2001.

The ENM is a serious public health problem. Thus, the precise mapping of its relative risk is an important tool to define adequate health policies to reduce it. A big challenge for statisticians is to propose models capable to provide good estimates of the associated risks in the presence of underreported data. Assuming the information from suspect areas as censored information, Bailey et al. (2005) consider the Censored Poisson model (CPM) proposed in Famoye and Wang (2004) for handling that kind of data.

As a limitation, the CPM assumes that all censored regions are precisely known *a priori*, which is not a simple task in many practical situations. Then, we propose in this work the Random Censoring Poisson model (RCPM) as an alternative for jointly estimate the relative risk and the underreporting probability in each region of the map, instead of using a fixed vector to indicate the censured ones. Proposing the RCPM was our main contribution.

The simulation studies involving the CPM (Section 2.3) was another contribution of this work because this kind of study for the CPM in the context of underreported data is not presented in Bailey et al. (2005) neither in other papers. We obtained some interesting results from that studies and they are summarized in Section 2.3.4.

Also, in Section 3.4 some simulated data were considered to compare the estimates for the relative risks provided by the proposed RCPM and the CPM with different specifications. In summary, we conclude that quality of the estimates provided by CPM depends on the proportion of correct information about the true censoring mechanism used to generate the data that is supplied to the model. RCPM and CPM provide similar estimates depending on the number of censored regions in the generated data and the proportion of such regions that is correctly informed to the CPM. Under RCPM, the underreporting probabilities $\boldsymbol{p}$ were estimated with a very small variability and, in general, the posterior estimates of $\boldsymbol{p}$ are quite close to its prior expected mean. Summarized conclusions are provided in Section 3.4.5.

The relative risk of ENM in Minas Gerais State was estimated in Chapter 4 using both RCPM and CPM. For the CPM we considered three different fixed censoring criteria. Independently on adopted censoring criterion, the CPM seems to provide better estimates in regions with suspect underreporting if compared with the maximum likelihood estimates under the standard Poisson model (Section 2.1). However, the estimation using the CPM seems to be highly affected by the pre-established censoring criterion, because the estimate for the risk is always dramatically increased when a region is censored.

Regarding to the estimates provided by the RCPM, in regions having a posterior estimate for the underreporting probability greater than 0.75 the RR's estimate is increased if compared with the maximum likelihood estimates under the standard Poisson model.

In non-censored regions under CPM or in regions where the posterior estimate for the underreporting probability is low (smaller than 0.45) under RCPM, the 95% Highest Posterior Density interval discloses there is a small variability in the estimation of the ENM's relative risk.

From all studies presented in this work, we notice that estimates using CPM are highly affected by the pre-establishment of a censoring criterion and the choice of a specific criterion as well as the comparison between the estimates provided by different criteria is quite complicated. Therefore, seems to be more appropriate to model the underreporting probabilities using the proposed RCPM than guaranteeing with probability 1.0 which are the censored regions.

The censoring mechanism used in the proposed model (Chapter 3) can be easily applied in other case studies. For our particular problem involving the ENM data in Minas Gerais State, several other specifications for the logit function in (3.4) can be thought. For example, the underreporting probabilities can be modeled using other quantities besides the Adequacy Index in a way to give higher underreporting probabilities for the regions in the North of MG than those probabilities that was obtained in this application. By doing this, we may achieve higher estimates for the relative risk in the North of the State, as expected for this region of MG by the experts in mortality. Also, a sensibility analysis involving other choices for the prior distributions assigned to all parameters of the RCPM must be done.

Extensions to more general specifications of the proposed model , e.g., including spatial random effects in the linear predictor of the relative risks or considering a clustering structure between the regions in the map, are some ideas for future research. Moreover, the RCPM presented in this work can be easily extended to model count data subject to overreporting or even more general misclassification.

Although we focused on the early neonatal mortality data in Minas Gerais State, Brazil; underreporting is not an exclusive problem of this dataset. Actually, 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 (Merli (1998) and Xu et al. (2014)) and several other countries in African, Asia and Latin America and the Caribbean according to the World Health Organization (WHO, 2006). Although on a smaller scale, underreporting of mortality and disease cases is also present in more developed countries such as Japan (Campbell et al., 2011), United States of America (Gould et al., 2002) and and Norway (Alfonso et al., 2015).

Finally, we want to emphasize that the study of methods and models to appropriately handle underreported mortality and health data is quite important due to the context involved since, if underreporting occurs and it is not accounted for, inference using the observed counts will be biased and risks will be underestimated and, consequently, appropriate control and intervention policies would be affected. Therefore, the problem addressed in this work has great relevance, mainly in the practical sense.

Bibliography

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.

Andrade, S. M., Soares, D. A., Matsuo, T., Souza, R. K. T., Mathias, T., and Iwakura, M. L. H. (2006). Condições de vida e mortalidade infantil no estado do Paraná, Brasil. *Caderno de Saúde Pública*, 22(1):181–189.

Araújo, N. and Loschi, R. H. (2013). Mapeamento da mortalidade neonatal precoce em Minas Gerais: Modelagem e SSVS em modelos espacias. Monografia de conclusão do curso de Graduação em Estatística, Departamento de Estatística, Universidade Federal de Minas Gerais, 2013, Belo Horizonte, Brazil. (Unpublished).

Assunção, R., Barreto, S., Guerra, H., and Sakurai, E. (1998). Mapas de taxas epidemiológicas: Uma abordagem Bayesiana. *Caderno de Saúde Pública*, 14(4):713–723.

Bailey, T. C. (2001). Spatial statistical methods in health. *Caderno de Saúde Pública*, 17(5):1083–1098.

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

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC.

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

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

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 4(82):733–746.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Statist Math*, 43:1–59.

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research, Volume II: The Design and Analysis Of Cohort Studies*. International Agency for Research on Cancer, Lyon.

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

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.

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. *Revista Brasileira de Epidemiologia*, 2(10):223–238.

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

Clayton, D. and Bernardinelli, L. (1996). *Geographic and Environmental Epidemiology: Methods for Small Area Studies*, chapter Bayesian Methods for Mapping Disease Risk, pages 205–220. Oxford University Press.

Denison, D. G. T. and Holmes, C. C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics*, 57:143–49.

Dvorzak, M. and Wagner, H. (2015). Sparse Bayesian modelling of underreported count data. *Statistical Modelling*.

Famoye, F. and Wang, W. (2004). Censored Generalized Poisson regression model. *Computational Statistics and Data Analysis*, 46:547–560.

França, E., Abreu, D., Campos, D., and Rausch, M. C. (2006). Avaliação da qualidade da informação sobre mortalidade infantil em Minas Gerais, em 2000-2002: Utilização de uma metodologia simplificada. *Revista Médica de Minas Gerais*, 16(1 Supl 2):S29–S35.

Frias, P., Pereira, P., Andrade, C., and Szwarcwald, C. (2008). Sistema de informações sobre mortalidade: estudo de caso em municípios com precariedade dos dados. *Caderno de Saúde Pública*, 10(24):2257–2266.

Gelfand, A. E. and Smith, F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

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.

Guimarães, E. A. A., Hartz, Z. M. A., Filho, A. I. L., Meira, A. J., and Luz, Z. M. P. (2013). Avaliação da implantação do sistema de informação sobre nascidos vivos em municípios de Minas Gerais, Brasil. *Caderno de Saúde Pública*, 29(10):2105–2118.

Hegarty, A. and Barry, D. (2008). Bayesian disease mapping using product partition models. *Statistics in Medicine*, 27:3868–3893.

Holmes, C. C., Denison, D. G. T., and Mallick, b. k. (1999). Bayesian partitioning for classification and regression. Technical report, Imperial College, London.

Knorr-Held, L. and Best, N. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of Royal Statistical Society: Series A*, 164:73–85.

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F., and Bertollini, R. (1999). *Disease Mapping and Risk Assessment for Public Health.* Wiley Chichester.

Lima, C. R. A., Schramm, J. M. A., Coeli, C. M., and Silva, M. E. M. (2009). Revisão das dimensões de qualidade de dados e métodos aplicados na avaliação dos sistemas de informação em saúde. *Caderno de Saúde Pública*, 25(10):2095–2109.

Machado, E., Alfradique, M., and Monteiro, L. (2006). Caracterização da rede hospitalar do Sistema Único de Saúde em Minas Gerais. *Fundação João Pinheiro de Belo Horizonte.*

Merli, M. (1998). Underreporting of births and infant deaths in rural China: Evidence from field research in one county of Northern China. *The China Quarterly*, 155:637–655.

Mollié, A. (1995). *Markov Chain Monte Carlo in Practice*, chapter Bayesian Mapping of Disease, pages 359–379. Chapman & Hall.

MS-Brasil (2004). Saúde Brasil - 2004: Uma análise da desigualdade em saúde. Technical report, Ministério da Saúde. Secretaria de Vigilância à Saúde. Departamento de Análise de Situação em Saúde.

Oliveira, G. L. and Loschi, R. H. (2013). Mapeamento da mortalidade neonatal precoce em Minas Gerais: O uso da censura para contornar o problema do sub-registro. Monografia de conclusão do curso de Graduação em Estatística, Departamento de Estatística, Universidade Federal de Minas Gerais, 2013, Belo Horizonte, Brazil. (Unpublished).

Ortiz, L. P. (2000). Metodologia de cálculo da taxa de mortalidade infantil no Brasil. Technical report, RIPSA - Rede Interagencial de Informações Para a Saúde. Relatório final do grupo de trabalho ad hoc relacionado à reunião do CTI NATALIDADE E MORTALIDADE realizada na Faculdade de Saúde Pública da USP/SP, nos dias 8 e 9 de maio de 2000.

Schramm, J. and Szwarcwald, C. (2000a). Diferenciais nas taxas de mortalidade neonatal e natimortalidade hospitalares no Brasil: um estudo com base no Sistema de Informações Hospitalares do Sistema Único de Saúde (SIH/SUS). *Caderno de Saúde Pública*, 16(4):1031–1040. a.

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

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.

Teixeira, L. V., Assunção, R. M., and Loschi, R. H. (2015). A generative spatial clustering model for random data through spanning trees. In *2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, November 14-17, 2015*, pages 997–1002.

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

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.

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

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

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.