Bayesian Analysis (2020) **TBA**, Number TBA, pp. 1–32

Bias Correction in Clustered Underreported Data

Guilherme Lopes de Oliveira*, Raffaele Argiento†, Rosangela Helena Loschi‡, Renato Martins Assunção§, Fabrizio Ruggeri¶, and Márcia D'Elia Branco‖

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

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

MSC2020 subject classifications: Primary 62F15; secondary 62J12.

1 Introduction

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

*Departamento de Computação, CEFET-MG, Belo Horizonte, Brazil, guilhermeoliveira@cefetmg.br
†Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milano, Italy, raffaele.argiento@unicatt.it
‡Departamento de Estatística, UFMG, Belo Horizonte, Brazil, loschi@est.ufmg.br
§Departamento de Ciência da Computação, UFMG, Belo Horizonte, Brazil, assuncao@dcc.ufmg.br
¶CNR IMATI, Milano, Italy, fabrizio@mi.imati.cnr.it
‖Departamento de Estatística, USP, São Paulo, Brazil, mbranco@ime.usp.br

 https://doi.org/10.1214/20-BA1244

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

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

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

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

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

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

Inspired by situations in which validation datasets are unaccessible and reliable prior information about the reporting process is only available for areas experiencing the best data quality, we propose a new hierarchical Bayesian approach for the CPM (Section 2). It considers that the areas composing the region of interest are ordered according to data quality categories. If it is reasonable to additionally cluster the areas into homogeneous groups, then the model becomes more parsimonious. The clusters may be defined based on experts' opinion or applying some clustering technique to data quality indicators provided by previous studies and surveys. In our model, the data quality clustering of the areas is a tool used to model their reporting probabilities. We leverage the clusters

to create a hierarchical structure in which the reporting probabilities decrease as we move from the best data quality areas to the worst ones. The novelty in our approach is that only an informative prior distribution about the reporting probability at areas experiencing the best data quality is required to ensure identifiability (Section 2.1). To model the event occurrence rates, we consider a set of potential covariates through a regression structure. Bayesian variable selection is incorporated into the model to identify regressors with a non-zero effect.

Extensive simulation studies are presented to evaluate the performance of the proposed model in different scenarios (Section 3). We apply the developed Bayesian methodology to estimate the early neonatal mortality rates in Minas Gerais State, Brazil, for the periods 1999–2001 and 2009–2011 (Section 4), where the death counts are known to be underreported (Campos, Loschi, and França, 2007). In this context, the proposed approach is attractive because neither validation datasets nor prior knowledge about the overall mean reporting probability is available. Section 5 closes the paper with our main conclusions.

2 Model specification

Consider a region divided into A areas and denote by T_i the total number of events at area i for $i = 1, \ldots, A$. Assume that $T_i \mid \lambda_i \overset{ind}{\sim} \mathcal{P}oisson(\lambda_i)$, where λ_i is the mean expected counts in the ith area. The relative risk for the event at area i is given by $\theta_i = \lambda_i / E_i$, where E_i is a known offset quantity representing the expected number of events in such area. The offset E_i allows for a variation in the population size over the areas. In the context of underreported data, T_i is not fully observed for, at least, part of the areas, so that the reported number of events Y_i corresponds only to a fraction of T_i. To consider this data feature, each event occurring in the ith area is associated to a binary random variable $Z_t \overset{ind}{\sim} \mathcal{B}ernoulli(\epsilon_i)$, $t = 1, \ldots, T_i$, that determines whether the event will be reported or not, where $\epsilon_i \in [0, 1]$ represents the associated reporting probability. The random variables in the sequence $Z_1, Z_2, \ldots, Z_{T_i}$ are assumed as being identically distributed, mutually independent and also independent of T_i. Consequently, $Y_i = \sum_{t=1}^{T_i} Z_t$ has a compound Poisson distribution in which $Y_i \mid T_i, \epsilon_i \overset{ind}{\sim} \mathcal{B}inomial(T_i, \epsilon_i)$ and $T_i \mid \theta_i \overset{ind}{\sim} \mathcal{P}oisson(E_i\theta_i)$. Marginalizing the joint distribution of (Y_i, T_i) over T_i, it follows that the observed count Y_i has the conditional distribution

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

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

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

One of the main goals in this work is to model $\boldsymbol{\epsilon}$ in situations where no validation dataset is available to guarantee model identifiability and whenever reliable prior information about the percentage of underreporting is only available for areas where data are known to be better reported. In this context, we assume that $\epsilon_i = g(H_{1i}, \ldots, H_{mi}) = 1 - \gamma - f(H_{1i}, \ldots, H_{mi})$, where $\gamma \in [0, 1)$ is the basal underreporting probability in the area with the best data quality and f is any non-negative function such that $f(H_{1i}, \ldots, H_{mi}) < 1 - \gamma$ for all i. The function f captures the increase in the basal underreporting probability explained by the covariates. If f equals to zero in the best area, then $f(H_{1l}, \ldots, H_{ml})$ denotes the increase in the underreporting probability for area l when compared to the best one. As in the model proposed by Stoner, Economou, and Drummond (2019), covariates $H_{1i}, \ldots, H_{mi}$ are assumed to be different from $X_{1i}, \ldots, X_{pi}$ to guarantee model identifiability. This model limitation may be avoided only if validation datasets are accessible as in Dvorzak and Wagner (2015). A further discussion on this issue is given in Section 2.1.

The definition of a general f which satisfies all these constraints is not a simple task. To facilitate its construction, we assume that it is possible to sort the areas according to their data quality. Additionally, we assume that the reporting probabilities are equal for areas where the covariates related to the reporting process experience similar values. For this purpose, we assume that the A areas are grouped into K known data quality clusters, where $K \leq A$. We allow for $K = A$ to account for situations in which there is no prior information for clustering the areas. However, if such information is available and $K < A$, we obtain a more parsimonious model and more data information to estimate the reporting probabilities throughout the areas.

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

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

at area i is given by

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

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

The proposed model has some interesting features. Firstly, to be identifiable, it only requires information about the reporting probabilities in the best areas (see discussion in Section 2.1). Besides that, ϵ_i is represented in terms of interpretable parameters, which facilitates prior elicitation. For a given area i, $\mathbf{h}_i = (1, 0, \ldots, 0)^T$ and $\mathbf{h}_i = (1, 1, \ldots, 1)^T$ represent the two most extreme situations. If $\mathbf{h}_i = (1, 0, \ldots, 0)^T$ then the ith area has the highest level of data quality. We will assume that data in such area are recorded with a higher probability ($\epsilon_i = 1 - \gamma_1$) if compared to the areas in the remaining data quality levels. At the other extreme, if $\mathbf{h}_i = (1, 1, \ldots, 1)^T$ then the ith area lies in the worst data quality category. Data in this region are recorded with a lower probability ($\epsilon_i = 1 - \gamma_1 - \cdots - \gamma_K$) if compared to those areas belonging to clusters with better data quality. Thus, the parameter γ_1 represents the probability of not recording an event in areas classified in the highest level of data quality. The parameter γ_2 is the increment on such probability for areas experiencing the second highest data quality level, and so on. Another attractive feature of the proposed model is that, although the clustering indicator variable cannot be used to also model the relative risks $\boldsymbol{\theta}$, the covariates used for clustering are indirectly taken into consideration when estimating $\boldsymbol{\theta}$, since the areas belonging to the same cluster are homogeneous w.r.t. such clustering covariates.

2.1 On model identifiability

The lack of identifiability of the compound Poisson model presented in expression (2.1) has been discussed by several authors (Whittemore and Gong, 1991; Moreno and Girón, 1998; Stamey, Young, and Boese, 2006; Papadopoulos and Silva, 2012; Dvorzak and Wagner, 2015; Schmertmann and Gonzaga, 2018; Stoner, Economou, and Drummond, 2019). All these previous works impose some constraints on $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ to attain model identifiability.

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

To the best of our knowledge, there are two approaches to obtain an identifiable model when sets of covariates, say $\boldsymbol{X}$ and $\boldsymbol{H}$, are used to model the relative risk $\boldsymbol{\theta}$ and the reporting probability $\boldsymbol{\epsilon}$, respectively. The first one requires extra information from independent validation datasets (Whittemore and Gong, 1991; Stamey, Young, and Boese, 2006; Dvorzak and Wagner, 2015). This is a rare situation in practice that, however, does not require the intersection of $\boldsymbol{X}$ and $\boldsymbol{H}$ to be empty. The second one,

adopted by Papadopoulos and Silva (2012) and Stoner, Economou, and Drummond (2019), creates some kind of linear separability of the covariates sets $\boldsymbol{X}$ and $\boldsymbol{H}$. Stoner, Economou, and Drummond (2019) build $\boldsymbol{X}$ and $\boldsymbol{H}$ by splitting the set of all available covariates into two disjoint sets based on experts' opinion. Hence, there is an empty intersection between the covariates in the sets $\boldsymbol{X}$ and $\boldsymbol{H}$ but this is not enough to guarantee identifiability. In their modeling framework, they also had available an informative prior distribution for the overall mean reporting rate which was sufficient to complete the identifiability conditions. Papadopoulos and Silva (2012) allow intersection between the two sets of covariates but impose prior information to establish appropriate constraints on the parametric space, such as restrictions on the signs or exclusion of some coefficients. This avoids the need for validation datasets.

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

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

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

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

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

Proposition 2.1. *The proposed model under the specification in expression (2.3) is identifiable if β_0 or one of the coordinates of vector $\boldsymbol{\gamma}$ is fixed at a known value.*

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

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

Another way to investigate model identifiability is to consider the associated Fisher information. The Fisher information plays an important role in the asymptotic theory of maximum likelihood estimation as well as in Bayesian reference analysis. Besides that,

Rothenberg (1971) showed that a model that belongs to the exponential family is globally identifiable if the Fisher information matrix is nonsingular. Let $\Lambda(j) = \left(1 - \sum_{l=1}^{j} \gamma_l\right)$ and $\mu_{ij} = E_i \exp\{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}\}\,\Lambda(j)$. The Fisher information matrix $\mathcal{I}(\boldsymbol{\Psi})$ resulting from expression (2.3) is given by

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

The $K \times K$ sub-matrix highlighted in bold will be considered in Section 2.2 to build the Jeffreys prior for $\boldsymbol{\gamma}$ given $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$.

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

Proof. Denote by $\mathcal{C}_\kappa$ the column vector of $\mathcal{I}(\boldsymbol{\Psi})$ associated to the parameter $\kappa \in \boldsymbol{\Psi}$ such that $\mathcal{I}(\boldsymbol{\Psi}) = \left[\mathcal{C}_{\beta_0} \ldots \mathcal{C}_{\beta_p}\ \mathcal{C}_{\gamma_1}\ \mathcal{C}_{\gamma_2} \ldots \mathcal{C}_{\gamma_K}\right]$. Let $\xi_0 = 1$, $\xi_1 = (1-\gamma_1)$ and $\xi_j = -\gamma_j$ for $j = 2, \ldots, K$. Assuming these non-null constants, it follows that $\xi_0\mathcal{C}_{\beta_0} + \xi_1\mathcal{C}_{\gamma_1} + \sum_{j=2}^{K}\xi_j\mathcal{C}_{\gamma_j} = 0$. Thus, $\mathcal{I}(\boldsymbol{\Psi})$ is a singular matrix and, from Theorem 3 in (Rothenberg, 1971), it follows that the associated statistical model is not globally identifiable. □

As previously shown in Proposition 2.1, model identifiability is achieved provided that the parameter γ_1 is fixed at a known value. In the general case, it is difficult to prove directly that the Fisher information matrix $\mathcal{I}(\boldsymbol{\Psi})$ is nonsingular when we fix γ_1. However, some special cases are amenable to analytic treatment and they are illuminating for this identifiability discussion as shown in Proposition 2.3.

Proposition 2.3. *Assume that the areas experience a common relative risk $\log(\theta_i) = \beta_0$, for $i = 1, \ldots, A$. If γ_1 is fixed at a known value $\gamma_1^0 \in [0, 1]$ then the Fisher information matrix associated with the model given in expressions (2.1) and (2.2) is nonsingular.*

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

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

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

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

Remark 2.1. *As suggested by an anonymous referee, an equivalent representation of our model is obtained considering the parameterization*

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

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

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

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

2.2 Prior distributions

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

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

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

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

where $GBeta(\alpha, \nu; a, b)$ denotes the generalized Beta distribution with probability density function (p.d.f.) given by $f(x \mid \alpha, \nu; a, b) = \frac{\Gamma(\alpha+\nu)}{\Gamma(\alpha)\Gamma(\nu)(b-a)} \left(\frac{x-a}{b-a}\right)^{\alpha-1} \left(1 - \frac{x-a}{b-a}\right)^{\nu-1}$, $x \in (a, b)$, $\alpha > 0$, $\nu > 0$, $a \in \mathbb{R}$, $b \in \mathbb{R}$. The generalized Beta distribution can be obtained as the linear transformation $X = a + (b-a)B$, where $B \sim Beta(\alpha, \nu)$. By letting $a_j = 0$ and $a_j^* = 1$ for all $j = 1, \ldots, K$, the prior distribution in expression (2.5) is the well-known stick-breaking representation of the Dirichlet process, in which we consider independent random variables $Z_j \sim Beta(\alpha_j, \nu_j)$, $j = 1, \ldots, K$, and we let $\gamma_1 = Z_1$ and $\gamma_j = Z_j \prod_{l=1}^{j-1}(1 - Z_l)$ for $j = 2, \ldots, K$. This is an advantageous feature we consider to facilitate the computational implementation of the generalized Beta prior distribution.

If we set $\alpha_j = \nu_j = 1$ for $j = 1, \ldots, K$, and $0 \leq a_j < a_j^* \leq 1$, $j = 1, \ldots, K$, the conditional prior distributions given in expression (2.5) corresponds to a simpler model which is based on conditional uniform distributions so that

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

The uniform prior distribution in expression (2.6) is more parsimonious and easier to be elicited. In turn, the generalized Beta prior distribution in expression (2.5) is more flexible and provides different shapes for the marginal prior distribution of each γ_j. Thus, the choice between the prior distributions given by expressions (2.5) and (2.6)

will depend on the information that the practitioner has available. In practice, the choice of all prior hyperparameters might be driven by experts' opinion or guided by results of previous studies. Special attention, however, should be given to the prior distribution of γ_1 as it plays an important role in the proposed model identification. As discussed in Section 2.1, it has to be informative, putting a significant probability mass in the subset of the parametric space indicated by the experts as containing the most likely values for such parameter.

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

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

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

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

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

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

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

Another way to model the prior uncertainty about the model parameters is to consider the Jeffreys' approach (Jeffreys, 1946). Let $Y_i \mid \theta_i, \epsilon_i \overset{ind}{\sim} \mathcal{P}oisson(E_i \theta_i \epsilon_i)$ in which $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$. We assume that, *a priori*, $\boldsymbol{\gamma}$ is independent of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ and we only focus on the Jeffreys prior for $\boldsymbol{\gamma}$. The Fisher information matrix for the vector $\boldsymbol{\gamma}$, given $\boldsymbol{\beta}$, is the bottom right $K \times K$ submatrix highlighted in bold in $\mathcal{I}(\boldsymbol{\Psi})$ which is given in Section 2.1. Consequently, the Jeffreys prior distribution for $\boldsymbol{\gamma}$ becomes

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

The challenge is to prove that the prior in expression (2.7) is a proper distribution and to investigate the level of prior information about γ_1 that is induced by the Jeffreys prior.

Proposition 2.4. *The Jeffreys prior distribution for $\boldsymbol{\gamma}$ given in expression (2.7) is proper.*

Proof. The proof of Proposition 2.4 follows straightforwardly by noticing that the Jeffreys prior given in expression (2.7) may be represented as $\pi_J(\boldsymbol{\gamma} \mid \boldsymbol{\beta}) \propto \psi(\gamma_1)\psi(\gamma_2 \mid \gamma_1)\cdots \psi(\gamma_K \mid \gamma_1, \ldots, \gamma_{K-1})$, where $\psi(\gamma_1)$ is the kernel of the generalized Beta distribution $GBeta(1, 1/2;\ 0, 1)$ and $\psi(\gamma_k \mid \gamma_1, \ldots, \gamma_{k-1})$ is the kernel of a $GBeta\left(1, 1/2; 0, 1 - \sum_{l=1}^{k-1} \gamma_l\right)$, for $k = 2, \ldots K$. Consequently, $\pi_J(\boldsymbol{\gamma} \mid \boldsymbol{\beta})$ is proper as it belongs to the generalized Beta family of distributions given in expression (2.5). □

Assuming the Jeffreys prior in expression (2.7), the prior expected value of γ_1 is 0.6667 and its marginal prior distribution concentrates most probability mass around large values (see Figure 1). It is expected that such prior does not provide good posterior estimates for the model parameters whenever the true percentage of underreported events in areas with the best data quality is small and far from that prior expected value. Particularly, it is not an appropriate prior to model the uncertainty about γ_1 in the case study addressed in the paper where the probability of underreporting in the best areas is expected to be close to zero. To illustrate the effect of the marginal Jeffreys prior distribution of γ_1 on the joint Jeffreys prior for $\boldsymbol{\gamma}$, we present in Figure 1 the joint Jeffreys prior distribution for parameters γ_1 and γ_2. As the prior associated to γ_1 is centered around large values, the most probable prior values for the vector (γ_1, γ_2) are associated to large values for γ_1 and small values for γ_2.

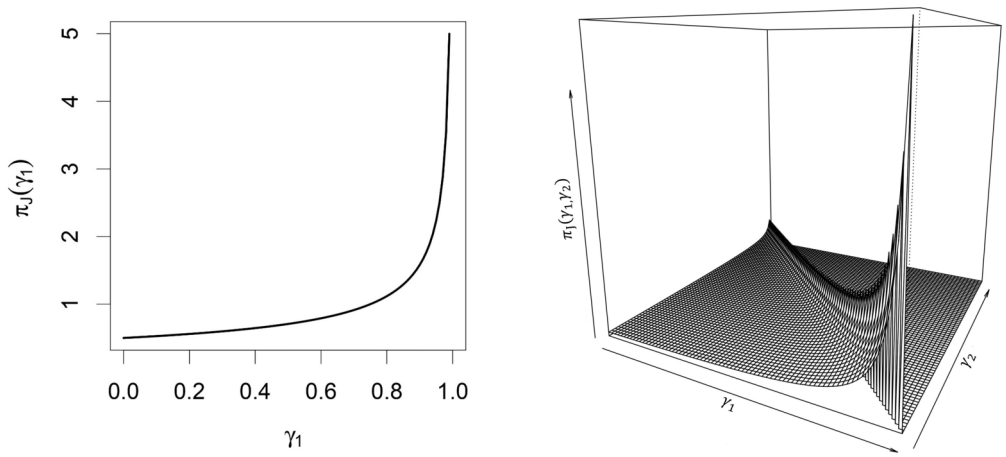


Figure 1: Marginal Jeffreys prior for γ_1 (left) and the joint Jeffreys prior for γ_1 and γ_2 (right).

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

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

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

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

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

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

3 Simulated data studies

In this section, we investigate the performance of the proposed model through Monte Carlo simulations. To mimic our case study presented in Section 4, we consider the map of Minas Gerais State that is composed of $A = 75$ areas. A total of 100 datasets are generated from Poisson distributions such that $Y_i \overset{ind}{\sim} \mathcal{P}oisson(E_i\theta_i\epsilon_i)$, for $i = 1, \ldots, 75$, where $\epsilon_i = 1 - \mathbf{h}_i^T\boldsymbol{\gamma}$ and the expected number of cases E_i is known and equal to the

one available for the case study. We also consider the same clustering indicator variable used in the case study, which has $K = 4$ data quality categories (clusters), partitioning the map in groups with a total of 28, 16, 14 and 17 areas, respectively, from the best to the worst category. This clustering variable is based on the adequacy index (AI) introduced by França *et al.* (2006). We provide a detailed explanation of the clustering construction in Section 4. We set $\boldsymbol{\gamma} = (0.05, 0.10, 0.15, 0.20)$ imposing that 5% of events are not reported in those areas classified at the highest level of data quality whereas only 50% of events are reported in those areas belonging to the worst data quality cluster. To generate the relative risks, we consider independent observations from five covariates such that $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_5 X_{5i}$, where $\beta_0 = 0.50$ and $\boldsymbol{\beta} = (-0.25, -0.25, 0, 0, 0.25)$. These covariates are different from the clustering covariate. They were selected from our real dataset such that part of them are correlated with the clustering covariate. All covariates considered here are provided in the Supplementary Material (Oliveira *et al.*, 2020).

When fitting the simulated datasets, three different structures are considered for the relative risk $\boldsymbol{\theta}$. In Model 1, we let $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_5 X_{5i}$, where $\beta_m \overset{iid}{\sim} N(0, 10)$ for $m = 0, \ldots, 5$. Model 2 differs from Model 1 by considering a variable selection scheme on the set of covariates through the SSVS prior distribution given in expression (2.8) with $\sigma^2_{spike} = 0.001$, $\sigma^2_{slab} = 10$ and $\rho_m = 0.5$ for $m = 1, \ldots, 5$. Model 3 differs from Model 2 by the inclusion of both local and spatially structured random effects in the log-linear regression such that $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_5 X_{5i} + u_i + s_i$, where $u_i \overset{iid}{\sim} \mathbf{N}(0, \sigma^2_u)$ is the local effect of area i and $\boldsymbol{s} = (s_1, \ldots, s_A)$ denotes the spatial effects having the ICAR prior distribution (Besag, York, and Mollié, 1991) with precision parameter $\tau_s = \sigma_s^{-2}$. The neighboring structure inherent to the map of case study in Section 4 is adopted to model the spatial effects $\boldsymbol{s}$ and we further assume that the model precision parameters are modeled as $1/\sigma^2_s \sim \text{Gamma}(0.5, 0.0005)$ and $1/\sigma^2_u \sim \text{Gamma}(2, 0.01)$.

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

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

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

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

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

Table 1: Bias, relative mean squared error (RMSE) and nominal coverage of 95% credible intervals (Cov.) for the estimated relative risks $\boldsymbol{\theta}$; Simulation Study I.

Table 1 summarizes the results. By eliciting an informative prior distribution only for parameter γ_1 (partially informative case), the proposed model provides good posterior estimates for the risks with bias and RMSE close of zero. The results are quite close to those obtained under informative prior for all components of parameter vector $\boldsymbol{\gamma}$ (fully informative case). In general, we observe a slight difference between results obtained under the generalized Beta prior and the conditional uniform distributions for $\boldsymbol{\gamma}$, where

the former has a greater number of hyperparameters to be chosen. Results under Models 1–3 are quite similar showing that spatial and local effects do not significantly influence the posterior inferences. This is an interesting result as the data are generated from a model that does not include any spatial or local correlation. It should be also mentioned that the non-significant (null) effect of covariates X_3 and X_4 (results not shown) is well identified even under Model 1 which does not consider variable selection.

Table 1 also shows that, as expected, the standard Poisson model fails in estimating the relative risks, $\boldsymbol{\theta}$, whenever applied to analyze underreported data. It produces very poor estimates, always underestimating the relative risks no matter the structure imposed to model them. The RMSE under such a model is reasonably small but the 95% credible intervals do not contain the true value of the relative risk for the great majority of the Monte Carlo replications, which means that the posterior distribution for $\boldsymbol{\theta}$ tends to put negligible probability mass around its true value.

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

The prior distribution for parameter γ_1 plays an important role in model identification and, consequently, in the quality of the posterior estimates. In this section, we reexamine the datasets considered in Section 3.1 fitting the proposed model with different partially informative prior distributions for $\boldsymbol{\gamma}$, that is, an informative prior distribution is considered only for the component γ_1. A sensitivity analysis is performed in order to evaluate the influence of such prior distribution on the posterior inference.

The evaluation metrics for the posterior estimates of $\boldsymbol{\theta}$ under six different conditional uniform priors for γ_1 (Table 2) show that the relative risks tend to be underestimated if, *a priori*, we elicited $\gamma_1 \sim U(0.0, 0.01)$ and $\gamma_1 \sim U(0.0, 0.05)$. Such prior distributions put all probability mass below 0.05 which is the true value of γ_1. On the other hand, the risks tend to be overestimated whenever the prior expectation exceeds the true value of γ_1. The highest the difference between the prior expectation $E(\gamma_1)$ and the true value of γ_1, the highest are the bias and RMSE of the posterior estimates of $\boldsymbol{\theta}$. This is not a surprising result and it evidences the importance of searching for reliable prior information about parameter γ_1 in practical situations.

Table 2 also shows that, if we assume $\gamma_1 \sim U(0, 0.05)$ or $\gamma_1 \sim U(0, 0.15)$, the prior means differ from the true value of γ_1 by the same amount. Although the latter prior imposes much higher prior variance than the former, the posterior estimates present similar absolute values for the bias and the RMSE in both cases. This suggests that quality of posterior estimates under the proposed model are more strongly related to the prior expectation of γ_1 than to its prior variance. Such an idea is supported by the results in Table 3 which exhibits some evaluation metrics related to posterior inference for $\boldsymbol{\theta}$ assuming different partially informative generalized Beta prior distributions for $\boldsymbol{\gamma}$. In all cases, $\gamma_1 \sim GB(\alpha_1, \nu_1; 0, 1)$ where hyperparameters α_1 and ν_1 are chosen such that this prior is centered around the true value of γ_1, that is, $E(\gamma_1) = 0.05$, but the prior uncertainty about γ_1 varies from 0.00002 to 0.00950.

Table 3 shows that the RMSE approaches zero in all cases. As expected, the bias tends to increase as the prior uncertainty about γ_1 increases. If the generalized Beta prior

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

Table 2: Bias and relative mean squared error (RMSE) for the estimated relative risks $\boldsymbol{\theta}$ assuming partially informative conditional uniform priors to $\boldsymbol{\gamma}$ with six levels of prior information on γ_1 ($E(\gamma_1)$ and $V(\gamma_1)$ are different in all cases); Simulation Study II.

distributions with $V(\gamma_1) = 0.00024$ and $V(\gamma_1) = 0.00226$ are assumed, the biases are much smaller than those observed in Table 2 under priors $\gamma_1 \sim U(0.0, 0.05)$ and $\gamma_1 \sim U(0.0, 0.15)$ whose variances are similar (respectively, $V(\gamma_1) = 0.00021$ and $V(\gamma_1) = 0.00188$). Moreover, the bias and RMSE under the prior $U(0.0, 0.30)$, which has variance equal to 0.0075, are much higher than those obtained when assuming a generalized Beta prior with a variance equal to 0.0095. In summary, these findings provide more evidence that the posterior inference is more influenced by the prior expectation of γ_1 than by its prior variance.

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

Table 3: Bias and relative mean squared error (RMSE) for the estimated relative risks $\boldsymbol{\theta}$ assuming partially informative generalized Beta priors to $\boldsymbol{\gamma}$ with six distinct levels of information on γ_1 ($E(\gamma_1) = 0.05$ (true γ_1) and a different prior variance in each case); Simulation Study II.

Table 4 exhibits the averaged posterior means for parameters β_0, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ under three out of the different partially informative conditional uniform prior distributions for γ_1 considered in previous studies. Results for Models 1–3 are quite similar, thus we only present the results obtained under Model 3. The vector of fixed effects $\boldsymbol{\beta}$ and variable selection parameter $\boldsymbol{\omega}$ are well estimated regardless of the prior distribution elicited for γ_1 but very little is learned about γ_1 from the data. The posterior mean of

γ_1 tends to be close to its prior expectation, reinforcing the importance of obtaining reliable prior information about this parameter. Posterior estimates for the remaining components of $\boldsymbol{\gamma}$ become worse as the prior expectation of γ_1 gets far from its true value and the prior variance of γ_1 increases.

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

Table 4: Averaged posterior means of β_0, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ under three different prior specifications on parameter γ_1; Simulation Study II.

Goodness of posterior estimation for parameters β_0 and γ_1 are closely related, which is not a surprising result given the identifiability issues discussed in Section 2.1. The intercept β_0 is overestimated (resp., underestimated) if γ_1 is also overestimated (resp., underestimated). Since β_0 directly affects the estimation of the relative risks $\boldsymbol{\theta}$, by overestimating (resp., underestimating) β_0, the relative risks $\boldsymbol{\theta}$ is overestimated (resp., underestimated) inducing the larger positive (resp., negative) bias shown in Table 2.

3.3 Simulation Study III: breaking the identification constraints

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

We consider the same four clusters used in the previous simulation studies, which are based on a variable called adequacy index (AI) available in our case study (Section 4). In the first scenario, named Categorical AI, the variable AI is considered in its discretized version with four categories indicating the clusters and the variable AI enter in this discretized form in both $\boldsymbol{X}$ and $\boldsymbol{H}$. In the second scenario, named Continuous AI, its discretized version is maintained in $\boldsymbol{H}$ but, for $\boldsymbol{X}$, we consider the original continuous AI

re-scaled to have a zero mean and a unit standard deviation. To generate the datasets, we set $\boldsymbol{\gamma} = (0.05, 0.10, 0.15, 0.20)$ and assume the covariates $X_{1i}, \ldots, X_{4i}$ as in the previous studies. In the Continuous AI scenario, we let $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_5 X_{5i}$, where X_{5i} is the AI in its continuous scale, $\beta_0 = 0.15$ and $\boldsymbol{\beta} = (-0.25, -0, 25, 0, 0, -0.25)$. In the Categorical AI scenario, we assume $\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_4 X_{4i} + \beta_{5,1} D_{1i} + \cdots + \beta_{5,3} D_{3i}$, where $\beta_0 = 0.15$ and $\boldsymbol{\beta} = (-0.25, -0.25, 0, 0, 0.25, 0.50, 0.75)$. The dummy variable D_{li} represents the lth level of the discretized AI for $l = 1, 2, 3$. To analyze the data, we consider the partially informative conditional uniform prior for $\boldsymbol{\gamma}$ in which $\gamma_1 \sim U(0, 0.10)$.

As expected, Table 5 shows that the posterior inferences for the relative risks are much worse if we break the identifiability constraints (Categorical AI case). However, such estimates do not lose quality if we consider strongly correlated variables to model $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ (Continuous AI case). In the Categorical AI case, Table 6 shows confounding between the parameters $\boldsymbol{\gamma}$ and the effects of the dummy variables, being all these parameters poorly estimated. This problem is not experienced by the parameters in the Continuous AI case. These findings are in perfect agreement with the theoretical identifiability results discussed in Section 2.1.

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

Table 5: Bias, relative mean squared error (RMSE) and nominal coverage of 95% credible intervals (Cov.) for the estimated relative risks $\boldsymbol{\theta}$; Simulation Study III.

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

Table 6: Averaged posterior means of β_0, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ under Model 2; Simulation Study III.

3.4 Comments on further simulation studies

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

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

Finally, from the study presented in Section 3.3 of the Supplementary Material, we note that, if the data are correctly recorded (that is, assuming $\epsilon_i = 1 \ \forall \ i$), the relative risks $\boldsymbol{\theta}$ are overestimated under our approach and the bias magnitude depends on the prior knowledge about $\boldsymbol{\gamma}$ (see Web Table 3). In this context, as expected, the standard Poisson model performs very well presenting both bias and RMSE close to zero. However, the standard Poisson model always underestimates the relative risks if counts are partially recorded (see Table 1 of the main paper), and the bias magnitude depends on the amount of underreporting in the data.

Therefore, it is important mentioning that the proposed model shows better results whenever fitted to analyze perfectly recorded data (in terms of bias and RMSE) than the standard Poisson model does whenever fitted to analyze underreported data. In practical situations, the relative risk estimates may guide the definition of government policies for control and intervention. Thus, the underestimation of such quantities leads to undesirable consequences, for instance, if we are mapping disease or mortality risks.

4 Early neonatal mortality data in Brazil

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

The counts were obtained from the *Sistema de Informações sobre Mortalidade* (SIM) and *Sistema de Informações sobre Nascidos Vivos* (SINASC) from the National Health

System of the Brazilian Ministry of Health (BMH). The 853 municipalities of MG were grouped in $A = 75$ microregions (areas) following the official division suggested by the BMH. Two periods of time comprising the two most recent Brazilian Demographic Censuses are considered, namely, 1999–2001 and 2009–2011.

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

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

To define the clustering structure, we consider the adequacy index (AI) introduced by França *et al.* (2006) as a measure of the quality of infant mortality data collected in Minas Gerais. Based on the adequacy index, França *et al.* (2006) proposed a partition of the $A = 75$ microregions of MG into $K = 4$ groups: MG1 (most adequate, $AI > 70.0$, 28 microregions), MG2 (group intermediate A, $50.1 < AI < 70.0$, 16 microregions), MG3 (group intermediate B, $20.0 < AI < 50.0$, 14 microregions) and MG4 (less adequate group, $AI < 20.0$, 17 microregions). We consider these four groups to analyze the ENM data in both periods, 1999–2001 and 2009–2011. Since there is an expectation of improved data reporting quality in recent years, the $K = 4$ clusters induced by this partition may be more heterogeneous in the period 1999–2001. In order to provide a sensitivity analysis and also attempting to reduce the effect of within cluster heterogeneity, we divide each of the previous groups in two new groups obtaining another clustering structure with $K = 8$ categories of data quality. The median of the AI within each of the four initial groups is considered for defining the new partition into eight groups. Panels (b) and (d) of Figure 2 display the groups defined in both cases (each color corresponds to a different group).

About prior elicitation

To complete the model specification a prior distribution must be elicited for each parameter, with special attention to the informative prior needed for parameter γ_1. According to experts' opinion, the reporting probability in areas experiencing the best data quality likely approaches one. Based on the information obtained from the specialists (local

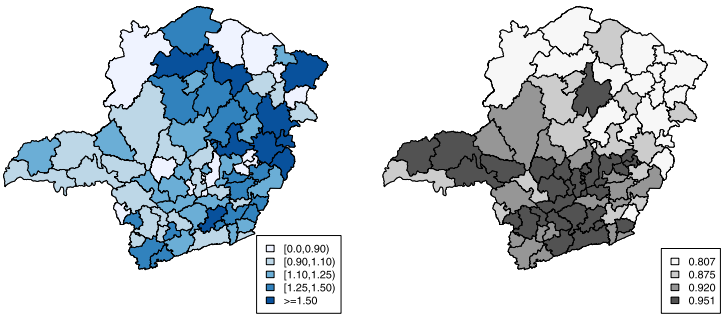
epidemiologists and health researchers) for both periods of interest, we adopt the conditional uniform prior distribution given in expression (2.6) eliciting an informative prior distribution only for parameter γ_1 (partially informative prior distribution). When considering the clustering structure with $K = 4$ data quality groups, we set $\gamma_1 \sim U(0, 0.10)$ for period 1999–2001 and, as an improvement on data reporting quality is expected in more recent years, for period 2009–2011 it is assumed $\gamma_1 \sim U(0, 0.05)$. When fitting the data with $K = 8$ clusters, we set the prior $\gamma_1 \sim U(0, 0.05)$ for both periods.

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

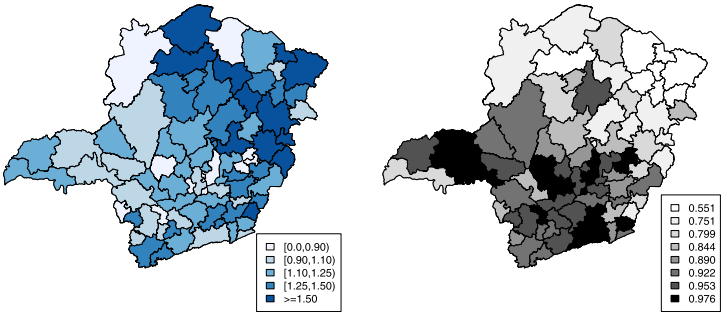
Posterior results

Figures 2 and 3 show the posterior estimates of the ENM risks in MG for periods 1999–2001 and 2009–2011, respectively. By fitting the proposed model, we estimate the probability of recording the events in each area, see Panels (b) and (d) of Figures 2 and 3. Panel (d) of Figure 2 show that, for the period 1999–2001, the posterior mean for the probability of recording an early neonatal death at areas with the worst data quality is 0.551. Such estimate increases to 0.806 in the period 2009–2011 (Panel (d), Figure 3) indicating an improvement in the data reporting process in North and Northeast areas. The same occurred for the other areas showing that an improvement in data reporting process spread out over the state. For those areas classified in the best data quality cluster, the estimated reporting probability tends to be close in both periods, which is expected as the posterior estimate for parameter ϵ_i in the best group is quite influenced by its prior mean (see the discussion in Sections 2.1 and 3.2).

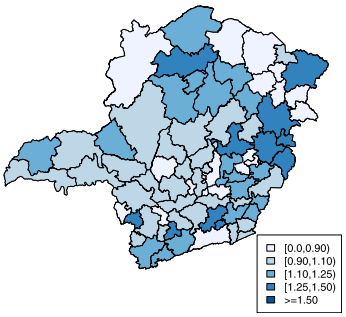
Posterior estimates for the relative risks under the standard Poisson model are displayed in Panel (e) of Figures 2 and 3. For the period 1999–2001 (Figure 2), such estimates shows that areas in the North and Northeast regions of Minas Gerais experienced the lowest ENM risks, being smaller than the risk obtained for Belo Horizonte city, the capital of the Minas Gerais State. This finding goes against the results obtained in some epidemiological studies that relate the quality of data to socioeconomic and access to health care indicators (e.g., Campos, Loschi, and França (2007)). Because the North and Northeast regions are the poorest and present the lowest socio-educational indicators in the state, experts believe that the ENM risks in such areas are much higher than those estimated through the standard Poisson model, evidencing the incapacity of this model to account for a high underregistration level. In relation to the most recent period 2009–2011 (Figure 3), the spatial distribution of the posterior estimates provided by the standard Poisson model are more compatible to what is expected by the specialists. The posterior estimates for the ENM risks in the poorest areas (North



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

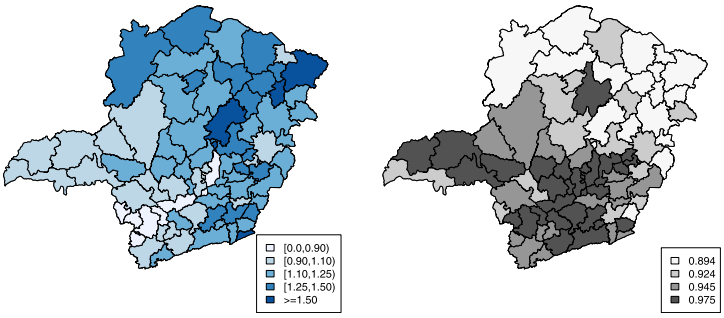


(c) $\hat{\boldsymbol{\theta}}$, proposed model, $K = 8$ (d) $\hat{\boldsymbol{\epsilon}}$, proposed model, $K = 8$

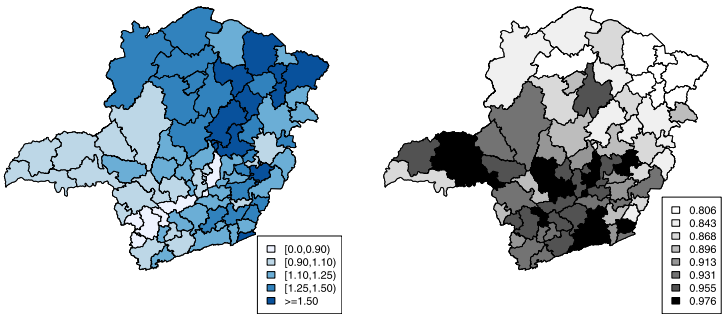


(e) $\hat{\boldsymbol{\theta}}$, standard Poisson model

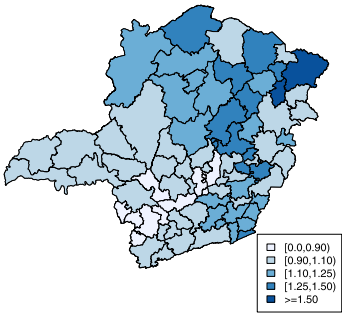
Figure 2: Posterior mean for the relative risks, $\boldsymbol{\theta}$, of early neonatal mortality (Panels (a) and (c)) and the reporting probabilities, $\boldsymbol{\epsilon}$, (Panels (b) and (d)) under the proposed model with $K = 4$ (Panels (a) and (b)) and $K = 8$ (Panels (c) and (d)) and the standard Poisson model (Panel (e)); Minas Gerais data, period 1999–2001.



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



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



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

Figure 3: Posterior mean for the relative risks, $\boldsymbol{\theta}$, of early neonatal mortality (Panels (a) and (c)) and the reporting probabilities, $\boldsymbol{\epsilon}$, (Panels (b) and (d)) under the proposed model with $K = 4$ (Panels (a) and (b)) and $K = 8$ (Panels (c) and (d)) and the standard Poisson model (Panel (e)); Minas Gerais data, period 2009–2011.

and Northeast) are higher than the ones obtained for more developed regions of Minas Gerais. It points to an improvement in the quality of the data reporting process as indicated by the estimates for the reporting probabilities obtained under the proposed model in both periods. Moreover, compared to the estimates for period 1999–2001 (Figure 2), the ENM risks for most regions in South and Southwest of Minas Gerais decrease by 2009–2011. These results are possibly indicating the advance in the socio-economic conditions and the access to health care in Minas Gerais.

Panels (a) and (c) of Figures 2 and 3 show that the proposed model provides estimates for the ENM risks in Minas Gerais that are more compatible with the findings in Campos, Loschi, and França (2007), especially in northeastern areas for both periods. Its performance is specially good when estimating the ENM risks in the period 1999–2001, in which data quality is more questionable. By accounting for underreporting, the proposed model corrects at least part of the underestimation experienced by the poorest microregions of the state providing more realistic estimates for the ENM risks in such areas. As expected, for areas experiencing a good data quality, estimation under both the proposed and the standard Poisson models are similar. As observed for the standard Poisson model, the maps for the ENM relative risks estimated under the proposed model in period 2009–2011 disclose a decrease in the risk for most microregions in South and Southwest of Minas Gerais if compared to period 1999–2001.

Table 7 summarizes the results under the fitted models. The log pseudo-marginal likelihood (LPML) criterion (Ibrahim, Chen, and Sinha, 2001) points that data from 1999–2001 are better fitted by the proposed model with $K = 8$ data quality clusters whereas for period 2009–2011 the proposed model with $K = 4$ provides the best data fit. The expected improvement in the quality of the data reporting process in the most recent period, 2009–2011, makes the microregions more homogeneous in relation to such data feature. Therefore, a smaller number of data quality categories is actually expected. For each period, only results related to the best fitted models are considered in the following analysis.

Assuming that a covariate X_m, $m = 1, \ldots, 5$, should be included into the model whenever $\hat{\omega}_m \geq 0.5$, where $\hat{\omega}_m$ denotes the posterior estimate for the associated inclusion probability, then Table 7 shows that different sets of covariates are significant to explain the ENM risks in the two analyzed periods. Under the best models, only the covariate MHDI shows to be significant (likely non-zero effect) to explain the ENM risk for the period 1999–2001 while, for the period 2009–2011, MHDI, Anomaly and Prenatal were significant. As expected in practice, the effect of the covariate MDHI is negative in both periods, indicating that the highest the MHDI, the smallest the ENM risk. The effect of MHDI is smaller in the period 2009–2011. Also for this most recent period, we observe that the ENM risk is smaller in areas with a high proportion of mothers who have made seven or more prenatal visits during the pregnancy. Furthermore, the positive effect associated to the proportion of children who were born with some congenital anomaly (Anomaly) indicates that such characteristic has been an important factor to the occurrence of early neonatal deaths in recent years. Covariates MomEduc and LowWeight, usually pointed out as important factors to explain the infant mortality rate, are not significant in the best model for both periods considered in our study.

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

Table 7: Posterior summaries for the regression effects β_0 and $\boldsymbol{\beta}$ under proposed and standard Poisson models; Minas Gerais data in both periods 1999–2001 and 2009–2011. We provide the posterior mean (Mean), the standard deviation (St.Dev.), the posterior probability of being positive ($\mathbb{P}(\beta > 0)$) and the posterior inclusion probability ($\hat{\omega}$).

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

5 Discussion

We presented a novel Bayesian modeling framework to analyze potentially underreported count data. We propose a clustering scheme that relates the reporting probabilities among the areas according to a previous data quality partitioning. Auxiliary variables and experts' opinion can be considered to assess data quality throughout the areas. One interesting feature of the proposed model is that, to ensure its identifiability,

only an informative prior for the underreporting probability in areas experiencing the best data quality is required. That is attractive because in the best areas information about the reporting probability tends to be easily accessed.

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

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

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

It is an interesting topic for future research to introduce partitioning models, such as Dirichlet process or product partition models, for underreported data. Such kind of models will allow us to also infer about the clusters throughout the estimation process. Extensions of the proposed model should also consider the situation in which there are

spatial patterns in the reporting process. By borrowing strength from spatial modeling and extreme learning machines, Prates (2019) introduce a hierarchical model to perform imputation over missing count data whose usage and adaptation for the context of underreporting is an interesting point for further investigation as well. Although not approached in this paper, the modeling of underreported count time series has been suggested in recent years, for instance, by Bracher and Held (2020) and Fernández-Fontelo *et al.* (2016). Another related problem that may interest readers is the estimation of animal abundance with differential probability of detection (see, e.g., Dorazio and Royle (2005); Hickey and Sollmann (2018)). In this context, hierarchical Poisson models are also used to model both the underlying process and the detection (reporting) probability.

Supplementary Material

Supplementary Materials for "Bias correction in clustered underreported data" (DOI: 10.1214/20-BA1244SUPP; .zip).

References

Alkema, L., and New, J. R. (2014). Global estimation of child mortality using a Bayesian B-spline bias-reduction model. *The Annals of Applied Statistics*, **8**(4), 2122–2149. MR3292491. doi: https://doi.org/10.1214/14-AOAS768. 3

Alexander, M., and Alkema, L. (2018). Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research*, **38**(15), 335–372. doi: https://doi.org/10.4054/DemRes.2018.38.15. 3

Bailey, T. C., Carvalho, M. S., Lapa, T. M., Souza, W. V., and Brewer, M. J. (2005). Modeling of under-detection of cases in disease surveillance. *Annals of Epidemiology*, **15**(5), 335–343. doi: https://doi.org/10.1016/j.annepidem.2004.09.013. 2, 3, 28

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**(1), 1–20. MR1105822. doi: https://doi.org/10.1007/BF00116466. 14, 15

Bracher, J. and Held, L. (2020). A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts. *arXiv:2003.05885 [stat.ME]*. 29

Campos, D., Loschi, R. H., and França, E. (2007). Early neonatal hospital mortality in Minas Gerais: Association with healthcare variables and the issue of underreporting (available in Portuguese). *Revista Brasileira de Epidemiologia*, **10**(2), 223–238. doi: https://doi.org/10.1590/S1415-790X2007000200010. 3, 4, 21, 23, 26

Caudill, B. S. and Mixon Jr., F. G. (1995). Modeling Household Fertility Decisions: Estimation and Testing of Censored Regression Models for Count Data. *Empirical Economics*, **20**(2), 183–196. doi: https://doi.org/10.1007/BF01205434. 2

Dorazio, R. M. and Royle, J. A. (2005). Estimating Size and Composition of Biological Communities by Modeling the Occurrence of Species. *Journal of the American Statistical Association*, **100**(470), 389–398. MR2170462. doi: https://doi.org/10.1198/016214505000000015. 29

Dvorzak, M. and Wagner, H. (2016). Sparse Bayesian modelling of underreported count data. *Statistical Modelling*, **16**(1), 24–46. MR3457686. doi: https://doi.org/10.1177/1471082X15588398. 2, 5, 6, 28

Fernández-Fontelo, A., Cabaña, A., Puig, P., and Moriña, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*, **35**(26), 4875–4890. MR3554999. doi: https://doi.org/10.1002/sim.7026. 29

França, E., Abreu, D., Campos, D. and Rausch, M. C. (2006). Avaliação da Qualidade da informação sobre a mortalidade infantil em Minas Gerais: Utilização de uma metodologia simplificada (available in Portuguese). *Revista Médica de Minas Gerais*, **16**(1 suppl. 2), S28–S35. 15, 22

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**(423), 881–889. doi: https://doi.org/10.1080/01621459.1993.10476353. 14

Gustafson, P., Gelfand, A. E., Sahu, S. K., Johnson, W. O., Hanson, T. E., Joseph, L., and Lee, J. (2005). On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables. *Statistical Science*, **20**(2), 111–140. MR2183445. doi: https://doi.org/10.1214/088342305000000098. 8

Hickey , J. R. and Sollmann, R. (2018). A new mark-recapture approach for abundance estimation of social species. *PLOS One*, **13**(12), e0208726. doi: https://doi.org/10.1371/journal.pone.0208726. 29

Huang, G. H. (2005). Model Identifiability. *Encyclopedia of Statistics in Behavioral Science*, John Wiley & Sons, Ltd, Chichester. Volume **3**, pp. 1249–1251. 8

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001) *Bayesian Survival Analysis*, Springer-Verlag, New York, p. 589. MR1876598. doi: https://doi.org/10.1007/978-1-4757-3447-8. 26

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society A*, **186**, 453–461. MR0017504. doi: https://doi.org/10.1098/rspa.1946.0056. 12

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika*, **56**, 331–347. MR0082427. doi: https://doi.org/10.1007/BF02296300. 8

Moreno, E. and Girón J. (1998). Estimating with incomplete count data: A Bayesian approach. *Journal of Statistical Planning and Inference*, **66**(1), 147–159. MR1617002. doi: https://doi.org/10.1016/S0378-3758(97)00073-6. 2, 3, 5, 6

Oliveira, G. L., Loschi, R. H., and Assunção, R. M. (2017). A random-censoring Poisson model for underreported data. *Statistics in Medicine*, **36**(30), 4873–4892. MR3734480. doi: https://doi.org/10.1002/sim.7456. 2, 28

Oliveira, G. L., Argiento, R., Loschi, R. H., Assunção, R. M., Branco, M. D. and Ruggeri, F. (2020). Supplementary material for "Bias correction in clustered underreported data". *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1244SUPP. 12, 14, 15, 21, 23

Papadopoulos, G. and Silva, J. M. C. S. (2012). Identification issues in some double-index models for non-negative data. *Economics Letters*, **117**(1), 365–367. doi: https://doi.org/10.1016/j.econlet.2012.06.001. 6, 7

Picci, G. (1977). Some connections between the theory of sufficient statistics and the identifiability problem. *SIAM Journal on Applied Mathematics*, **33**(3), 383–398. MR0464455. doi: https://doi.org/10.1137/0133025. 8

Prates, M. O. (2019). Spatial extreme learning machines: An application on prediction of disease counts. *Statistical Methods in Medical Research*, **28**(9), 2583–2594. MR4000182. doi: https://doi.org/10.1177/0962280218767985. 29

R Core Team (2015). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 (2015). Available at https://www.R-project.org/. 15

Rothenberg, T. J. (1971). Identification on parametric models. *Econometrica*, **39**(3), 577–591. MR0436944. doi: https://doi.org/10.2307/1913267. 9, 10

Schmertmann, C. and Gonzaga, M. R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*, **55**(4), 1363–1388. doi: https://doi.org/10.1007/s13524-018-0695-2. 3, 5, 6, 11

Silva, G. D. M. da, Bartholomay, P., Cruz, O. G. and Garcia, L. P. (2017). Evaluation of data quality, timeliness and acceptability of the tuberculosis surveillance system in Brazil's microregions. *Ciência & Saúde Coletiva [online]*, **22**(10), 3307–3319. doi: https://doi.org/10.1590/1413-812320172210.18032017. 3

Stamey, J. D., Young, D. M., and Boese, D. (2006). A Bayesian hierarchical model for Poisson rate and reporting-probability inference using double sampling. *Australian & New Zealand Journal of Statistics*, **48**(2), 201–212. MR2253916. doi: https://doi.org/10.1111/j.1467-842X.2006.00434.x. 2, 6, 28

Stoner, O., Economou, T., Drummond, G. (2019). A Hierarchical Framework for Correcting Under-Reporting in Count Data. *Journal of the American Statistical Association*, **114**(528), 1481–1492. MR4047275. doi: https://doi.org/10.1080/01621459.2019.1573732. 3, 5, 6, 7, 28

Whittemore, A. S. and Gong, G. (1991). Poisson Regression with Misclassified Counts: Application to Cervical Cancer Mortality Rates. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **40**(1), 81–93. 2, 5, 6, 28

World Health Organization (WHO) (2006). Neonatal and perinatal mortality: Country, regional and global estimates. *World Health Organization (WHO) Library Cataloguing-in-Publication Data*. 3

Acknowledgments

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