PLOS ONE

RESEARCH ARTICLE

A Bayesian unified framework for risk estimation and cluster identification in small area health data analysis

K. C. Flórez¤, A. Corberán-Vallet*, A. Iftimi, J. D. Bermúdez

Department of Statistics and Operations Research, University of Valencia, Valencia, Spain

¤ Current address: Department of Mathematics and Statistics, Universidad del Norte, Barranquilla, Colombia
* Ana.Corberan@uv.es

OPEN ACCESS

Citation: Flórez KC, Corberán-Vallet A, Iftimi A, Bermúdez JD (2020) A Bayesian unified framework for risk estimation and cluster identification in small area health data analysis. PLoS ONE 15(5): e0231935. https://doi.org/10.1371/journal.pone.0231935

Editor: Holger Fröhlich, University of Bonn, Bonn-Aachen International Center for IT, GERMANY

Received: June 30, 2019

Accepted: April 3, 2020

Published: May 7, 2020

Data Availability Statement: Data underlying the study is available on the Kaggle repository (simulation study: https://www.kaggle.com/josedbermudez/simulated-spatial-data-districts-of-valencia-spain; varicella case study: https://www.kaggle.com/josedbermudez/varicella-2013-data-in-districts-of-valencia-spain).

Funding: This work has been supported by Grant Number MTM2017-83850-P from the Spanish Ministry of Economy, Industry, and Competitiveness.

Abstract

Many statistical models have been proposed to analyse small area disease data with the aim of describing spatial variation in disease risk. In this paper, we propose a Bayesian hierarchical model that simultaneously allows for risk estimation and cluster identification. Our model formulation assumes that there is an unknown number of risk classes and small areas are assigned to a risk class by means of independent allocation variables. Therefore, areas within each cluster are assumed to share a common risk but they may be geographically separated. The posterior distribution of the parameter representing the number of risk classes is estimated using a novel procedure that combines its prior distribution with an efficient estimate of the marginal likelihood of the data given this parameter. An extension of the model incorporating covariates is also shown. These covariates may incorporate additional information on the problem or they may account for spatial correlation in the data. We illustrate the performance of the proposed model through both a simulation study and a case study of reported cases of varicella in the city of Valencia, Spain.

Introduction

In the last decades there has been an increasing interest in the area of disease mapping, with the consequent development of numerous statistical techniques for the analysis of public health data. Health data usually consist of aggregated counts of disease within administrative units (small areas) such as zip codes, municipalities, etc. The objective of disease mapping is then to investigate geographic disease variation across the predefined study region.

It is usually assumed that disease counts are described by a discrete probability model. For relatively rare diseases, we can assume a Poisson likelihood with a mean which is a function of the expected counts of disease and the area-specific relative risks. The expected counts of disease represent the background population effect and they are usually estimated using a standard population rate. Once estimated, they are assumed fixed and known and the emphasis is placed on the study of the unknown relative risks, which measure the local deviation of the disease. For finite populations, we could consider a Binomial distribution instead ([1], chapter 5).

Competing interests: The authors have declared that no competing interests exist.

A wide range of statistical models have been developed to provide suitable relative risk estimates. By borrowing information across the areas, these models provide smoothed risk surfaces and improve local estimates. The most common approach to relative risk modeling is to assume a logarithm link to a linear predictor which is a function of spatial random effects and possible covariate effects. For instance, the convolution model decomposes the log of the relative risk as a function of spatially correlated and uncorrelated random effects [2]. This formulation has been found to be a robust and an appropriate model to describe disease variation and it is widely used in practice [3, 4]. On some occasions, a spatial trend model capturing large-scale variation of risk over the study region may provide an accurate description of the data ([1], chapter 5). Distance-risk models can be used when a specific point source is suspected to be responsible for an increased disease risk [5]. An alternative model-based approach for smoothing risks in a spatial context is based on the use of splines. Recently, penalized splines and area-specific random effects have been combined to model large-scale spatial trend together with individual variation [6, 7]. Goicoa et al. [8] provide a comparison of CAR and P-spline models in terms of smoothing and detection of high risk areas.

Identification of areas of significantly elevated risk also plays an important role in public health, since it facilitates the implementation of targeted public health interventions. One of the most widely used tests for cluster detection is the spatial scan statistic [9] and its extensions (see, for instance, [10]). Often, it is convenient to consider clusters as a residual feature of the data, and so different residual diagnostics have been proposed to identify extreme regions ([1], chapter 6). The posterior probability that the relative risk exceeds a reference threshold has also been used to classify areas as high risk areas [11, 12]. A major concern with the usefulness of these measures is that they are sensitive to the model chosen. Besides, they attempt to detect clusters from a spatially smoothed risk surface, which may be problematic.

Models that explicitly describe the clustering behaviour of the data have also been designed. Some recent studies address risk estimation and cluster identification simultaneously. In [13] the authors propose a non-parametric mixture model within an empirical Bayes context to identify population heterogeneity. Knorr-Held and Raßer [14] propose a Bayesian partition model where small areas are combined in clusters attending a distance measure that ensures that clusters are connected. Denison and Holmes [15] present a similar Bayesian model to estimate the risk surface. The main difference is that clusters are defined using the Voronoi tessellation. A related partition model is proposed in [16], where the authors introduce a class of Markov connected component field priors to incorporate some prior information about clusters. However, the model is not intended to provide a flexible modeling of the risk surface. Hidden Markov models have also been proposed in a disease mapping context to analyze spatial heterogeneity of disease count data. For instance, [17] present a class of hidden Markov random field (HMRF) models related to an underlying finite-mixture model for the Poisson rates. To allow for spatial correlation, the allocation variables are modeled through a Potts model. One feature of this model is that disconnected areas can have the same label. An alternative frequentist approach for direct classified risk mapping based on a discrete HMRF model can be found in [18]. A different approach has been recently proposed in [19]. In that paper, the authors propose a two-step methodology where a spatially adjusted hierarchical agglomerative clustering algorithm is applied to data prior to the study period to elicit potential cluster structures. Then, for each cluster configuration, a Bayesian hierarchical model is fitted to the study data and the final cluster structure is chosen based on the deviance information criterion. A Bayesian spatio-temporal model can be found in [20].

In this paper, we develop a flexible parametric Bayesian hierarchical model that simultaneously allows for risk estimation and cluster detection in a purely spatial context. We assume that there is an unknown number k of risk levels that describe the risk surface and small area

count data are assigned to a risk class by means of independent and identically distributed Multinomial indicator vectors. Therefore, geographically separated small areas with a similar risk can be grouped in the same class. It is important to emphasize that, unlike previous model formulations, we do not impose any spatial correlation, which allows us to explore different risk structures. Conditioning on k, we develop an MCMC algorithm to sample from the joint posterior distribution of the model parameters. A novel procedure is then developed to estimate the value of k. In particular, it is obtained from its posterior distribution, which is computed as a function of the marginal likelihood of the data given k and its prior distribution. The proposed model can be straightforwardly extended to incorporate information from covariates. These covariates can account for spatial correlation in the data if present.

We organize this paper as follows. In Section 2 we present the model formulation and describe our MCMC implementation. In Section 3 we analyze the performance of the model using a simulation study. Section 4 provides an application to reported varicella cases in the city of Valencia, Spain. We conclude with a general discussion of the proposed methodology and provide directions for future research.

Model formulation and Bayesian analysis

Let us assume that the study region is divided in m contiguous non-overlapping small areas, and let y_i represent the count of disease observed in the i-th area, $i = 1, 2, \ldots, m$. We also assume that there is a piecewise constant risk surface that describes geographic disease variation, and so counts of disease are modeled with the following mixture of Poisson distributions:

$$f(y_i|k, \eta, p, e_i) = \sum_{j=1}^{k} \text{Po}(y_i|e_i\eta_j)p_j, \tag{1}$$

where e_i is the expected count of disease, k is the number of risk classes (that can be viewed as clusters), $\eta = (\eta_1, \eta_2, \ldots, \eta_k)'$ is the vector of risks, and $p = (p_1, p_2, \ldots, p_k)'$ is the vector of probabilities associated with the risk classes, with $p_j \geq 0$ and $\sum_{j=1}^{k} p_j = 1$. If the small areas in the study region have a similar disease risk, then the components of the risk vector η will take similar values. It may even be the case that only one latent risk ($k = 1$) suffices to describe counts of disease across the entire study region. On the other hand, a risk surface exhibiting abrupt jumps will dominate.

Let $z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})'$ be the latent indicator vector that assigns the small area i to one of the k risk classes; that is, z_{ij} is equal to one if the relative risk for area i corresponds to η_j and zero otherwise. Eq (1) can then be formulated as:

$$\begin{aligned} y_i|k, z_i, \eta, e_i &\sim \text{Po}(y_i|e_i z_i'\eta), \\ z_i|p &\sim \text{Mult}(z_i|n = 1, p). \end{aligned} \tag{2}$$

Since a common prior distribution is assumed for all the indicator vectors z_i, $i = 1, 2, \ldots, m$, this model formulation implies that the prior probabilities of the areas belonging to every possible class are the same.

Bayesian analysis of the model when k is known

Let us suppose first that the number of risk classes k is known. Prior distributions for the model parameters η and p can be specified as follows. In order to mitigate the label switching problem common in mixture models, we propose to order the risk classes. This can be easily

achieved by using the transformation:

$$\eta_j^* = \frac{\eta_j}{1+\eta_j}, \quad j = 1, \ldots, k,$$

which transforms the interval $[0, +\infty)$ into $[0, 1]$, and introducing a vector of cut-off points $v = (v_0 = 0, v_1, \ldots, v_{k-1}, v_k = 1)'$ so that:

$$f(\eta^*|v) = \prod_{j=1}^{k} \mathrm{Un}(v_{j-1}, v_j) = \prod_{j=1}^{k} \frac{1}{v_j - v_{j-1}} \mathrm{I}_{(v_{j-1}, v_j)}(\eta_j^*),$$

where $\mathrm{I}_{A(.)}$ denotes the indicator function of the set A. The prior distribution for the risk vector η is then given by:

$$f(\eta|v) = \prod_{j=1}^{k} \frac{1}{(v_j - v_{j-1})} \frac{1}{(1+\eta_j)^2} \mathrm{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)}(\eta_j). \tag{3}$$

The prior distribution that we propose for the vector v is based on the distances between a cut-off point and the following, that is $d_j = v_j - v_{j-1}$. Note that these distances are positive and add up to one, so a natural prior distribution for the vector $d = (d_1, d_2, \ldots, d_k)'$ is the Dirichlet distribution, $d \sim Dir(d|\gamma)$. We assign γ a particular value, specifically $\gamma_j = 1\ \forall j$, which corresponds to the flat Dirichlet distribution. Under that assumption, $E(d_j) = 1/k$. However, other values of γ are also possible. The prior distribution for the vector v is then given by:

$$f(v) = \prod_{j=1}^{k} (v_j - v_{j-1})^{\gamma_{j-1}}, \tag{4}$$

if $v_0 = 0 \leq v_1 \leq \ldots \leq v_{k-1} \leq v_k = 1$.

As a prior distribution for p we consider the Dirichlet distribution:

$$p \sim Dir(p|\alpha), \tag{5}$$

where α is also assigned a particular value. The choice of α here is important, since it may affect the posterior results. Values larger than 1 may prevent some classes from being empty (see, for instance, [21] and the references therein). Under the assumption that the number of risk classes k is known, we can use a weakly informative prior, such as $\alpha_j = 2\ \forall j$, which facilitates the estimation of the risks within each class and, consequently, convergence of MCMC chains. Other values of α are also possible (for more details, see the sensitivity analysis performed in the case study section).

The joint posterior distribution for the model parameters is given by:

$$\begin{aligned} f(z,p,\eta,v|y,k,e) &\propto f(y|k,z,\eta,e)f(z|p)f(p)f(\eta|v)f(v) \\ &\propto \left[\prod_{i=1}^{m}(z_i'\eta)^{y_i}\exp\{-e_i z_i'\eta\}\right]\left[\prod_{i=1}^{m}z_i'p\right]\left[\prod_{j=1}^{k}p_j^{\alpha_j-1}\right] \\ &\quad \prod_{j=1}^{k}\frac{1}{(v_j-v_{j-1})}\frac{1}{(1+\eta_j)^2}\mathrm{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}},\frac{v_j}{1-v_j}\right)}(\eta_j) \\ &\quad \prod_{j=1}^{k}(v_j-v_{j-1})^{\gamma_j-1}\mathrm{I}_{(v_{j-1},v_{j+1})}(v_j). \end{aligned} \tag{6}$$

This posterior distribution is analytically intractable but can be sampled using MCMC simulation techniques.

MCMC algorithm. To sample from the joint posterior distribution of the model parameters, we suggest an MCMC simulation algorithm based on the Gibbs sampling procedure, which requires the full conditional posterior distributions to be known [22].

The latent indicator vector $z_i \in \{E_1, E_2, \ldots, E_k\}$, E_j being the j-th column of the identity matrix I_k. The probability of each value is given by:

$$\begin{aligned} f(z_i = E_j|y,k,p,\eta,v,e) &\propto f(y_i|k,z_i=E_j,\eta,e_i)f(z_i=E_j|p) \\ &\propto \eta_j^{y_i}\exp\{-e_i\eta_j\}p_j. \end{aligned}$$

So, the full conditional posterior distribution for each vector z_i, $i = 1, 2, \ldots, m$, is the discrete distribution defined as:

$$f(z_i = E_j|y,k,p,\eta,v,e) = \frac{\eta_j^{y_i}\exp\{-e_i\eta_j\}p_j}{\sum_{l=1}^{k}(\eta_l^{y_i}\exp\{-e_i\eta_l\}p_l)}, \quad j = 1,\ldots,k. \tag{7}$$

The full conditional posterior distribution for parameter p is given by:

$$\begin{aligned} f(p|y,k,z,\eta,v,e) &\propto f(z|p)f(p) \\ &\propto \prod_{j=1}^{k}p_j^{(\sum_{i=1}^{m}z_{ij})+\alpha_j-1}; \end{aligned} \tag{8}$$

that is, $p \sim Dir(p|\alpha + z_{+.})$, where $z_{+.}$ is the resultant vector after adding the columns of the indicator matrix z.

The full conditional posterior distribution of the latent risk η_j, $j = 1, 2, \ldots, k$, can be obtained as:

$$\begin{aligned} f(\eta_j|y,k,z,p,\eta_{-j},v,e) &\propto \prod_{i:z_i=E_j}f(y_i|k,z_i=E_j,\eta_j,e_i)f(\eta_j|v) \\ &\propto (\eta_j)^{y_{(j)}}\exp\{-e_{(j)}\eta_j\}\frac{1}{(1+\eta_j)^2}\mathrm{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}},\frac{v_j}{1-v_j}\right)}(\eta_j) \\ &\propto \mathrm{Ga}[\eta_j|y_{(j)}+1,e_{(j)}]\quad\frac{1}{(1+\eta_j)^2}\mathrm{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}},\frac{v_j}{1-v_j}\right)}(\eta_j), \end{aligned} \tag{9}$$

where $y_{(j)}$ and $e_{(j)}$ represent, respectively, the sum of the observed and expected counts of

disease of those areas that are assigned to the j-th risk class. Note that if the j-th risk class is empty, the posterior distribution of parameter η_j corresponds to the prior distribution.

Finally, the full conditional posterior distribution of each cut-off point v_j, $j = 1, 2, \ldots, k-1$, is given by:

$$\begin{aligned} f(v_j|y,k,z,p,\eta,v_{-j},e) &\propto f(\eta|v)f(v) \\ &\propto \frac{1}{v_j - v_{j-1}} \mathrm{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)}(\eta_j) \\ &\quad \frac{1}{v_{j+1} - v_j} \mathrm{I}_{\left(\frac{v_j}{1-v_j}, \frac{v_{j+1}}{1-v_{j+1}}\right)}(\eta_{j+1}) \\ &\quad (v_j - v_{j-1})^{\gamma_j - 1}(v_{j+1} - v_j)^{\gamma_{j+1}-1} \mathrm{I}_{(v_{j-1}, v_{j+1})}(v_j) \\ &\propto (v_j - v_{j-1})^{\gamma_j - 2}(v_{j+1} - v_j)^{\gamma_{j+1}-2} \mathrm{I}_{\left(\frac{\eta_j}{1+\eta_j}, \frac{\eta_{j+1}}{1+\eta_{j+1}}\right)}(v_j). \end{aligned} \tag{10}$$

Note that only the last indicator function is necessary, since the previous ones imply $v_{j-1} < \frac{\eta_j}{1+\eta_j} < v_j$ and $v_j < \frac{\eta_{j+1}}{1+\eta_{j+1}} < v_{j+1}$.

Details on how to simulate from these full conditional posterior distributions are given in S1 Appendix.

Bayesian analysis of the model when k is unknown

In practice, however, the number of risk classes is unknown and so it has to be estimated. Previous approaches to estimation of k include the use of a reversible jump MCMC algorithm [14, 17], randomized model search strategies [16] or a penalty based approach that fixes k to be overly large and penalises values in the extreme risk classes [20]. In the context of Bayesian finite mixture models, [21] propose a criterion to determine the appropriate number of latent classes based on the posterior distribution of the class proportions. Concretely, the mixture model is first estimated with a relatively large number of latent classes. The true number of latent classes is then estimated as the posterior mode of the number of non-empty classes (where a class is defined as empty in an MCMC iteration if the proportion of observations assigned to that class is below a certain cut-off). This criterion, which can be easily computed using standard software for Bayesian analysis, has proved to perform well if the hyperparameters of the Dirichlet prior on the class proportions is sufficiently vague.

Here, we pursue a novel alternative methodology, which provides a discrete posterior distribution $f(k|y, e)$ on the integers $\{1, 2, \ldots, K\}$. Our alternative is based on the marginal likelihood of the data given the parameter k:

$$f(y|k,e) = \int\int f(y|k,p,\eta,e) f(p,\eta|k)\, dp d\eta. \tag{11}$$

Note that this marginal likelihood can be approximated by Monte-Carlo integration using a sample $\{p^{(n)}, \eta^{(n)}\}_{n=1}^{N}$ from the prior distribution of parameters p and η. However, this approach may not be efficient if the prior distribution is non-informative. In order to reduce the Monte-Carlo error, we define the following procedure. Let us first assume that $k = 1$. In that case, the only parameter of the model is η, and the marginal likelihood can be formulated

as:

$$
\begin{aligned}
f(y|k=1,e) &= \frac{\prod_{i=1}^{m} e_i^{y_i}}{\prod_{i=1}^{m} y_i!}\int_0^{\infty} \eta^{\sum_{i=1}^{m} y_i}\exp\{-\eta\sum_{i=1}^{m} e_i\}\frac{1}{(1+\eta)^2}d\eta \\
&= \frac{\prod_{i=1}^{m} e_i^{y_i}}{\prod_{i=1}^{m} y_i!}\Gamma\left(1+\sum_{i=1}^{m} y_i\right)(\sum_{i=1}^{m} e_i)^{-(1+\sum_{i=1}^{m} y_i)}\int_0^{\infty}\frac{1}{(1+\eta)^2}\mathrm{Ga}\left(\eta|1+\sum_{i=1}^{m} y_i,\sum_{i=1}^{m} e_i\right)d\eta \quad (12)\\
&= \frac{\prod_{i=1}^{m} e_i^{y_i}}{\prod_{i=1}^{m} y_i!}\Gamma\left(1+\sum_{i=1}^{m} y_i\right)(\sum_{i=1}^{m} e_i)^{-(1+\sum_{i=1}^{m} y_i)}E((1+\eta)^{-2}).
\end{aligned}
$$

Given a sample $\{\eta^{(n)}\}_{n=1}^{N}$ from $\mathrm{Ga}(\eta|1+\sum_{i=1}^{m} y_i, \sum_{i=1}^{m} e_i)$, Eq (12) can be estimated as:

$$
\hat{f}(y|k=1,e) = \frac{\prod_{i=1}^{m} e_i^{y_i}}{\prod_{i=1}^{m} y_i!}\Gamma\left(1+\sum_{i=1}^{m} y_i\right)(\sum_{i=1}^{m} e_i)^{-(1+\sum_{i=1}^{m} y_i)}\frac{1}{N}\sum_{n=1}^{N}\frac{1}{(1+\eta^{(n)})^2}.
$$

In the general case $k > 1$, $g(\eta, p, d|y, k, e) = g(\eta|y, k, d, e)g(p|y, k, d, e)g(d|k, e)$ is defined as a naive approximation to the posterior distribution of the model parameters and it is used as an importance function. We next describe this importance function when $k = 2$. Its generalization for values of $k > 2$ is straightforward. If $k = 2$, $\eta = (\eta_1, \eta_2)'$, $p = (p_1, 1 - p_1)'$, and $d = (d_1, 1 - d_1)'$. In particular, we assume that $g(d_1|k) = \mathrm{Be}(d_1|\gamma_1, \gamma_2)$. Once the value of d_1 has been observed, each small area is assigned to the first risk class if $\frac{y_i}{e_i} < \frac{d_1}{1-d_1}$. Let m_1 be the number of small areas assigned to the first risk class, and $m_2 = m - m_1$ the number of small areas assigned to the second risk class. Function $g(p_1|y, k, d, e)$ is defined as $\mathrm{Be}(p_1|\alpha_1 + m_1, \alpha_2 + m_2)$. Finally, $g(\eta_1|y, k, d, e)$ and $g(\eta_2|y, k, d, e)$ are defined, respectively, from the empirical distribution of $\{y_i|e_i\}_{i:y_i/e_i<d_1/(1-d_1)}$ and $\{y_i|e_i\}_{i:y_i/e_i>d_1/(1-d_1)}$. The marginal likelihood function can then be expressed as:

$$
\begin{aligned}
f(y|k,e) &= \int\int\int f(y|k,p,\eta,d,e)f(p,\eta,d|k)dpd\eta dd \\
&= \int\int\int f(y|k,p,\eta,e)f(p,\eta|k,d)f(d|k)dpd\eta dd \\
&= \int\int\int f(y|k,p,\eta,e)f(p,\eta|k,d)\frac{g(\eta,p,d|y,k,e)}{g(\eta,p|y,k,d,e)}dpd\eta dd \\
&= E\left(\frac{f(y|k,p,\eta,e)f(p,\eta|k,d)}{g(\eta,p|y,k,d,e)}\right),
\end{aligned} \quad (13)
$$

which can be easily estimated using a sample from $g(\eta, p, d|y, k, e)$.

Once the marginal likelihood function of the data given the parameter k has been estimated for $k = 1, 2, \ldots, K$, the value of k can be obtained from its posterior distribution:

$$
f(k|y,e) \propto \hat{f}(y|k,e)f(k) \quad (14)
$$

where $f(k) = q_k$ is a discrete prior distribution on the integers $\{1, 2, \ldots, K\}$.

The advantage of this procedure is that it allows for cluster identification and risk estimation simultaneously. If we are interested in identifying the number of risk classes, we can select the value of k within the set $\{1, 2, \ldots, K\}$ with the highest posterior probability. This criterion is similar to the one proposed in [21] in the sense that one value of k is selected from the posterior distribution. However, if the final objective of the study is to estimate the relative risk of each small area i, $i = 1, 2, \ldots, m$, we can derive relative risk estimates accounting for model

uncertainty through posterior averaging. That is, given the parameters of the model, the relative risk of area i is:

$$\theta_i = \sum_{j=1}^{k} \eta_j P(z_i = E_j | y, k, p, \eta, \nu, e) \tag{15}$$

Using a sample generated from the joint posterior distribution of the model parameters with the Gibbs sampling procedure, the relative risks can be estimated as:

$$\hat{E}(\theta_i | y, k, e) = \frac{1}{N}\sum_{n=1}^{N} \theta_i^{(n)} = \frac{1}{N}\sum_{n=1}^{N} (z_i^{(n)})' \eta^{(n)}.$$

In the general case of k unknown, we can use the posterior distribution of parameter k to estimate the relative risk for area i as:

$$\hat{E}(\theta_i | y, e) = \sum_{k=1}^{K} \hat{E}(\theta_i | y, k, e) f(k | y, e). \tag{16}$$

Extension of the model with covariates

Let us know assume that we have information from L covariates, x_i being the $L \times 1$ vector of covariate information corresponding to the i-th small area, $i = 1, 2, \ldots, m$. In this case, counts of disease are modeled as:

$$\begin{aligned} y_i | k, z_i, \eta, \beta, e_i, x_i &\sim \text{Po}(y_i | e_i \cdot \exp\{x_i'\beta\} \cdot z_i'\eta), \\ z_i | p &\sim \text{Mult}(z_i | n = 1, p), \end{aligned} \tag{17}$$

where the difference with respect to Eq (2) is the incorporation of the extra term $\exp\{x_i'\beta\}$ in the mean of the Poisson distribution.

The prior distribution assumed for parameter β is the multivariate Gaussian distribution with zero mean vector and covariance matrix Σ. As hyperprior for parameter Σ, we propose an inverse Wishart distribution with ν degrees of freedom and scale matrix Ψ. A typically used relatively uninformative inverse Wishart prior sets $\nu = L + 1$ and $\Psi = I_L$ (see, for instance, [23]). The particular case $L = 1$ corresponds to an inverted Gamma distribution $\sigma_\beta^2 \sim Ga^{-1}(a, b)$ with $a = \nu/2$ and $b = \Psi/2$.

The full conditional posterior distributions for parameters z_i, $i = 1, 2, \ldots, m$, p, η_j, and ν_j, $j = 1, 2, \ldots, k$, remain as those in Eqs (7)–(10), the only difference is that e_i is substituted by $e_i \cdot \exp\{x_i'\beta\}$.

The full conditional posterior distributions for parameters β and Σ are given by:

$$\begin{aligned} f(\beta | y, k, z, p, \eta, \nu, \Sigma, e, x) \quad \propto \quad & \exp\left\{ \left(\sum_{i=1}^{m} y_i x_i \right)' \beta \right\} \exp\left\{ -\sum_{i=1}^{m} e_i \cdot \exp\{x_i'\beta\} \cdot z_i'\eta \right\} \\ & \exp\{-\frac{1}{2}\beta'\Sigma^{-1}\beta\}. \end{aligned} \tag{18}$$

$$f(\Sigma | y, k, z, p, \eta, \nu, \beta, e, x) \quad \propto \quad |\Sigma|^{-\frac{\nu+L+2}{2}} \exp\left\{ -\frac{1}{2} tr\left((\beta\beta' + \Psi)\Sigma^{-1} \right) \right\}. \tag{19}$$

Simulation from the full conditional posterior distribution of parameter β can be done using the Metropolis algorithm. In particular, we propose to use the uniform distribution in a

ball in $\mathbb{R}^L$ centered at the previous simulated value of β and radius r as the proposal density. The full conditional posterior distribution of parameter Σ is an inverse Wishart distribution with $v+1$ degrees of freedom and scale matrix $\beta\beta' + \Psi$. Simulation from this distribution is straightforward.

As explained in the introduction, most of the statistical models that have been proposed to explain the spatial pattern in small area disease data incorporate spatially autocorrelated random effects, which account for any spatial autocorrelation remaining in the data after the known covariate effects have been accounted for. Because many covariates vary spatially, spatial confounding between spatially structured random effects and fixed effects is a well-known problem. One possible solution is to use a multi-stage modelling process (see, for instance, [24]). It is important to emphasize here that, because the proposed model does not impose any spatial correlation, it is not affected by collinearity problems, and so fixed-effects can be properly estimated.

Simulation study

We present here some of the results obtained in an extensive simulation study that we conducted to assess the effectiveness of the proposed model to recover the true underlying risk surface (for more details, see [25]). All the analysis was performed using the free statistical software R.

Data

We used the city of Valencia (Spain), which consists of $m = 86$ contiguous boroughs, as the base map to generate the observed disease count data at borough level. To calculate the expected counts of disease, we used the population size of each borough and assumed an incidence rate $r = 30$ per 10^4 persons. In *Scenario 1*, the parameters assumed for the simulation were $k = 3$, $\eta = (0.5, 1.0, 3.5)$, and $p = (0.2, 0.3, 0.5)$. In *Scenario 2*, the parameters assumed for the simulation were $k = 7$, $\eta = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$, and $p = (1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7)$. To generate the observed disease counts within each scenario, we first assigned the small areas to the different risk levels simulating from a discrete distribution with probabilities given by the corresponding vector p. Then, each y_i, $i = 1, 2, \ldots, m = 86$, was simulated from a Poisson distribution with mean $e_i z_i' \eta$, where e_i was calculated taking into account the population size of the i-th borough and the assumed value of r, and $z_i'\eta$ is the risk corresponding to the risk class where the i-th borough was assigned.

For comparative purposes, we also simulated data from the convolution model [2]. In *Scenario 3*, the relative risk for the i-th borough (θ_i) was simulated as:

$$\theta_i = \exp\{\rho + u_i + v_i\} \tag{20}$$

where $\rho \sim N(0, \sigma_\rho^2)$ is the overall level of the relative risk in the study region; $u = (u_1, u_2, \ldots, u_m)'$ represents the spatially correlated random effect, following a CAR model with variance σ_u^2; and $v = (v_1, v_2, \ldots, v_m)'$ is assumed to be a realization of a multivariate Gaussian distribution with zero mean vector and covariance matrix $\sigma_v^2 I_m$. The values of the standard deviances were $(\sigma_\rho, \sigma_u, \sigma_v) = (0.01, 0.05, 0.1)$. The observed disease count for the i-th borough was then simulated from a Poisson distribution with mean $e_i\,\theta_i$, where e_i was calculated taking into account the population size of the borough and an incidence rate $r = 30$ per 10^4 persons.

Finally, we also considered a scenario where different risks were assigned to each borough without imposing any spatial correlation between risks in nearby areas. In particular, in

Scenario 4, the relative risk for the i-th borough was simulated as:

$$\theta_i = \exp\{\rho + v_i\} \tag{21}$$

where parameters ρ and v_i are defined as those in Eq (20). The values of the standard deviances were $(\sigma_\rho, \sigma_v) = (0.01, 0.1)$. This scenario represents a situation where the relative risks are not spatially correlated but, unlike our model formulation, the small areas are not assigned to risk classes.

To allow for sampling variability, we simulated 100 data sets for each scenario. The results presented are averaged over these 100 realizations.

Evaluation measures

We mainly used two measures to evaluate the performance of our model: The root mean squared error (RMSE) and the percentage of correspondence.

Let θ_i and $\hat{\theta}_i$ be the true and the estimated relative risk for the i-th small area, $i = 1, 2, \ldots, m$. The RMSE is defined as $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\theta_i - \hat{\theta}_i)^2}$.

In the analysis of each data set with our model, we first assumed that the value of k was known and simulated 5000 samples from the posterior distribution of the model parameters using the MCMC simulation algorithm described in S1 Appendix. The values of k that we considered ranged from $k = 1$ to $k = 10$; that is, we sequentially implemented the MCMC algorithm for $k = 1, 2, \ldots, K = 10$. We calculated then the posterior distribution of parameter k as previously explained (see Eq (14)) using 100000 samples in the Monte-Carlo integration step. Finally, the estimated relative risk for each borough was calculated using Eq (16). The resulting RMSE is represented by RMSE_{mod}. For comparative purposes, we also show the RMSE_{smr} based on the standardized mortality rate ($\hat{\theta}_i = y_i/e_i$) and the RMSE_{BYM} that was obtained by fitting the convolution model to the simulated dataset. To carry out this analysis, we used the free statistical software WinBUGS.

The percentage of correspondence is calculated as the percentage of data sets within each scenario that satisfy that the value of k with the highest posterior probability corresponds to the true value of k. Note that this measure can only be calculated for *Scenario* 1 and *Scenario* 2, since the data were simulated from our model and so we know the true value of k.

Results

For *Scenario* 1, the median and interquartile range of the estimated values of η and p when we assume that k is known and equal to 3 are displayed in Table 1. As can be seen, when we condition on the true number of risk classes, the risks within each class are accurately estimated and the proportion of areas assigned to each risk class coincide with the real probabilities of the classes.

We next consider k as an additional parameter of the model and compute its posterior distribution. Table 2 shows the percentage of data sets (out of the 100) that select each possible

Table 1. Scenario 1: Median (interquartile range) of the estimated values of η and p when we assume $k = 3$. True values are $\eta = (0.5, 1.0, 3.5)$ and $p = (0.2, 0.3, 0.5)$.

$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$
0.49 (0.06)	1.00 (0.06)	3.50 (0.07)
$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
0.22 (0.07)	0.29 (0.07)	0.49 (0.06)

https://doi.org/10.1371/journal.pone.0231935.t001

Table 2. Scenario 1: Percentage of data sets that satisfy that the highest probability of the posterior distribution of k coincides with each possible value of k considered.

k	1	2	3	4	5	6	7	8	9	10
	0%	0%	94%	4%	2%	0%	0%	0%	0%	0%

https://doi.org/10.1371/journal.pone.0231935.t002

Table 3. Scenario 2: Median (interquartile range) of the estimated values of η and p when we assume $k = 7$. True values are η = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5) and p = (1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7).

$\hat{\eta_1}$	$\hat{\eta_2}$	$\hat{\eta_3}$	$\hat{\eta_4}$	$\hat{\eta_5}$	$\hat{\eta_6}$	$\hat{\eta_7}$
0.18 (0.05)	0.46 (0.08)	0.75 (0.08)	1.09 (0.12)	1.65 (0.14)	2.56 (0.18)	4.38 (3.38)
$\hat{p_1}$	$\hat{p_2}$	$\hat{p_3}$	$\hat{p_4}$	$\hat{p_5}$	$\hat{p_6}$	$\hat{p_7}$
0.13 (0.03)	0.14 (0.03)	0.16 (0.04)	0.16 (0.04)	0.13 (0.03)	0.14 (0.04)	0.14 (0.05)

https://doi.org/10.1371/journal.pone.0231935.t003

value of k considered (from 1 to 10) as the most probable value. In this scenario, a percentage of correspondence equal to 94% is achieved, being the mean of the posterior probability of the value $k = 3$ equal to 0.79.

The corresponding results for *Scenario 2* are presented in Tables 3 and 4. As in the previous scenario, when we condition on the true number of risk classes ($k = 7$ here), the estimated risks within each class and the probabilities of these classes are close to the real values. As expected, however, the difficulty of identifying the risk classes increases with the number of risk classes used in the simulation and also when there is a low separation between the risks. Under this scenario, the values of k that are most often selected are $k = 7$, 8, and 9. In this case, the mean of the posterior probability of the value $k = 7$ is 0.20.

A comparison with the results provided by the criterion proposed in [21] to select the number of risk classes is shown in S2 Appendix.

The mean of three RMSE previously described are shown in Table 5. This table also displays the 95% Bayesian prediction interval for the difference between the $RMSE_{BYM}$ and the $RMSE_{mod}$ that would be obtained after applying the convolution model and the proposed model to a new data set. As expected, the $RMSE_{smr}$ is the highest one, while the differences between the $RMSE_{mod}$ and the $RMSE_{BYM}$ are not significant except when the data show a marked clustering behaviour. On those occasions, the proposed model provides more satisfactory results.

Table 4. Scenario 2: Percentage of data sets that satisfy that the highest probability of the posterior distribution of k coincides with each possible value of k considered.

k	1	2	3	4	5	6	7	8	9	10
	0%	0%	0%	0%	3%	12%	19%	25%	27%	14%

https://doi.org/10.1371/journal.pone.0231935.t004

Table 5. Mean of the RMSEs obtained in the estimation of the relative risks for the small areas and 95% Bayesian prediction interval for the difference between the $RMSE_{BYM}$ and the $RMSE_{mod}$ that would be obtained after applying the convolution model and the proposed model to a new data set.

	$RMSE_{smr}$	$RMSE_{mod}$	$RMSE_{BYM}$	95% Pred Int
Scenario 1	0.422	0.204	0.397	[0.063,0.322]
Scenario 2	0.342	0.293	0.320	[-0.033,0.088]
Scenario 3	0.286	0.105	0.100	[-0.023,0.012]
Scenario 4	0.284	0.098	0.094	[-0.020,0.011]

https://doi.org/10.1371/journal.pone.0231935.t005

When counts of disease are either spatially correlated or do not show a clustering behaviour, the proposed model is able to capture such behaviour and the relative risk estimates are similar to those provided by the convolution model.

Case study

In this section we study the number of reported varicella cases in the city of Valencia for year 2013. These data were obtained from the Surveillance Service and Epidemiological Control, General Division of Epidemiology and Health Surveillance—Department of Public Health, Generalitat Valenciana. Since the data provided were aggregated counts of disease at borough level, we did not have access to any identifying information of patients.

Varicella, also known as chickenpox, is an acute and highly contagious viral, airborne disease. It is primarily a disease of children, characterized by blister-like rash, itching, tiredness, and fever. It usually resolves by itself within a couple of weeks. Immunity following infection is considered to be long-lasting and reinfections are rare.

The city of Valencia (the third largest city in Spain) consists of 19 districts divided into 87 boroughs. Districts 17, 18 and 19 are thinly populated and far from the urban core. Hence, we consider here the analysis of varicella data in 70 boroughs of Valencia (corresponding to districts 1-16) for year 2013. The number of cases registered in these districts represents over 98% of the total.

In our analysis of the data, we have first assumed that the value of k is known and we have simulated 5000 samples from the posterior distribution of the model parameters using the MCMC simulation algorithm described in S1 Appendix. We performed a sensitivity analysis to assess the influence of the values assumed for the hyperparameters on posterior results. We considered values of γ and α different to the ones proposed in the description of the Bayesian analysis of the model. Concretely, we also assumed $\gamma_j = 2$ and $\alpha_j = 0.4$ or $\alpha_j = 1$, $\forall j$. In all cases, the correlation among the posterior estimates of the parameters was greater than 0.998. The results presented here correspond to the values $\gamma_j = 1$ and $\alpha_j = 2$. The values of k that we have considered range from $k = 1$ to $k = 10$. We have calculated then the posterior distribution of parameter k using 100000 samples in the Monte-Carlo integration step. This posterior distribution is shown in Table 6.

If the objective of the study is to identify clusters of disease, we would select the value of $k = 5$ as the number of risk classes. The posterior estimates of the risks and the probabilities associated with the risk classes are given in Table 7.

The estimated posterior probabilities $f(z_i = E_j|y, k = 5, e)$, E_j being the j-th column of the identity matrix I_5, allow us to assign each small area i to the risk class corresponding to the highest probability. Fig 1 maps the clustering behavior of the data, where each color

Table 6. Case study: Posterior distribution of parameter *k*.

k	1	2	3	4	5	6	7	8	9	10
	0	0	0.01	0.10	0.40	0.20	0.13	0.08	0.05	0.03

https://doi.org/10.1371/journal.pone.0231935.t006

Table 7. Case study: Median (interquartile range) of the estimated values of η and p when we assume $k = 5$.

$\hat{\eta_1}$	$\hat{\eta_2}$	$\hat{\eta_3}$	$\hat{\eta_4}$	$\hat{\eta_5}$
0.46 (0.22)	0.82 (0.26)	1.12 (0.30)	1.85 (0.84)	5.34 (0.62)
$\hat{p_1}$	$\hat{p_2}$	$\hat{p_3}$	$\hat{p_4}$	$\hat{p_5}$
0.19 (0.19)	0.34 (0.21)	0.26 (0.24)	0.10 (0.12)	0.05 (0.03)

https://doi.org/10.1371/journal.pone.0231935.t007

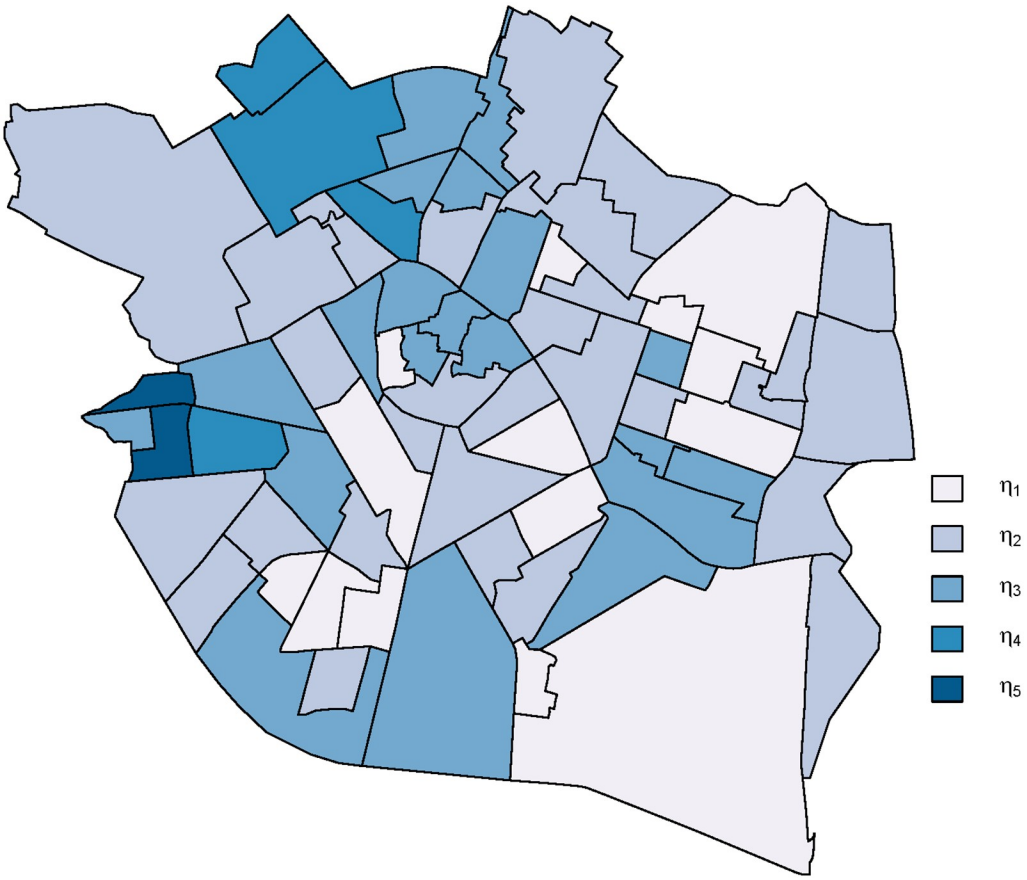

Fig 1. Case study: Clustering behavior of the varicella data for the 70 boroughs of the city of Valencia. The value of $k = 5$ has been selected as the number of risk classes.

https://doi.org/10.1371/journal.pone.0231935.g001

corresponds to a risk level and areas represented with the same color indicate that they belong to the same risk class. Hence, geographic disease variation is described by a piecewise constant risk surface.

If we are interested in obtaining relative risk estimates for each borough, we can derive them using Eq (16), which allows us to incorporate model uncertainty. A summary of the estimated relative risks for the 70 boroughs of Valencia is shown in Table 8. Fig 2 maps the estimated relative risks. For comparative purposes, we also include the results obtained by the convolution model. Note that the estimated relative risks with our model are very similar to those provided by the convolution model. The log-pseudo marginal likelihood ($\sum_{i=1}^{m} log(CPO_i)$) corresponding to our model and that of the convolution model are, respectively, -178.30 and -213.41. So, the fit provided by the proposed model is slightly better than that of the convolution model. It is important to emphasize here that, even though our model does not impose any spatial correlation, it properly describes the spatial distribution of varicella in the city of Valencia.

Table 8. Case study: Summary of the estimated relative risks for the 70 boroughs of Valencia. The results obtained with the convolution model are also included.

	Min	Q1	Median	Mean	Q3	Max
Proposed model	0.46	0.74	0.89	1.06	1.06	5.34
Convolution model	0.49	0.67	0.89	1.06	1.11	5.62

https://doi.org/10.1371/journal.pone.0231935.t008

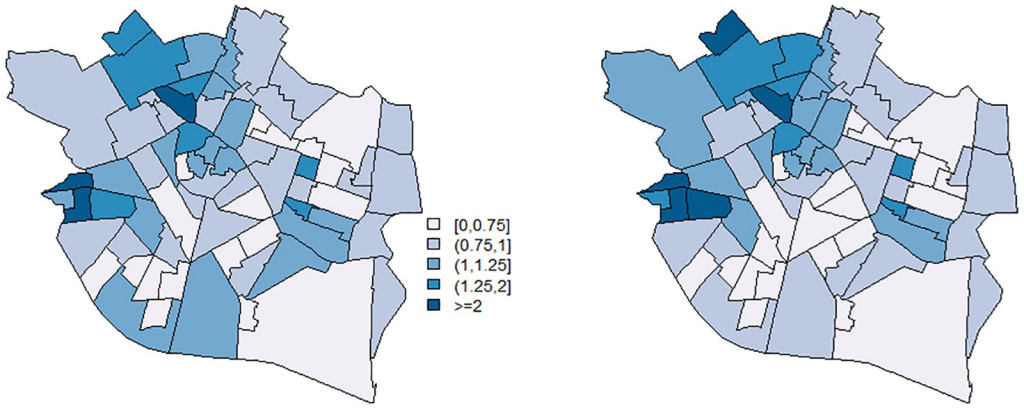

Fig 2. Case study: Estimated relative risks for the 70 boroughs of the city of Valencia. Left: Results obtained with our model. Right: Results obtained by applying the convolution model.

https://doi.org/10.1371/journal.pone.0231935.g002

We now show the results obtained in the analysis of the data when we include one covariate. In particular, for each borough, we have information corresponding to the percentage of population aged 0-4, which constitutes one of the high-risk age groups. Posterior point estimate of parameter β is 0.19, being the 95% confidence interval equal to [0.05, 0.31]. This estimate is consistent with the estimate obtained after fitting a model including the covariate effect and uncorrelated random effects (namely, $log(\theta_i) = \alpha + \beta x_i + v_i$, $v_i \sim N(0, \sigma_v^2)$): $\hat{\beta} = 0.15$, and $CI_{95\%}(\beta) = [-0.007, 0.309]$. From 2006 to 2013, varicella vaccine was available in pharmacies and young children could be vaccinated according to parents' criteria. A more informative covariate would be the percentage of unvaccinated population aged 0-4. However, this information is not available at the borough level. Nevertheless, the proposed model is capable of estimating the effect of the available covariate and the residual relative risks for each borough. Fig 3 maps the total impact of covariate ($E(\exp\{\beta x_i\}|y, e, x)$) and the mean of the Poisson distribution without the term e_i (that is, $E(\exp\{\beta x_i\} \cdot z_i'\eta|y, e, x)$). As expected, for each borough, the estimated mean of the Poisson distribution is practically the same as the relative

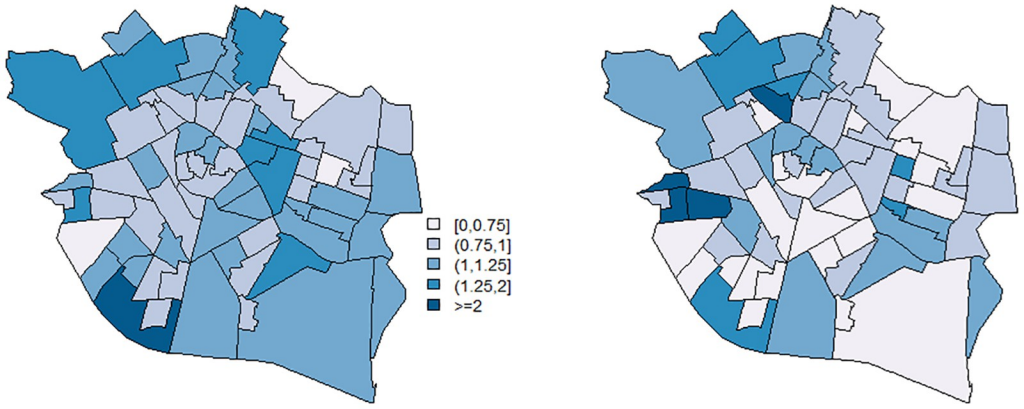

Fig 3. Case study. Left: Total impact of covariate. Right: Estimated mean of the Poisson distribution without the term e_i for the 70 boroughs of the city of Valencia.

https://doi.org/10.1371/journal.pone.0231935.g003

risks estimated with the proposed model without covariate information. The corresponding log-pseudo marginal likelihood is -176.45.

Discussion

We have proposed a Bayesian Poisson mixture model that allows for relative risk estimation and cluster detection. By looking at the posterior distribution of each indicator vector z_i, the small areas can be grouped together in, possibly multiple, risk classes. However, posterior averaging over both the generated samples and the number of risks classes considered leads to a smoothly varying risk surface.

Our model formulation allows that areas belonging to a specific risk class are disconnected. This idea has been explored in previous papers, but some kind of spatial dependence in the definition of the prior distributions for the allocation variables is always assumed [17, 18]. A major novelty of the proposed methodology is that we do not impose any spatial correlation at any level of the model hierarchy. However, as shown in the case study, the model can properly explain the spatial distribution of the data under study. The results obtained in a simulation study also demonstrate the good performance of our procedure in a variety of situations encompassing both smooth and discontinuous cases. This flexibility is of utter importance because, in practice, there is little prior information about the underlying risk surface.

It is also important to emphasize that we have treated the number of risks classes k as an additional parameter of the model. We have proposed here a novel methodology based on the marginal likelihood of the data given parameter k to estimate its posterior distribution. The integral to obtain such marginal distribution is solved by Monte Carlo integration using an approximation to the posterior distribution of the model parameters. As shown, this is an efficient and straightforward to program procedure.

If the effect of known explicative variables is taken into account, the proposed model aids in the identification of unexplained components of risk. Because our model formulation does not impose spatial correlation, it avoids the spatial confounding problem and provides a suitable framework to estimate the fixed-effect coefficients associated with spatially-structured covariates. Covariate information could also be used to introduce some spatial correlation in the model (for instance distance to putative foci of risk) or to define informative prior distributions for the indicator vectors. A very fruitful area for further research would be the incorporation and selection of multiple covariates.

It would also be valuable to extend the model to the spatio-temporal domain. This would allow us to describe the spatial distribution of disease risk and also its evolution over time. A possible extension can be obtained by defining, for each small area, a set of temporally correlated indicator vectors $\{z_{it}\}_{t=1}^{T}$. Because the assignment of each small area to a risk class can change, the model so defined would allow the relative risks to evolve in time.

Supporting information

S1 Appendix. Gibbs sampling procedure.
(PDF)

S2 Appendix. Identification of the number of risk classes: A comparison.
(PDF)

S1 File.
(ZIP)

Acknowledgments

We thank the Surveillance Service and Epidemiological Control, General Division of Epidemiology and Health Surveillance—Department of Public Health, Generalitat Valenciana, for providing the varicella data.

Author Contributions

Data curation: A. Iftimi.

Formal analysis: K. C. Flórez, A. Corberán-Vallet, A. Iftimi, J. D. Bermúdez.

Investigation: K. C. Flórez, A. Corberán-Vallet, J. D. Bermúdez.

Methodology: K. C. Flórez, A. Corberán-Vallet, J. D. Bermúdez.

Software: K. C. Flórez, A. Corberán-Vallet, J. D. Bermúdez.

Writing – original draft: K. C. Flórez, A. Corberán-Vallet, J. D. Bermúdez.

References

1. Lawson AB. Bayesian Disease Mapping. Hierarchical Modeling in Spatial Epidemiology, 2nd Edition. Boca Raton: CRC Press; 2013.
2. Besag J, York J, Mollié A. Bayesian image restoration with two applications in spatial statistics. Ann Inst Stat Math. 1991; 43:1–59. https://doi.org/10.1007/BF00116466
3. Lawson AB, Biggeri AB, Boehning D, Lesaffre E, Viel JF, Clark A, et al. Disease mapping models:an empirical evaluation. Stat Med. 2000; 19:2217–2241. https://doi.org/10.1002/1097-0258(20000915/30)19:17/18<2217::aid-sim565>3.0.co;2-e PMID: 10960849
4. Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res. 2005; 14:35–59. https://doi.org/10.1191/0962280205sm388oa PMID: 15690999
5. Wakefield JC, Morris SE. The Bayesian modeling of disease risk in relation to a point source. J Am Stat Assoc. 2001; 96:77–91. https://doi.org/10.1198/016214501750332992
6. Lee DJ, Durbán M. Smooth-CAR mixed models for spatial count data. Comput Stat Data Anal. 2009; 53:2968–2979. https://doi.org/10.1016/j.csda.2008.07.025
7. Perperoglou A, Eilers PHC. Penalized regression with individual deviance effects. Comput Stat. 2010; 25:341–361. https://doi.org/10.1007/s00180-009-0180-x
8. Goicoa T, Ugarte MD, Etxeberria J, Militino AF. Comparing CAR and P-spline models in spatial disease mapping. Environ Ecol Stat. 2012; 19:573–599. https://doi.org/10.1007/s10651-012-0201-8
9. Kulldorff M. A spatial scan statistic. Commun Stat.—Theory Methods. 1997; 26:1481–1496. https://doi.org/10.1080/03610929708831995
10. Wakefield J, Kim A. A Bayesian model for cluster detection. Biostatistics. 2013; 14:752–765. https://doi.org/10.1093/biostatistics/kxt001 PMID: 23476026
11. Richardson S, Thomson A, Best NG, Elliott P. Interpreting posterior relative risk estimates in disease mapping studies. Environ Health Perspect. 2004; 112:1016–1025. https://doi.org/10.1289/ehp.6740 PMID: 15198922
12. Hossain M, Lawson AB. Cluster detection diagnostics for small area health data:with reference to evaluation of local likelihood models. Stat Med. 2006; 25:771–786. https://doi.org/10.1002/sim.2401 PMID: 16453370
13. Schlattmann P, Böhning D. Mixture models and disease mapping. Stat Med. 1993; 12:1943–1950. https://doi.org/10.1002/sim.4780121918 PMID: 8272672
14. Knorr-Held L, Raßer G. Bayesian detection of clusters and discontinuities in disease maps. Biometrics. 2000; 56:13–21. https://doi.org/10.1111/j.0006-341x.2000.00013.x PMID: 10783772
15. Denison DGT, Holmes CC. Bayesian partitioning for estimating disease risk. Biometrics. 2001; 57:143–149. https://doi.org/10.1111/j.0006-341x.2001.00143.x PMID: 11252589
16. Gangnon RE, Clayton MK. Bayesian detection and modeling of spatial disease clustering. Biometrics. 2000; 56:922–935. https://doi.org/10.1111/j.0006-341x.2000.00922.x PMID: 10985238
17. Green PJ, Richardson S. Hidden Markov models and disease mapping. J Am Stat Assoc. 2002; 97:1055–1070. https://doi.org/10.1198/016214502388618870

18. Charras-Garrido M, Abrial D, De Goer J, Dachian S, Peyrard N. Classification method for disease risk mapping based on discrete hidden Markov random fields. Biostatistics. 2012; 13:241–255. https://doi.org/10.1093/biostatistics/kxr043 PMID: 22133757

19. Anderson C, Lee D, Dean N. Identifying clusters in Bayesian disease mapping. Biostatistics. 2014; 15:457–469. https://doi.org/10.1093/biostatistics/kxu005 PMID: 24622038

20. Lee D, Lawson AB. Cluster detection and risk estimation for spatio-temporal health data. 2019 [cited 2019 June 30]. Available from: https://arxiv.org/abs/1408.1191.

21. Nasserinejad K, van Rosmalen J, de Kort W, Lesaffre E. Comparison of criteria for choosing the number of classes in Bayesian finite mixture models. PLoS ONE. 2017; 12(1):e0168838. https://doi.org/10.1371/journal.pone.0168838 PMID: 28081166

22. Casella G, George EI. Explaining the Gibbs Sampler. Am Stat. 1992; 46:167–174. https://doi.org/10.1080/00031305.1992.10475878

23. Alvarez I, Niemi J, Simpson M. Bayesian inference for a covariance matrix. Conference on Applied Statistics in Agriculture. 2014. https://doi.org/10.4148/2475-7772.1004.

24. Baer DR, Lawson AB. Evaluation of Bayesian multiple stage estimation under spatial CAR model variants. J Stat Comput Simul. 2019; 89:98–144. https://doi.org/10.1080/00949655.2018.1536755

25. Flórez-Lozano KC. Modelo de conglomerados para el análisis bayesiano de datos epidemiológicos en áreas pequeñas. Doctoral Dissertation, The University of Valencia. 2016. Available from http://roderic.uv.es/handle/10550/53930?show=full.