

**Supplementary material for the manuscript “Spatial
autoregressive modelling of epidemiological data: Geometric
mean model proposal”**

Mabel Morales-Otero^{1,2}, Christel Faes³ and Vicente Núñez-Antón⁴

January 2025

The material contained herein is supplementary to the article named
in the title and published in SORT-Statistics and Operations
Research Transactions Volume 49(1).

¹ Institute of Data Science and Artificial Intelligence (DATAI), University of Navarra, Calle Universidad 6, 31009, Pamplona, Spain. mmoralesote@unav.es

² TECNUN School of Engineering, University of Navarra, Manuel Lardizabal Ibilbidea 13, 20018, Donostia-San Sebastián, Spain.

³ I-BioStat, Center for Statistics, Hasselt University, Agoralaan gebouw D, 3590 Diepenbeek, Belgium. christel.faes@uhasselt.be

⁴ Department of Quantitative Methods, University of the Basque Country (UPV/EHU), Avenida Lehen-dakari Aguirre 83, 48015 Bilbao, Spain. vicente.nunezanton@ehu.eus

1. Introduction

We include some supplementary material that can be useful to the readers of the manuscript “Spatial autoregressive modelling of epidemiological data: Geometric mean model proposal”. In Section 2 we describe the algorithm we have followed to implement a Gibbs sampler for the simulation study. Section 3 includes a table that summarizes how many times each model was selected as the best fitting one in the simulation study. In Section 4 we present the heatmaps of the matrices considered in this paper. Section 5 includes the predicted incidence maps for some of the fitted models. In Section 6 we present the scatterplots of the observed versus the predicted rates, obtained from some of the fitted models. Section 7 includes some figures about the distributional assumptions in the fitted models. In Section 8, we provide some additional details about the BYM2 and Leroux models. Finally, in Section 9, we include a detailed discussion about the effect of the geometric mean of the rates on the estimated disease rate, considering different values for the spatial parameter estimate.

2. Gibbs sampling algorithm

We present the algorithm we have followed to implement a Gibbs sampler for simulating Poisson spatially autocorrelated data. First, we define a set of initial values for the parameters β , ρ and τ and, then, on each iteration we draw Poisson samples, where the mean is conditioned on the values of the previous iteration. For $k = 1$ we draw Poisson samples from an uncorrelated mean:

$$\begin{aligned}\mu_i^{(1)} &= \exp(\beta + \log(P_i) + v_i^{(1)}) \\ y_i^{(1)} &\sim \text{Poi}(\mu_i^{(1)}) \\ r_i^{(1)} &= \frac{(y_i^{(1)} + 1)}{P_i} \quad \text{for } i = 1, \dots, n\end{aligned}\tag{S1}$$

Then, if we perform a total number of S^* iterations, for $k = 2, \dots, S^*$, the algorithm follows, so that:

$$\begin{aligned}\mu_1^{(k)} &= \exp\{\beta + \log(P_1) + \rho \sum_{j=1}^n w_{1j} \log(r_j^{(k-1)}) + v_1^{(k)}\} \\ y_1^{(k)} &\sim \text{Poi}(\mu_1^{(k)}) \\ r_1^{(k)} &= \frac{(y_1^{(k)} + 1)}{P_1} \\ \mu_2^{(k)} &= \exp\{\beta + \log(P_2) + \rho (w_{21} \log(r_1^{(k)}) + \sum_{j=2}^n w_{2j} \log(r_j^{(k-1)})) + v_2^{(k)}\} \\ y_2^{(k)} &\sim \text{Poi}(\mu_2^{(k)}) \\ r_2^{(k)} &= \frac{(y_2^{(k)} + 1)}{P_2} \\ &\vdots\end{aligned}$$

$$\begin{aligned}
\mu_h^{(k)} &= \exp\{\beta + \log(P_h) + \rho(\sum_{j=1}^{h-1} w_{hj} \log(r_j^{(k)}) + \sum_{j=h}^n w_{hj} \log(r_j^{(k-1)})) + v_h^{(k)}\} \\
y_h^{(k)} &\sim \text{Poi}(\mu_h^{(k)}) \\
r_h^{(k)} &= \frac{(y_h^{(k)} + 1)}{P_h} \\
&\vdots \\
\mu_n^{(k)} &= \exp\{\beta + \log(P_n) + \rho \sum_{j=1}^{n-1} w_{nj} \log(r_j^{(k)}) + v_n^{(k)}\} \\
y_n^{(k)} &\sim \text{Poi}(\mu_n^{(k)}) \\
r_n^{(k)} &= \frac{(y_n^{(k)} + 1)}{P_n}
\end{aligned} \tag{S2}$$

3. WAIC and CPO counts

In Table S1, we have included the counts for the number of times that the information criteria and the predictive accuracy values were smaller in each case so that we can check how many times each model was selected as the best fitting one.

Table S1. Number of times that the information criteria and the predictive accuracy values selected each of the models fitted to the simulated datasets as the best fitting ones.

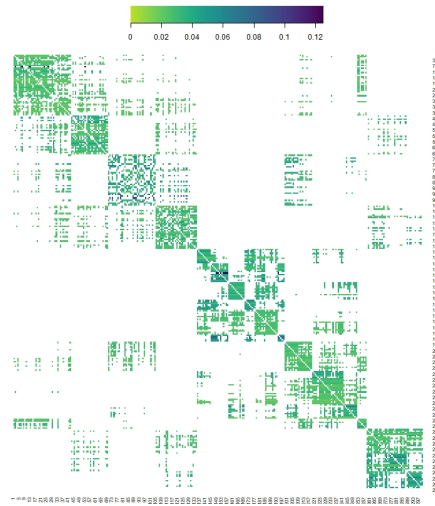
True values		Mobility	Contiguity	Inverse distance	BYM2	Leroux
$\beta = -2, \rho = 0.5, \tau = 15$	WAIC CPO	248 250	0 0	0 0	0 0	2 0
$\beta = -2, \rho = 0.5, \tau = 5$	WAIC CPO	250 154	0 0	0 0	0 79	0 17
$\beta = -0.5, \rho = 0.5, \tau = 5$	WAIC CPO	250 248	0 2	0 0	0 0	0 0
$\beta = -0.5, \rho = 0.5, \tau = 15$	WAIC CPO	249 249	0 0	0 0	0 0	1 1
$\beta = -2, \rho = 0.2, \tau = 15$	WAIC CPO	211 238	0 0	0 0	35 0	1 12
$\beta = -2, \rho = 0.2, \tau = 5$	WAIC CPO	247 143	0 107	3 0	0 0	0 0
$\beta = -0.5, \rho = 0.2, \tau = 5$	WAIC CPO	201 144	0 52	0 0	4 3	45 51
$\beta = -0.5, \rho = 0.2, \tau = 15$	WAIC CPO	154 166	0 33	0 0	1 0	95 84
$\beta = -2, \rho = 0.9, \tau = 15$	WAIC CPO	248 250	0 0	2 0	0 0	0 0
$\beta = -2, \rho = 0.9, \tau = 5$	WAIC CPO	230 249	1 0	19 0	0 1	0 0
$\beta = -0.5, \rho = 0.9, \tau = 5$	WAIC CPO	203 250	9 0	19 0	2 0	0 0
$\beta = -0.5, \rho = 0.9, \tau = 15$	WAIC CPO	168 250	0 0	0 0	0 0	82 0

4. Heatmaps

Figure S1 shows the heatmaps of the weights matrices considered here. Heatmaps use colours to represent the values of the weights for each matrix. Thus, white would indicate that the weights are zero for those municipalities. That is, heatmaps are graphical representations for the individual weights in each of the different weights matrix structures. More specifically, matrices following the inverse and the negative exponential distance only have zeros in their diagonal, for the weights w_{ii} , $i = 1, \dots, n$, whereas the rest of the matrices have a larger percentage of weights that are zero. Moreover, matrices presenting the largest number of connected areas are the ones following the contiguity of order three and the mobility matrix. Finally, matrices where the contiguity of order one criterion is considered are the ones that present the smallest number of connected municipalities and, hence, the ones having the largest percentage of weights that are zero.

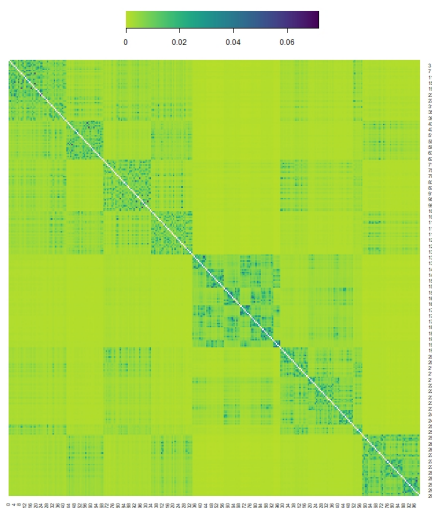


(a) Heatmap of the spatial weights matrix following contiguity of order 1.

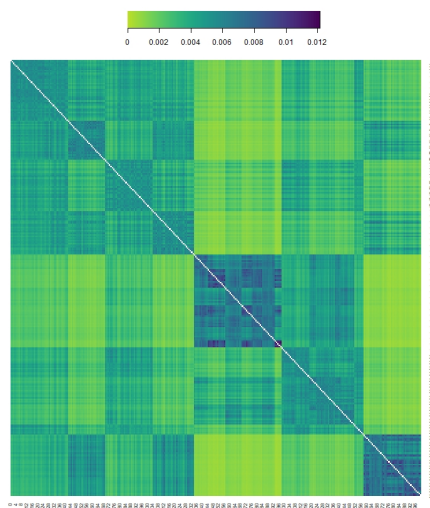


(b) Heatmap of the spatial weights matrix following contiguity of order 3.

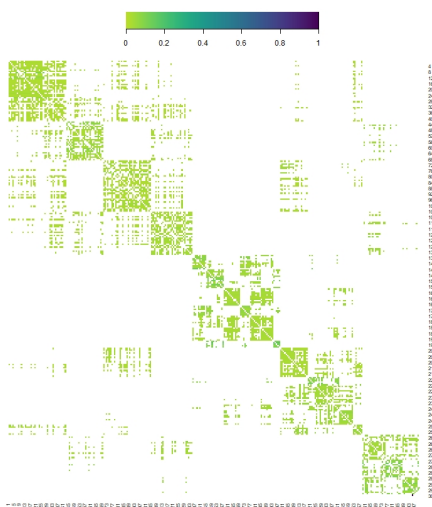
Figure S1. Heatmaps of the spatial weights matrices considered.



(c) Heatmap of the spatial weights matrix following inverse distance.



(d) Heatmap of the spatial weights matrix following negative exponential.



(e) Heatmap of the spatial weights matrix following distance band.



(f) Heatmap of the similarity spatial weights matrix combining contiguity order 1 and the variable ***budgetmeters***.

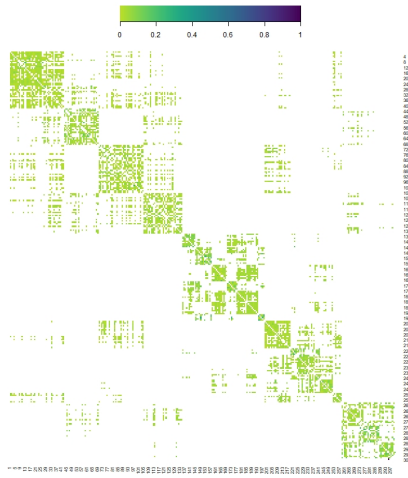
Figure S1. Heatmaps of the spatial weights matrices considered (Continued).



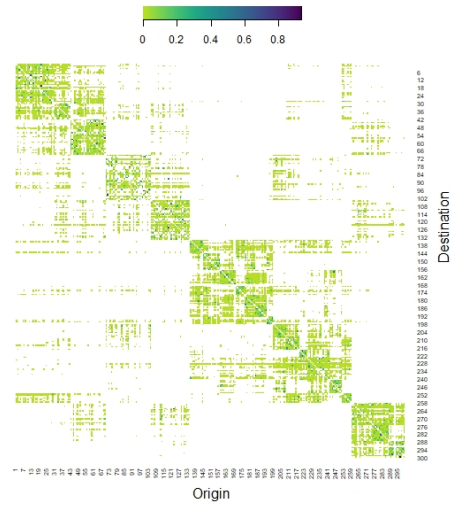
(g) Heatmap of the similarity spatial weights matrix combining distance band and the variable *budgetmeters*.



(h) Heatmap of the similarity spatial weights matrix combining contiguity of order 1 and the variable *single_house*.



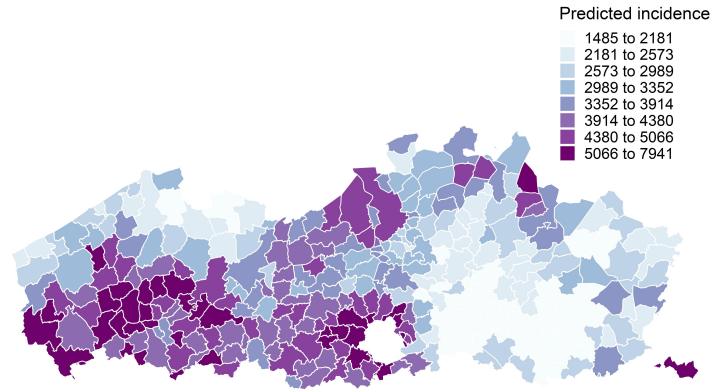
(i) Heatmap of the similarity spatial weights matrix combining distance band and the variable *single_house*.



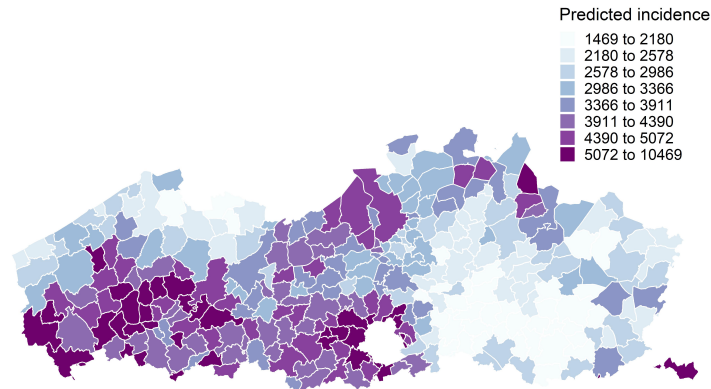
(j) Heatmap of the mobility matrix.

Figure S1. Heatmaps of the spatial weights matrices considered (Continued).

5. Predicted incidence maps



(a) Predicted incidence obtained from the model using the spatial matrix following the contiguity of order one criterion, fitted to the COVID-19 data in Flanders.

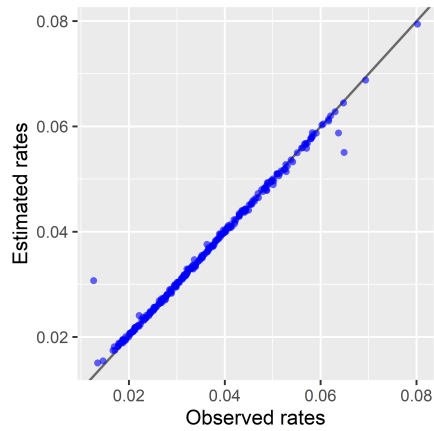


(b) Predicted incidence obtained from the model using the mobility matrix, fitted to the COVID-19 data in Flanders.

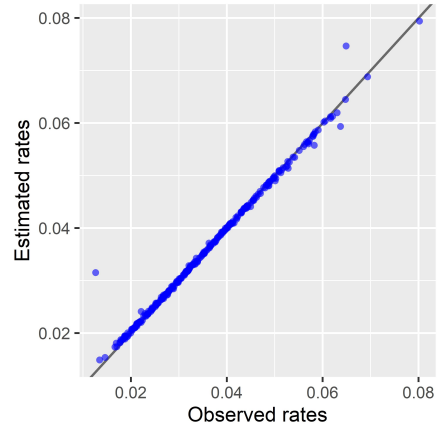
Figure S2. Predicted incidence obtained from some of the geometric mean spatial conditional normal Poisson models considered, fitted to the COVID-19 data in Flanders.

6. Scatterplots

The scatterplots of the observed versus the predicted rates, obtained from the fitting of some of the models considered are included in Figure S3. Here, we can see how the fitted models show high accuracy in the prediction of the incidence rates.

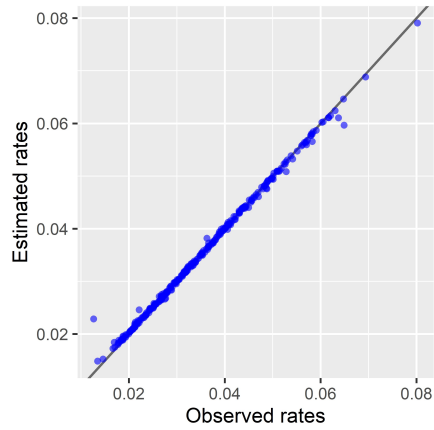


(a) Model where the spatial matrix follows the contiguity of order one criterion.

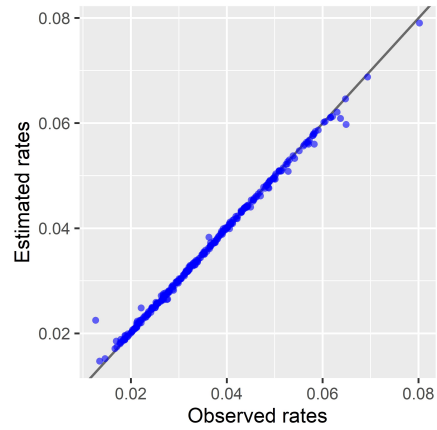


(b) Geometric mean model where the spatial matrix follows the contiguity of order three criterion.

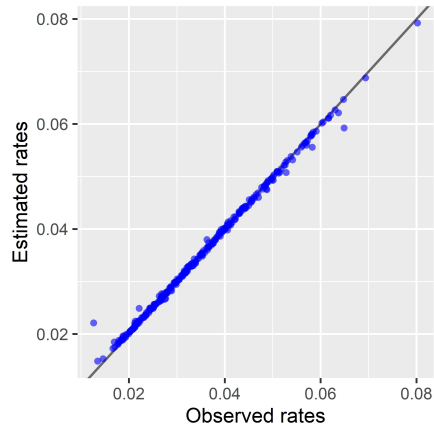
Figure S3. Scatterplots of the observed versus the predicted rates obtained for some of the fitted models.



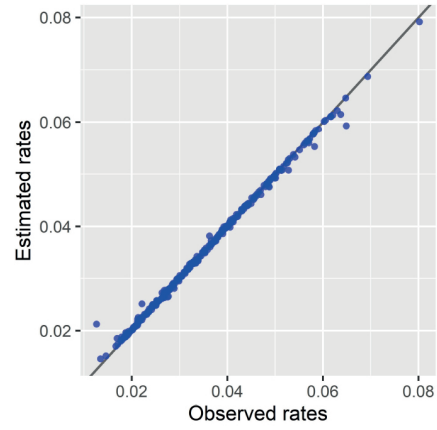
(c) Model where the spatial matrix follows the negative exponential distance criterion.



(d) Geometric mean model where the spatial matrix follows the distance bands criterion.

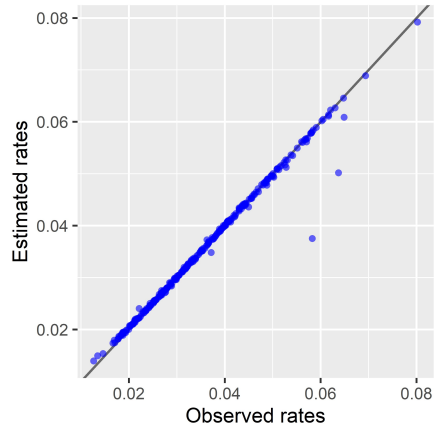


(e) Model for the similarity spatial weights matrix combining distance band and the variable **budgetmeters**.

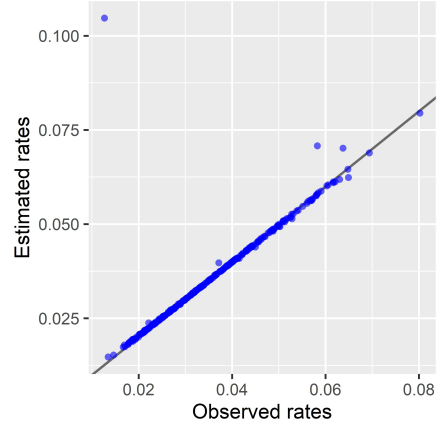


(f) Geometric mean model for the similarity spatial weights matrix combining distance band and the variable **budgetmeters**.

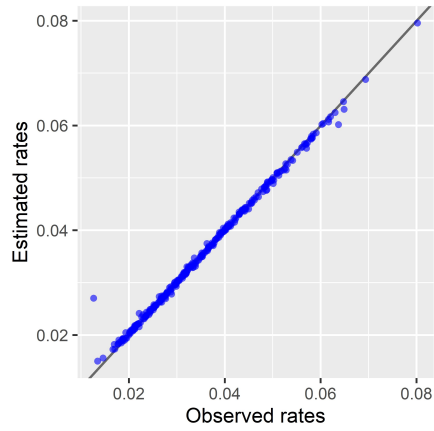
Figure S3. Scatterplots of the observed versus the predicted rates obtained for some of the fitted models (Continued).



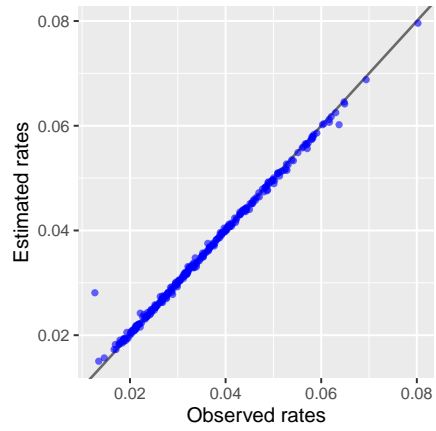
(g) *Model for the mobility matrix.*



(h) *Geometric mean model for the mobility matrix.*



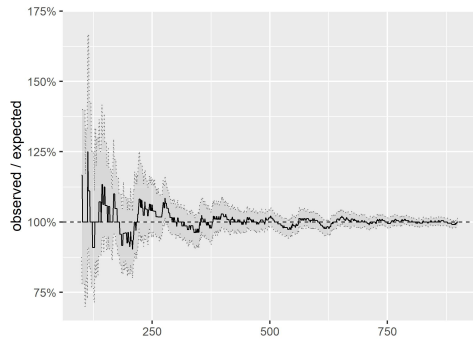
(i) *BYM2 model.*



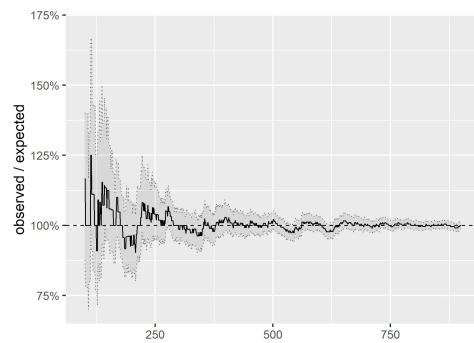
(j) *Leroux model.*

Figure S3. Scatterplots of the observed versus the predicted rates obtained for some of the fitted models (Continued).

7. Distributional assumptions in the fitted models

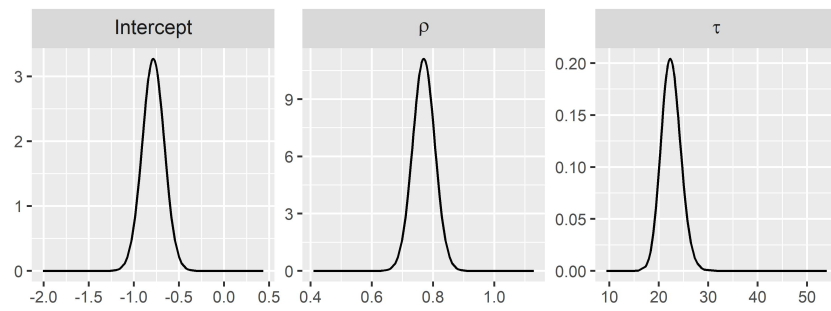


(a) Geometric mean model where the spatial matrix follows the contiguity of order one criterion.

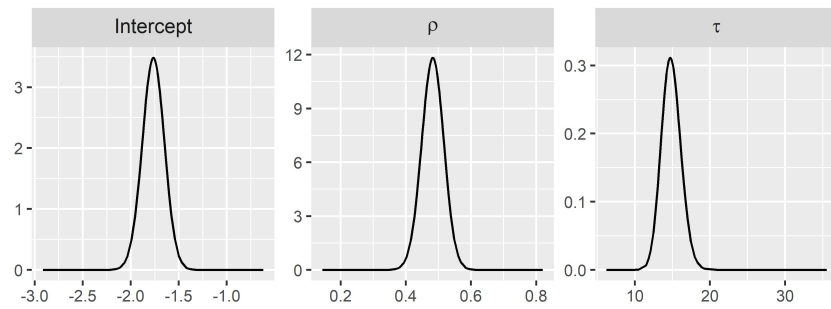


(b) Geometric mean model for the mobility matrix.

Figure S4. Distribution check for some of the fitted models.



(a) *Geometric mean model where the spatial matrix follows the contiguity of order one criterion.*



(b) *Geometric mean model for the mobility matrix.*

Figure S5. *Posterior densities from the parameters estimated for some of the fitted models.*

8. Review of the BYM2 and Leroux spatial models

In order to be able to provide some comparison of the proposed models with other commonly used models in the disease mapping literature, we have also considered the fitting of the BYM2 model (Riebler et al., 2016) and the Leroux spatial model (Leroux et al., 2000). In the BYM2 model, it is assumed that $(Y_i | v_i, \eta_i) \sim \text{Poi}(\mu_i)$, with conditional mean $E(Y_i | v_i, \eta_i) = \mu_i = P_i r_i$ following the regression structure:

$$\log(\mu_i) = \log(P_i) + \beta + \frac{1}{\sqrt{\tau_s}} \left(\sqrt{1 - \phi_s} v_i + \sqrt{\phi_s} \eta_i \right), \quad (\text{S3})$$

where v_i and η_i are unstructured normally and intrinsic conditionally autoregressive (ICAR) distributed random effects, respectively, but with variance scaled to approximately one. In addition, τ_s is a precision parameter that controls for the variance contribution from the sum of the two random effects and ϕ_s is a mixing parameter that captures the proportion of the variance explained by the spatially structured random effect. Note that $1 - \phi_s$ represents the proportion of the variance explained by the unstructured random effect.

For the parameters τ_s and ϕ_s , penalized complexity priors are generally assumed (Simpson et al., 2017). Comparison of the spatial conditional model with the BYM and the BYM2 models has been previously performed by Morales-Otero and Núñez-Antón (2021). Their results showed that, when compared to the spatial conditional model, they offered a similar fit in terms of information criteria. However, the BYM and BYM2 models did not provide additional information about the type and strength of spatial autocorrelation that was present in the data.

In the Leroux spatial model, it is assumed that $(Y_i | u_i) \sim \text{Poi}(\mu_i)$, with conditional mean $E(Y_i | u_i) = \mu_i = P_i r_i$ following the regression structure:

$$\log(\mu_i) = \log(P_i) + \beta + u_i, \quad (\text{S4})$$

where $\mathbf{u} = (u_1, \dots, u_n)$ is a set of spatially structured random effects. Unlike the BYM2 model, which includes two sets of random effects, one representing spatially structured variability and another for unstructured variability, the Leroux model uses a unified random effects structure. These effects follow a multivariate normal distribution, so that $\mathbf{u} \sim N\left(0, \tau_u^{-1} (\phi_u \mathbf{Q} + (1 - \phi_u) \mathbf{I})^{-1}\right)$, with \mathbf{Q} being the precision matrix derived from the spatial weights matrix following the contiguity of order one criterion and \mathbf{I} being the identity matrix. In addition, τ_u is the precision parameter of the random effects and ϕ_u is the spatial dependence parameter, which controls for the balance between the spatial structure of the effects, represented by the precision matrix, \mathbf{Q} , and the unstructured part, represented by the identity matrix, \mathbf{I} .

Fitting this model requires matrix manipulations, which can scale poorly with the number of spatial units, making it computationally intensive for large datasets. Values of $\hat{\phi}_u$ closer to zero indicate little to no spatial autocorrelation, while values closer to one

indicate strong spatial autocorrelation. Although $\hat{\phi}_u$ ranges from zero (purely unstructured) to one (purely spatially structured), its interpretation can be less intuitive when it is compared to models where the structured and unstructured effects are modelled explicitly as separate components, such as the spatial conditional or the BYM2 model.

Note that, in these models, the spatial neighbourhood structure that is usually assumed is the one based on contiguity of first order. Although some other structure might be specified, these models require the spatial matrix to be symmetrical (Wall, 2004; Lee, 2013). It may be useful to mention here that any matrix can be appropriately symmetrized by using well known and commonly used methods.

9. Effect of the geometric mean of the rates on the estimated rate

In this section, for the geometric mean model proposed in Section 2.2 in the manuscript, we provide a detailed analysis of the influence of the geometric mean of the rates on the estimated disease rate, where we consider different values for the spatial parameter estimate.

Note that equation (3) in the manuscript presents the formula for the geometric mean $\overline{\mathbf{Rates}}_i$ of the rates, where the spatial parameter $\hat{\rho}$ is not included (i.e., it does not depend on this estimated value) and, hence, the geometric mean does not change with respect to the estimated value of $\hat{\rho}$. In addition, we believe it is relevant to mention that, as described in Section 2.2, this parameter is a coefficient measuring the influence of the geometric mean of the neighbouring regions on the disease rate for a given region.

Let us start this reasoning within a very simple scenario where no fixed effects (covariates) and no unstructured heterogeneity (random effects v_i) are considered. In this case, we would have that $\hat{r}_i = \overline{\mathbf{Rates}}_i^{\hat{\rho}}$. In this case, when $\hat{\rho} = 1$, the rate in a given region is linearly related to the geometric mean of the rates in the neighbouring regions. However, we should note that, in general, there is also a fixed effects component and a residual error component in the model and, therefore, the element $\overline{\mathbf{Rates}}_i^{\hat{\rho}}$ would have a multiplicative effect of the rate for a given region. Figure S6 illustrates this specific effect and, in addition, other cases of particular interest in the proposed model.

Furthermore, based on the information included in Figure S6, we can analyse in more detail some specific cases for the possible estimated value of the spatial parameter ρ :

- When $\hat{\rho} = 0$, the multiplicative factor in \hat{r}_i is:

$$\overline{\mathbf{Rates}}_i^{\hat{\rho}} = 1,$$

and, therefore, in this case, the spatial variation is not described by the neighbourhood structure. This is illustrated by the solid blue line in Figure S6. In this case, there is no impact of the rate in the neighbouring region on the rate in the specific region under study and, thus, there would be no clear evidence of the existence of spatial autocorrelation.

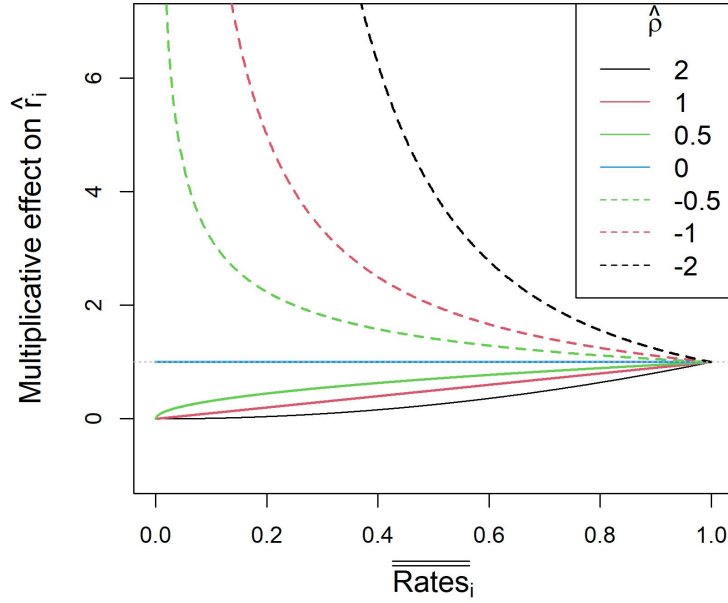


Figure S6. Visualization of the effect of the geometric mean of the rates on the estimated rate, considering different values for the spatial parameter estimate (i.e., $\hat{\rho}$).

- When $\hat{\rho} = 1$, this implies that there exists a linear relationship between the rate in a specific region and that in its neighbouring regions. Consequently, regions having a larger rate in neighbouring regions would also have a larger rate in the specific region under study, so that:

$$\hat{r}_i \propto \overline{\overline{\mathbf{Rates}_i}}$$

This specific case is illustrated by the solid red line in Figure S6.

- When $\hat{\rho} > 0$, the $\overline{\overline{\mathbf{Rates}_i}}^{\hat{\rho}}$ component would be increasing. This is illustrated with the solid green, red and black lines cases in S6, corresponding to $\hat{\rho} = 0.5$, $\hat{\rho} = 1$ and $\hat{\rho} = 2$, respectively. In a similar way, as was the case for $\hat{\rho} = 1$, this means that the rate in a specific region is expected to be smaller if the rate in neighbouring regions is smaller.
- When $\hat{\rho} \neq 1$, this relationship deviates from linearity. More specifically, when $\hat{\rho} < 1$, this implies that regions with lower neighbourhood rates are expected to have lower rates themselves, while medium and high neighbourhood rates are expected to have only a limited influence on the rate in the specific region under study (i.e., with a multiplicative effect close to 1). On the other hand, when $\hat{\rho} > 1$, specific regions are expected to have a higher rate when neighbouring regions

have higher rates, although with values lower than in the neighbouring regions. This latter case is illustrated by the black solid line in Figure S6 for $\hat{\rho} = 2$.

- Finally, when $\hat{\rho} < 0$, evidence of the existence of negative autocorrelation is obtained. That is, low rates in the neighbouring regions are associated with a large multiplicative factor and, therefore, a high rate in the specific region. These cases are illustrated by the dotted lines in Figure S6, for the specific cases of $\hat{\rho} = -0.5$, $\hat{\rho} = -1$ and $\hat{\rho} = -2$, respectively.

References

- Lee, D. (2013). CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, **55**(13), 1–24.
- Leroux, B.G., Lei, X. and Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (M.E. Halloran and D. Berry, eds.). Springer: New York, USA, 179–191.
- Morales-Otero, M. and Núñez-Antón, V. (2021). Comparing Bayesian spatial conditional overdispersion and the Besag-York-Mollié models: Application to infant mortality rates. *Mathematics* (Special issue on Spatial Statistics with its Applications), **9**(3), 282.
- Riebler, A., Sørbye, S.H., Simpson, D.P. and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, **25**(4), 1145–1165.
- Simpson, D.P., Rue, H., Riebler, A., Martins, T.G. and Sørbye, S.H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, **32**(1), 1–28.
- Wall, M.M. (2004) A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, **121**(2), 311–324.