SORT 49 (1) January-June 2025, 1-28 DOI:

Spatial autoregressive modelling of epidemiological data: Geometric mean model proposal

Mabel Morales-Otero[1,2], Christel Faes[3] and Vicente Núñez-Antón[4]

Abstract

We propose the geometric mean spatial conditional model for fitting spatial public health data, assuming that the disease incidence in one region depends on that of neighbouring regions, and incorporating an autoregressive spatial term based on their geometric mean. We explore alternative spatial weights matrices, including those based on contiguity, distance, covariate differences and individuals' mobility. A simulation study assesses the model's performance with mobility-based spatial correlation. We illustrate our proposals by analysing the COVID-19 spread in Flanders, Belgium, and comparing the proposed model with other commonly used spatial models. Our approach demonstrates advantages in interpretability, computational efficiency, and flexibility over the commonly used and previously existing methods.

MSC: *62J05, 62H11, 62M30, 92D30, 62F15.*

Keywords: *Bayesian approaches, COVID-19 incidence, Epidemiology, Spatial modelling.*

1. Introduction

The analysis of spatial data has become widely spread in epidemiology, specially because location can be an important surrogate for lifestyle, environment, as well as genetic

[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 Lehendakari Aguirre 83, 48015 Bilbao, Spain. vicente.nunezanton@ehu.eus

Received: May 2024.

Accepted: March 2025.

and other factors and, therefore, it can provide important insights for public health data analysis. Autoregressive models proposals for analysing spatial data include the Conditional Autoregressive (CAR) model, the auto-Poisson scheme (Besag, 1974) and the Simultaneous Autoregressive (SAR) model (Whittle, 1954), which incorporate the spatial correlation by assuming a conditional covariance structure for an unobservable component included in the regression structure. In addition, the spatial conditional overdispersion models include a spatial lag of the response variable in the regression model specification, which allows to capture the spatial dependence directly observed on neighbouring regions (see Cepeda-Cuervo, Córdoba and Núñez-Antón, 2018; Morales-Otero and Núñez-Antón, 2021). In the case of time series data, Zeger and Qaqish (1988) consider Poisson models that include the logarithm of the past counts in the log-mean regression specification, Knorr-Held and Richardson (2003) propose different autoregressive specifications when including the past counts and Held, Höhle and Hofmann (2005) propose an autoregressive model using an identity link.

An alternative to these models is given by spatial regression models for count data that make use of a spatially structured random effect, which is structured according to a given spatial weights matrix. In this context, two of the most popular models in spatial disease mapping are the Besag-York-Mollié (BYM) model (Besag, York and Mollié, 1991) and the BYM2 model (Riebler et al., 2016). The BYM model incorporates spatial dependence by means of two unobserved latent effects, namely a spatially unstructured random effect and a spatially structured random effect following an Intrinsic Conditional Autoregressive (ICAR) prior (Besag, 1974). In the BYM2 model the latent effect is a weighted average of these two random effects. Another random effects model frequently found in the literature is the Leroux model (Leroux, Lei, and Breslow, 2000). These models are generally estimated using Bayesian inferential methods.

In the aforementioned models, the relationship between two regions is described by a spatial weights matrix, for which several different specifications have been developed (see Anselin, 2002). In most cases, this matrix is fixed and previously specified, a choice that may have an impact on the results of the analysis. Therefore, it is very important for researchers to be able to study how to best describe the spatial structure of the data. Traditionally, spatial weights matrices are based on the adjacency of regions or on the distance among regions. However, there may be situations where the association is not given by the geographical proximity but, instead, it depends on some other connectivity structure or even on the specific characteristics of the regions under study.

In this sense, several authors have explored the use of different weights matrices. Earnest et al. (2007) studied the influence of different specifications of spatial weights matrices on the smoothing properties of the CAR model, obtaining considerable differences in the reported results, which provided a clear evidence about the importance of the proper choice of the spatial structure. In addition, Case, Hines, and Rosen (1993) proposed the use of a similarity matrix based on the inverse of the difference of the values that a given covariate takes in each region, which improved the performance of their fitted models. Ejigu and Wencheko (2020) proposed a weights matrix that took into ac-

count geographical proximity and covariate information simultaneously, which led to a better justification and motivation of the spatial structure present in the data under study.

After the beginning of the pandemic, several authors concentrated their efforts on the different statistical modelling proposals to study COVID-19 data. For example, Sahu and Böhning (2022) proposed a joint spatio-temporal model to analyse the weekly number of cases and deaths related to COVID-19, also presenting different specifications for the spatial and temporal random effects. Konstantinoudis et al. (2022) analysed the weekly number of deaths for several regions in Europe during the period going from 2015 to 2019, fitting a hierarchical Poisson model with a BYM2 specification to these data, thus, being able to evaluate the excess of mortality during the COVID-19 pandemic. Fritz et al. (2022) proposed a Poisson autoregressive model similar to the one in Held et al. (2005), and analysed data from Germany on COVID-19 infections, hospitalizations and intensive care units occupation. Additional references include D'Angelo, Abbruzzo, and Adelfio (2021), Johnson, Ravi, and Braneon (2021) and Natalia et al. (2022), among others. Furthermore, purely spatial approaches have also been used, such as the proposals in Konstantinoudis et al. (2021), where they studied the relationship between COVID-19 related deaths and long-term exposure to air-pollution, fitting a BYM2 model to data concerning the first wave of the disease in England. Other researchers have used the mobility of individuals among regions as a connectivity structure for modelling COVID-19 data. For example, Slater et al. (2022) combined geographical proximity and human mobility data on the BYM specification to spatially model COVID-19 case counts in the regions of Castilla-León and Madrid in Spain from March to June 2020.

In this paper, we propose a geometric mean extension of the spatial conditional models in Cepeda-Cuervo et al. (2018) and Morales-Otero and Núñez-Antón (2021) to account for the spatial autocorrelation that may be present in the data. The spatial conditional model is described in Section 2.1, and the extension is motivated and introduced in Section 2.2. Additionally, we also investigate the use of several spatial weights matrices in the computation of the spatial lag and propose some new possible structures to be implemented, which are discussed in Section 2.3. A simulation study is included in Section 3. The usefulness of our methodological proposals and their comparison with other commonly used spatial models is provided in Section 4. More specifically, a comparison with the BYM2 and Leroux spatial models is included in Section 4.3. In Section 5, we end with a discussion.

2. Methodology

This section reviews the spatial conditional overdispersion models proposed in the literature. Thereafter, we propose an extension of this model and discuss possible weights matrices that could describe the underlying spatial dependency structure.

2.1. Review of the spatial conditional model

The spatial conditional overdispersion models were developed to fit spatial count data, allowing to capture overdispersion and to explain the spatial dependence that may exist in the data, as suggested by Cepeda-Cuervo et al. (2018). These authors assume that the dependent variable Y_i, for regions $i = 1,\ldots,n$, follows a conditional distribution $f(y_i \mid y_{\sim i})$, where y_i represents the observed count in region i and, $y_{\sim i}$, the values in all of the neighbouring regions of the i-th region (without including the i-th region itself). A spatial autoregressive term, more specifically, the lag of the response variable, is incorporated in the regression model specification for the conditional mean $\mathrm{E}(Y_i \mid Y_{\sim i})$. The inclusion of such spatial dependence in the model can explain part of the overdispersion.

In an epidemiological context, interest often goes towards the modelling of the rates of a disease. In this case, Morales-Otero and Núñez-Antón (2021) assumed that the conditioned response variable $(Y_i \mid Y_{\sim i}, \nu_i)$, the total number of cases for $i = 1,\ldots,n$, follows a Poisson distribution, with conditional mean μ_i, so that $\mathrm{E}(Y_i \mid Y_{\sim i}, \nu_i) = \mu_i = \mathrm{P}_i r_i$. Here, P_i represents the population size and r_i the disease rate in the i-th region, for $i = 1,\ldots,n$. They proposed the following regression structure for the conditioned means:

$$\log(\mu_i) = \log(\mathrm{P}_i) + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \rho\mathbf{W}_i\mathbf{Rates} + \nu_i, \tag{1}$$

where an autoregressive component is included for the rates, (i.e., $\mathbf{W}_i\mathbf{Rates} = \sum_{j=1}^{n} w_{ij}\mathrm{Rates}_j$), which is a weighted average of the observed rates $\mathrm{Rates}_j = y_j/\mathrm{P}_j$, with weights specified by the spatial weights matrix $\mathbf{W}$. Here, $\mathbf{x}_i$ is a $1 \times p$ vector of explanatory variables for the i-th observation, $\boldsymbol{\beta} \in \Re^p$ a $p \times 1$ vector of unknown regression parameters that need to be estimated and $\rho \in \Re$ the unknown spatial autoregressive parameter. These parameters and variables belong to the set of all real numbers, as no constraints are imposed. In addition, a normally distributed random effect $\nu_i \sim N(0,\tau)$, with $\tau > 0$, is included to allow for additional unstructured overdispersion in the counts. Note that the assumed spatial structure is given by the matrix $\mathbf{W}$, where its elements, w_{ij}, are weights that represent the strength of the relationship between regions i and j. Section 2.3 includes a detailed description about the different ways these weights can be defined.

2.2. Geometric mean spatial conditional model

Zeger and Qaqish (1988) proposed several models to account for temporal autocorrelation in time series data, including one for count data, where they suggested the use of a Poisson model that incorporates the logarithm of the past counts in the regression model for the logarithm of the mean instead of the past counts. Knorr-Held and Richardson (2003) proposed the use of the term $\log(y_{t-1}+1)$ in order to overcome the issue of the nonexistence of the logarithm, so that it is equal to zero when there are no cases. Held et al. (2005) proposed to regress the mean directly on the past counts instead, but assuming an identity link.

Following the ideas in Zeger and Qaqish (1988) and Knorr-Held and Richardson (2003), we propose the following geometric mean spatial conditional model for count data. As before, we assume a Poisson model for the conditioned response outcomes, that is $(Y_i \mid Y_{\sim i}, \nu_i) \sim \text{Poi}(\mu_i)$, with conditional mean $\text{E}(Y_i \mid Y_{\sim i}, \nu_i) = \mu_i = \text{P}_i\text{r}_i$, following the regression model:

$$\log(\mu_i) = \log(\text{P}_i) + \mathbf{x}_i^{\text{T}}\boldsymbol{\beta} + \rho \mathbf{W}_i \log(\mathbf{Rates}) + \nu_i \tag{2}$$

Here, we believe it is important to mention that, in the presence of zero counts, it would be necessary to use $\log(y_j + 1)$ and $\log(P_j + 1)$ when computing the observed rates (see equation(1)). This model closely resembles the model in equation (1), but here the autoregressive component is a weighted average of the logarithms of the rates, instead of the rates. It can be easily seen that the smoothed estimates of the rates are estimated as:

$$\begin{aligned} \widehat{r_i} &= \exp(\mathbf{x}_i^{\text{T}}\hat{\boldsymbol{\beta}}) \exp\left(\frac{1}{n_i}\sum_{j=1}^{n} w_{ij}^{*} \log(\text{Rates}_j)\right)^{\hat{\rho}} \exp(\nu_i) \\ &= \exp(\mathbf{x}_i^{\text{T}}\hat{\boldsymbol{\beta}}) \overline{\overline{\mathbf{Rates}}}_i^{\hat{\rho}} \exp(\nu_i), \end{aligned} \tag{3}$$

with w_{ij}^{*} representing the non-standardized spatial weights, n_i being the number of neighbours of region i, and $\overline{\overline{\mathbf{Rates}}}_i$ being the geometric mean of the rates included in the vector of the observed rates **Rates**. Note that the geometric mean of a sample $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is defined as $(\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$, which can also be expressed as $\exp\left[\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right]$, when $x_i > 0$, for $i = 1, \ldots, n$.

This can also be generalized to the case where the spatial weights matrix is given by some criterion where the weights w_{ij} are not necessarily equal to 0 or 1. This could be the case, for example, in cases where we use criteria based on distance among regions or on the mobility matrix. In these cases, we would have a weighted geometric mean of the rates included in the vector of the observed rates, so that:

$$\overline{\overline{\mathbf{Rates}}}_i = \exp\left(\frac{\sum_{j=1}^{n} w_{ij}^{*} \log(\text{Rates}_j)}{\sum_{j=1}^{n} w_{ij}^{*}}\right) \tag{4}$$

Therefore, the estimated value obtained for the spatial parameter ρ would represent how the incidence rate in the regions resembles the (weighted) geometric mean of the rates in their neighbours. Consequently, the use of the logarithm of the rates in the autoregressive component has an important epidemiological interpretation. For a better understanding of this effect, in Section 9 of the supplementary material, we have included a detailed description and better motivation of the effect of the geometric mean of the rates on the estimated disease rate, considering different values of the estimated spatial parameter.

2.3. Spatial weights matrices

As already stated, the models proposed in Section 2 do not impose or need any restrictions when specifying the spatial weights matrix and, therefore, they are very flexible,

allowing the use of a wide range of spatial structures. Moreover, this flexibility makes them a valuable tool for exploring different spatial weights matrices in a specific dataset. This section discusses different possible choices for specifying the weights w_{ij} used in the proposed model in equation (2).

2.3.1. Spatial weights matrices based on contiguity

The spatial structure based on contiguity or adjacency is defined by the spatial weights matrix $\mathbf{W}$, where $w_{ij} = 1$, if region i is adjacent or a neighbour to region j, and $w_{ij} = 0$, otherwise. Different criteria can be assumed to specify whether two regions are adjacent. For example, the Queen contiguity criteria assumes that regions i and j are neighbours if they share at least one point in their boundaries. Most commonly the spatial weights matrix is standardized by rows, so that if region i is adjacent to region j, then $w_{ij} = 1/n_i$, where n_i is the number of neighbours region i has. In this way, the spatial lag $\mathbf{W}_i\mathbf{y}$ can be viewed as a spatial average of the values that the variable takes in all of its neighbouring locations.

First order contiguity is specified when we consider that regions i and j are neighbours if they share at least one point in their boundaries. This specification is also known as Queen contiguity criterion. Extending this criteria by considering that i and j are neighbours if they share a common neighbour, we can define second order contiguity. Third order contiguity can be specified the same way, when it is assumed that regions i and j are adjacent if they share a common neighbour of order two. Contiguity of higher order is also possible to specify by following these ideas.

2.3.2. Spatial weights matrices based on distance

An alternative way to define a spatial structure is to consider a spatial weights matrix where its elements are defined as a function of the distance among the central points of the polygons representing the regions, called the centroids, s_i $(i = 1,\ldots,n)$. Inverse distance weights are specified as $w_{ij} = 1/\|s_i - s_j\|$, with $\|s_i - s_j\|$ being the Euclidean distance between regions i and j. In addition, in the negative exponential criteria the weights are defined so that $w_{ij} = \exp(-\|s_i - s_j\|)$.

Finally, we can also define the distance band weights, with band width given by a critical threshold h. In particular, it is considered that regions i and j are neighbours if their centroid lies within the chosen band. Let s_i be the centroids of the regions under study, for a given threshold h, then $w_{ij} = 1$ if the Euclidean distance between s_i and s_j is smaller than h, that is $\|s_i - s_j\| < h$, and 0 otherwise.

2.3.3. Covariate-based similarity (or difference) matrices

Ejigu and Wencheko (2020) proposed a weights matrix $\mathbf{W}$, which not only takes into account geographical proximity, but also a specific covariate's information. Given an environmental variable e_i $(i = 1,\ldots,n)$ for n regions with centroids s_i, they define the following structure for the weights:

$$w_{ij} = \exp\{-[\alpha|e_i - e_j| + (1-\alpha)\|s_i - s_j\|]\}, \tag{5}$$

where α is a previously selected fixed value between zero and one, $|e_i - e_j|$ is the absolute difference in the value of the environmental covariate between regions i and j and $\|s_i - s_j\|$ is the Euclidean distance between the centroids of regions i and j. The elements in the diagonal of this matrix are zero and it is row standardized. As α approaches zero, the weights give more relevance to the geographical distance, and, when it approaches one, the covariate differences receive more importance.

Following this idea, we also propose an alternative covariate-based similarity matrix, where we will consider both environmental and socio-economic variables to impact the weight amongst regions. Let $\mathbf{W}$ be a traditional weights matrix based on contiguity, distance, or any other criteria, with elements w_{ij}, and $\mathbf{D}$ an $n \times n$ matrix with elements $d_{ij} = 0$ if $i = j$ and:

$$d_{ij} = \exp(-|e_i - e_j|), \text{ for } i \neq j, \tag{6}$$

We then propose the use of the matrix $\mathbf{W} \circ \mathbf{D}$, which is the Hadamard (or element-wise) product of matrices $\mathbf{W}$ and $\mathbf{D}$. In this way, small weights are given to neighbouring regions with large differences in the values of the covariate and to distant regions, while large weights are given to neighbouring regions with similar covariate information and that are geographically close to each other.

A potential concern might arise regarding whether specifying covariate-based similarity matrices in the model described by equation (2), while also including these covariates as independent variables, could lead to endogeneity problems. As discussed by Case et al. (1993), when the weights matrix $\mathbf{W}$ is constructed based on similarities or differences in covariates between municipalities, and the vector of observations for the covariates captures within-municipality variations, this design ensures that the elements of the weights matrix are orthogonal to the explanatory variables. Therefore, by construction, this approach eliminates any induced correlation between the covariates and the error term, thus addressing potential endogeneity issues.

2.3.4. Mobility matrix

The previous proposals presented here for the weights matrices are a representation of how close (in space) and/or how similar (in terms of covariate information) regions are. Another characteristic to define the weights matrix is to assess how much contact there was amongst individuals in the different regions. This is of special interest when considering, for example, an outcome that depends on the contact behaviour, such as is the case in infectious disease incidence. As a proxy for the contact behaviour, and based on mobile phone data, the mobility amongst regions can be used. That is, each element m_{ij} in the mobility matrix $\mathbf{M}$ is defined as the mean proportion of time that people from region i have spent in region j in a given time period. This matrix would then clearly represent a different type of connectivity structure among regions.

2.4. Model estimation and selection

All models considered here are fitted using the integrated nested Laplace approximation (INLA) approach, in the R-INLA package. It should be noted, however, that, in general,

any software methodology that allows for estimation of a generalized linear mixed model can be used to implement this model. This is a great advantage of the proposed method, as one is not restricted to complex estimation tools for fitting spatial models.

In addition, it could be worth addressing the potential risk of spatial confounding in the proposed geometric mean spatial model. Spatial confounding arises when covariates share similar spatial patterns with unobserved spatial processes or random effects. In our model, however, the spatial lag of the logarithm of the observed rates is used as an explanatory variable, directly incorporating the observed spatial structure. Since no additional spatially structured random effects are employed and the spatial structure is assumed to be fully observed, the model theoretically mitigates the issue of spatial confounding. The spatial dependence is captured through the geometric mean of neighbouring observations, minimizing the risk of confounding spatial random effects with covariate effects.

Model comparison is carried out by using the Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2010), where the smallest values indicate the best fitting model. Additionally, we also use the Conditional Predictive Ordinate (CPO) diagnostic (Pettit, 1990), which is a leave-one-out predictive measure. More specifically, for each observation i, the CPO_i is computed, so that it reflects the posterior probability of observing that value, given the other observations. In this way, we would be able to compute a global value by using the sum of the logarithms for the resulting CPO_i values (i.e., $\text{CPO} = -\sum_{i=1}^{n} \log(\text{CPO}_i)$) As in the case of the WAIC, the model with the smallest CPO value would be considered as the best fitting one.

Furthermore, these model selection criteria ensure a balance between model fit and complexity by penalizing overly complex models, helping in this way to prevent possible overfitting. To further assess the model's generalizability, cross-validation techniques like CPO evaluate predictive performance by measuring how well the model generalizes to unseen data. This approach ensures that the model does not overfit the observed data and is capable of making accurate predictions under new scenarios.

In addition, for all the estimated parameters, noninformative prior distributions are assumed. In particular, for the fixed effects and for the precision parameters, we assume independent normal $N(0, 1 \times 10^5)$ distributions, and gamma $G(1 \times 10^{-4}, 1 \times 10^{-4})$ distributions, respectively.

3. Simulation study

As already mentioned in Section 1, most spatial modelling applications make use of a spatial weights matrix following traditional criteria, such as the ones based on contiguity or distance among the regions. However, in this section, we wish to assess the performance of our proposed models in the selection of the weights matrix, as well as to study the sensitivity of the parameters to a misspecified neighbourhood matrix.

Therefore, we have carried out a simulation study, where we induce correlation in the response variable following the mobility matrix structure. For this purpose, we have

implemented a Gibbs sampling algorithm, which allowed us to generate spatially autocorrelated Poisson data by repeatedly sampling from conditional distributions (see Jackson and Sellers, 2008). In our specific case, we define a set of initial values for the parameters β, ρ and τ and, on each iteration, we draw Poisson samples, where the mean is conditioned on the values of the previous iteration. Additional details on the algorithm described below can be found in Section 2 in the supplementary material. We would like to remark that the data have been simulated using the mobility matrix provided in the dataset corresponding to the COVID-19 cases in the municipalities of Flanders, which will be analysed in Section 4. Moreover, this spatial structure has also been used to obtain the contiguity of order one and the inverse distance spatial weights matrices.

We have defined twelve different scenarios, given the true values for the parameters, which can be consulted in the first column to the left in Table 1. For each case, we have simulated $S^* = 500$ datasets (with the number of regions $n = 300$), and discarded half of them, so that $S = 250$ simulations for each scenario remained. Model (2) has been fitted to each of the simulated dataset, considering three different specifications for the spatial structure, one using the mobility matrix to compute the spatial lag, another one using the contiguity of order one spatial weights matrix, and a third one using the inverse distance spatial weights matrix. In addition, we have also fitted the BYM2 and Leroux models, both using the contiguity of order one spatial weights matrix, which is the standard specification for such models. For further details about these models, refer to Section 8 in the supplementary material.

Table 1 reports the bias, mean squared error (MSE) and the coverage of the estimates obtained from fitting each model to the simulated datasets. This table includes only the results obtained for the geometric mean model using the three different spatial weights structures mentioned above. The BYM2 and Leroux models have different formulations and produce estimates that are not directly comparable to those of the geometric mean model. Therefore, they are excluded from this analysis and will be considered later when evaluating and comparing the models' goodness of fit in the specific dataset under study.

For the scenarios where the parameters' true values were $\beta = -2$ and $\rho = 0.5$ (i.e., first two scenarios), the smallest bias was obtained for the estimations for the model using the mobility matrix, indicating that this is the model where the resulting estimates are closer to the true values of the parameters. In these scenarios the coverage percentages in the models using the mobility matrix are also the largest, indicating that most of the credible intervals of the estimated parameters in these models contain the true values. However, when the true value for β changed to -0.5 (i.e., third and fourth scenarios), the smallest bias and the best coverage were obtained for the model using the contiguity criterion for the weights matrix, which seems to suggests that the value given to the intercept β is having a significant impact on the results. This substantial influence can be attributed to the fact that the model does not include any covariates apart from the offset (i.e., the logarithm of the population in each municipality), the intercept itself, and the spatial lag of the logarithm of the rates. Consequently, the intercept determines the baseline level of expected counts across municipalities, directly influencing the scale of

the predicted counts, the overall variability, and the relative contribution of the spatial lag term.

In the scenarios where the true value for ρ is set to 0.2, the bias of the estimates considerably increases when using the mobility matrix. In fact, the estimations with smallest bias are obtained for the model using the inverse distance criterion for the spatial matrix and, moreover, the coverage is very high for all the models. This can be due to the fact that here we are setting a small value for the spatial parameter and, thus, forcing the mobility connectivity structure to have a smaller relevance in the simulated data.

In addition, for the parameters' true values $\beta = -2$ and $\rho = 0.9$, the smallest bias of the estimates and the best coverage were also obtained for the mobility matrix. Given that, in this case, we are setting a large value for the spatial parameter, more relevance is given to this structure. However, when $\beta = -0.5$ and $\rho = 0.9$, the models using the contiguity and the mobility matrix produce similar values for the bias of the estimations and the coverage for the contiguity matrix highly improves, meaning that, for this specific setting, the spatial structure is not so clearly defined.

The sensitivity of the results to small variations in the parameter ρ is due to the absence of additional covariates or effects in the model, making, therefore, the spatial autoregressive term the primary source of variation. The corresponding result would be that even minor adjustments to ρ can significantly influence the dynamics of the overall model and the resulting estimates.

Finally, from the results included in Table 1, for the precision parameter τ, no significant changes were observed in the bias of the estimations or in the coverage when changing this parameter's value from 5 to 15.

Regarding the predictive accuracy of the models, we can evaluate it by computing the mean squared predictive error (MSPE) of the simulated rates for each simulated dataset $\left[\text{MSPE}_s = \sum_{i=1}^{n}(r_i^{(s)} - \hat{r}_i^{(s)})^2/n\right]$ (Carroll et al., 2015). In this way, we can obtain an average for the model fitted for each of the 250 datasets generated for each scenario, so that $\text{MSPE} = \sum_{\text{s}=1}^{\text{S}}(\text{MSPE}_\text{s})/\text{S}$. Note that the models with the lowest values of the MSPE would be considered as the best fitting ones. The results obtained are included in Table 2, where we can see that, in general, the MSPE is small in every scenario, but the smallest values are mostly obtained for the models in which the mobility matrix was used to compute the spatial lag of the log-rates.

Moreover, we have counted the number of times that the information criteria values were smaller in each case so that we can check how many times the "correct" model was selected as the best fitting one. Most of the times, with a very few exceptions, the model where the mobility matrix was used, was selected with the smallest WAIC and CPO values. This indicates that we can indeed, based on the model selection criteria, select the underlying true neighbourhood matrix. These results are included in Table S1 in the supplementary material.

From the results obtained in the simulation study, we can conclude that it is essential to evaluate whether the spatial structure used in a study is the most adequate one. For most of the spatial modelling applications, the spatial weights matrix employed to de-

Table 1. *Results obtained from the models using different weights matrices, fitted to the simulated datasets.*

Fitted model:		Mobility			Contiguity			Inverse distance		
True value		β	ρ	τ	β	ρ	τ	β	ρ	τ
$\beta = -2$,	Bias	0.172	0.044	0.806	-0.718	-0.187	-6.863	12.064	3.049	-6.362
$\rho = 0.5$,	MSE	0.031	0.002	0.846	0.524	0.036	47.160	145.754	9.312	40.547
$\tau = 15$	Coverage	97%	96%	100%	6%	3%	0%	0%	0%	0%
$\beta = -2$,	Bias	0.347	0.088	0.319	-0.487	-0.130	-1.303	10.892	2.751	-1.217
$\rho = 0.5$,	MSE	0.123	0.008	0.113	0.245	0.017	1.705	118.791	7.577	1.489
$\tau = 5$	Coverage	84%	82%	100%	68%	67%	0%	0%	0%	0%
$\beta = -0.5$,	Bias	0.232	0.233	0.406	0.037	0.027	-0.219	2.623	2.639	-0.296
$\rho = 0.5$,	MSE	0.054	0.054	0.166	0.001	7.769e-04	0.049	6.883	6.966	0.088
$\tau = 5$	Coverage	0%	0%	100%	100%	100%	100%	0%	0%	100%
$\beta = -0.5$,	Bias	0.187	0.189	1.181	-0.010	-0.018	-1.285	2.612	2.632	-1.367
$\rho = 0.5$,	MSE	0.035	0.036	1.409	2.215e-04	4.489e-04	1.669	6.824	6.932	1.886
$\tau = 15$	Coverage	0%	0%	100%	100%	100%	100%	0%	0%	100%
$\beta = -2$,	Bias	0.138	0.055	0.855	-0.539	-0.219	-0.318	0.224	0.088	-0.317
$\rho = 0.2$,	MSE	0.020	0.003	0.783	0.293	0.048	0.153	0.282	0.045	0.153
$\tau = 15$	Coverage	100%	100%	100%	0%	0%	100%	100%	100%	100%
$\beta = -2$,	Bias	0.229	0.092	0.307	-0.501	-0.203	0.045	-0.066	-0.029	0.045
$\rho = 0.2$,	MSE	0.054	0.009	0.097	0.252	0.042	0.005	0.130	0.021	0.005
$\tau = 5$	Coverage	100%	100%	100%	68%	60%	100%	100%	100%	100%
$\beta = -0.5$,	Bias	0.083	0.131	0.308	-0.095	-0.156	0.171	-0.071	-0.118	0.169
$\rho = 0.2$,	MSE	0.007	0.017	0.095	0.009	0.024	0.030	0.006	0.017	0.029
$\tau = 5$	Coverage	100%	100%	100%	100%	100%	100%	100%	100%	100%
$\beta = -0.5$,	Bias	0.082	0.130	0.928	-0.104	-0.171	0.440	-0.058	-0.095	0.438
$\rho = 0.2$,	MSE	0.007	0.017	0.869	0.011	0.029	0.203	0.007	0.017	0.201
$\tau = 15$	Coverage	100%	100%	100%	100%	100%	100%	100%	100%	100%
$\beta = -2$,	Bias	0.106	0.012	110.710	0.584	0.117	-14.665	63.681	7.041	-14.651
$\rho = 0.9$,	MSE	0.013	2.227e-04	38995.408	0.640	0.018	215.050	4085.262	49.935	214.643
$\tau = 15$	Coverage	84%	92%	68%	60%	48%	0%	0%	0%	0%
$\beta = -2$,	Bias	0.191	0.020	13.815	0.410	0.093	-4.615	61.910	6.859	-4.664
$\rho = 0.9$,	MSE	0.039	4.853e-04	275.350	0.506	0.013	21.701	3860.889	47.386	21.756
$\tau = 5$	Coverage	89%	93%	67%	69%	59%	0%	0%	0%	0%
$\beta = -0.5$,	Bias	0.353	0.075	0.327	0.295	0.054	-2.981	15.885	3.387	-3.219
$\rho = 0.9$,	MSE	0.125	0.006	0.132	0.089	0.003	8.891	252.426	11.474	10.362
$\tau = 5$	Coverage	0%	0%	100%	100%	100%	0%	0%	0%	0%
$\beta = -0.5$,	Bias	0.174	0.037	0.647	0.194	0.033	-11.888	15.331	3.287	-12.256
$\rho = 0.9$,	MSE	0.031	0.001	0.819	0.040	0.001	141.331	235.405	10.822	150.208
$\tau = 15$	Coverage	43%	45%	100%	100%	100%	0%	0%	0%	0%

Table 2. *Average of the MSPE values obtained from the models using different weights matrices, fitted to the simulated datasets.*

True values	Mobility	Contiguity	Inverse distance	BYM2	Leroux
$\beta = -2, \rho = 0.5, \tau = 15$	2.323e-06	5.173e-05	5.601e-05	1.977e-05	1.973e-05
$\beta = -2, \rho = 0.5, \tau = 5$	9.659e-07	2.994e-05	3.190e-05	1.092e-05	1.042e-05
$\beta = -0.5, \rho = 0.5, \tau = 5$	1.765e-06	1.223e-05	1.212e-05	5.402e-06	5.622e-06
$\beta = -0.5, \rho = 0.5, \tau = 15$	2.292e-06	4.135e-05	4.204e-05	1.696e-05	1.724e-05
$\beta = -2, \rho = 0.2, \tau = 15$	2.764e-06	1.793e-05	1.796e-05	1.555e-05	1.554e-05
$\beta = -2, \rho = 0.2, \tau = 5$	1.683e-06	1.004e-05	1.005e-05	8.350e-05	8.402e-05
$\beta = -0.5, \rho = 0.2, \tau = 5$	3.518e-06	6.521e-06	6.540e-06	5.322e-06	5.516e-06
$\beta = -0.5, \rho = 0.2, \tau = 15$	5.497e-06	1.594e-05	1.598e-05	1.291e-06	1.336e-06
$\beta = -2, \rho = 0.9, \tau = 15$	6.441e-06	9.049e-06	9.917e-06	1.735e-05	1.852e-05
$\beta = -2, \rho = 0.9, \tau = 5$	4.254e-06	9.247e-06	9.594e-06	1.714e-05	2.275e-06
$\beta = -0.5, \rho = 0.9, \tau = 5$	8.752e-07	5.207e-05	4.737e-05	2.345e-05	6.582e-06
$\beta = -0.5, \rho = 0.9, \tau = 15$	2.551e-06	1.071e-04	9.862e-05	1.788e-05	1.482e-05

scribe the spatial structure of the data under study is the one following the contiguity of order one criterion. However, we believe it has been clearly shown that this is not always necessarily the best choice.

In this specific study, it has been shown that when the mobility matrix is the underlying structure, and the model is misspecified, in general, the bias of the estimations is larger than the bias obtained for the model using the mobility matrix. Moreover, information criteria values such as the WAIC and CPO and, also predictive accuracy measures such as the MSPE, have favoured the correctly specified model, selecting it as the best fitting one in almost all cases. Overall, the simulation study illustrates the fact that our proposed model effectively identifies the correct spatial structure when properly specified. However, this does not imply that our model is the correct or best model under a given setting, which was not the original purpose in the simulation study.

4. Illustration of methodology

4.1. Data Exploration

We investigate the spatial distribution of COVID-19 from September 2020 until January 2021 amongst the Flemish municipalities. Figure 1 shows the observed incidence of COVID-19 per 100,000 inhabitants in Flanders' municipalities in the time period con-

sidered, which was the time of the second wave in Belgium. It can be observed that not all municipalities presented the same impact in the second COVID-19 wave. In this study, we wish to assess whether the spatial correlation pattern of the incidence during the second wave of the disease was linked to any social demographics.

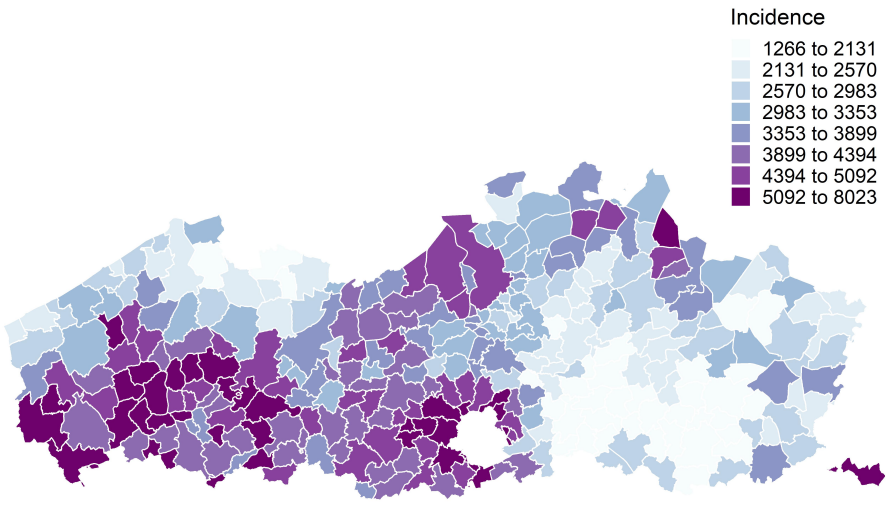


Figure 1. *Spatial distribution of the incidence of COVID-19 per 100,000 inhabitants in Flanders' municipalities from 2020-09-01 to 2021-01-31.*

The data under analysis includes information on the 300 municipalities of Flanders in Belgium, which is available at the website of the Belgian Institute for Public Health (Sciensano) (https://epistat.wiv-isp.be/covid/). Table 3 includes some descriptive statistics for the variables available across municipalities. The outcome of interest is the number of confirmed COVID-cases from 2020-09-01 to 2021-01-11, summarized by the variable `N.cases`. The population size in the municipality is denoted as `P`, and `incidence` is the number of COVID-19 cases in this time period per 100,000 inhabitants. There are also two additional variables available which can be considered as proxies for the socio-economic status and demography of the municipality. These are the percentage of households with a discount on the electricity meter (`budgetmeters`) and the percentage of single-parent households (`single_house`).

4.2. Model Estimates

The COVID-19 incidence map in Figure 1 suggests the presence of spatial autocorrelation in the data, as municipalities with similar values of COVID-19 incidence are grouped together in space. Therefore, in order to be able to characterize the spatial pattern of the second wave of COVID-19 in the data under study, we will further analyse these data by implementing spatial models that account for the spatial dependence. In addition, since we have seen the importance of testing different spatial structures for the

Table 3. *Descriptive statistics for the variables available across municipalities*

	Median	Mean	SD	Min.	Max.
N.cases	544.000	801.243	1570.136	1.000	24387.000
P	15036.500	22097.143	36156.404	79.000	529247.000
incidence	3358.868	3564.184	1235.740	1265.823	8023.070
single_house	7.885	8.021	1.245	5.300	15.300
budgetmeters	1.125	1.229	0.678	0.000	6.860

weights matrix with the simulation study carried out in the previous section, here we will also explore different possible choices for the spatial weights matrix to be included in the fitted models.

We fit both the spatial conditional normal Poisson model in equation (1) (Section 2.1) and the proposed geometric mean spatial conditional normal Poisson model in equation (2) (Section 2.2). As one of our objectives is to be able to select the spatial structure that best accommodates the spatial underlying process in the data, we use the different weights matrices described in Section 2.3, and compare the fitting of the different models by using their WAIC and CPO values. Note that, in this specific application, we do not include any covariates in the linear predictor, as we focus on the spatial modelling by means of the autoregressive terms and on the comparison of the performance of such models.

We believe it is important to mention the fact that, at the beginning of this research, the variables available were included in the model as covariates. However, the results obtained suggested that they did not offer any improvements in models' fitting in terms of information criteria. Therefore, in this specific study, we decided to only employ them when computing the proposed weights matrices based on similarities. It should also be noted that, in this study, we do not aim to identify any risk factor in the spreading of the infection, but to investigate the spatial correlation that may exist in the data and find the structures that best accommodate it.

The results obtained for the fitting of these models are included in Tables 4 and 5, which were fitted by considering ten different options for the weights matrix. First, we have used the spatial weights matrices based on the adjacency among municipalities (contiguity of first and third order). Second, weights were based on the distance among the centroids of the municipalities (inverse distance, negative exponential distance and distance band method). Third, the weights matrices were based on the product between covariate differences and traditional spatial weights, as proposed in Section 2.3. For these similarity matrices, the spatial weights matrices considered are the ones based on contiguity or first order, and that based on the distance band. The variables used to measure whether municipalities have a similar socio-economic status are single_house and budgetmeters. Finally, the mobility matrix was also considered. The heatmaps for these matrices are included in Figure S1 in the supplementary material, where we can clearly see the different structures they represent.

Table 4. *Results obtained after fitting the spatial conditional normal Poisson models to the COVID-19 incidence data in Flanders, for the different weights matrices considered.*

Weights matrix			$\hat{\beta}$	$\hat{\rho}$	$\hat{\tau}$
Contiguity of order 1	WAIC = 2947.7	Mean	-4.378	27.928	27.830
	CPO = 1807.4	SD	(0.041)	(1.121)	(2.440)
		95% CI	(-4.459,-4.297)	(25.728,30.129)	(23.287,32.868)
Contiguity of order 3	WAIC = 2942.7	Mean	-4.425	29.220	18.100
	CPO = 1868.4	SD	(0.060)	(1.637)	(1.542)
		95% CI	(-4.542,-4.307)	(26.006,32.434)	(15.217,21.271)
Inverse distance	WAIC = 2941.2	Mean	-5.649	63.083	19.319
	CPO = 1859.3	SD	(0.120)	(3.333)	(1.646)
		95% CI	(-5.885,-5.413)	(56.538,69.629)	(16.242,22.705)
Negative exponential	WAIC = 2941.4	Mean	-6.006	73.132	12.660
	CPO = 1921.9	SD	(0.221)	(6.152)	(1.062)
		95% CI	(-6.440,-5.573)	(61.049,85.214)	(10.672,14.843)
Distance band	WAIC = 2938.4	Mean	-4.514	31.480	24.663
	CPO = 1822.6	SD	(0.051)	(1.370)	(2.120)
		95% CI	(-4.613,-4.415)	(28.790,34.172)	(20.706,29.031)
$W \circ D$ single_house and	WAIC = 2946.9	Mean	-4.349	27.121	26.956
Contiguity of order 1	CPO = 1813.9	SD	(0.041)	(1.112)	(2.355)
		95% CI	(-4.430,-4.268)	(24.938,29.305)	(22.567,31.814)
$W \circ D$ single_house and	WAIC = 2940.3	Mean	-4.490	30.970	27.609
Distance band	CPO = 1811.3	SD	(0.046)	(1.246)	(2.396)
		95% CI	(-4.580,-4.400)	(28.524,33.418)	(23.142,32.550)
$W \circ D$ budgetmeters and	WAIC = 2949.4	Mean	-4.360	27.500	29.005
Contiguity of order 1	CPO = 1814.6	SD	(0.040)	(1.074)	(2.556)
		95% CI	(-4.438,-4.282)	(25.392,29.609)	(24.247,34.284)
$W \circ D$ budgetmeters and	WAIC = 2938.7	Mean	-4.513	31.657	27.009
Distance band	CPO = 1813.9	SD	(0.047)	(1.293)	(2.336)
		95% CI	(-4.606,-4.420)	(29.120,34.196)	(22.652,31.823)
Mobility	WAIC = 2972.6	Mean	-4.270	25.113	22.010
	CPO = 1849.2	SD	(0.045)	(1.213)	(1.974)
		95% CI	(-4.357,-4.182)	(22.726,27.491)	(18.342,26.093)

When comparing the models' fit related to the different weights matrices included in Table 4, it can be seen that parameter estimates can differ considerably. The estimated value for the autoregressive parameter ρ is large and statistically significant, according to its 95% credible interval, in all models, an indication that there is a clear sign for the existence of spatial autocorrelation. Interpretation of the value of the estimated parameter is difficult, however.

Table 5. *Results obtained after fitting the geometric mean spatial conditional normal Poisson models to the COVID-19 incidence data in Flanders, for the different weights matrices considered.*

Weights matrix			$\hat{\beta}$	$\hat{\rho}$	$\hat{\tau}$
Contiguity of order 1	WAIC = 2945.1	Mean	-0.786	0.770	22.488
	CPO = 1920.6	SD	(0.122)	(0.036)	(1.938)
		95% CI	(-1.027,-0.546)	(0.699,0.840)	(18.871,26.479)
Contiguity of order 3	WAIC = 2942.1	Mean	-0.885	0.740	15.724
	CPO = 1918.5	SD	(0.162)	(0.048)	(1.331)
		95% CI	(-1.202,-0.567)	(0.647,0.834)	(13.235,18.459)
Inverse distance	WAIC = 2941	Mean	3.753	2.108	19.144
	CPO = 1857.8	SD	(0.380)	(0.112)	(1.630)
		95% CI	(3.006,4.500)	(1.887,2.328)	(16.097,22.494)
Negative exponential	WAIC = 2941.4	Mean	4.919	2.451	12.709
	CPO = 1922.7	SD	(0.695)	(0.205)	(1.067)
		95% CI	(3.554,6.283)	(2.049,2.854)	(10.712,14.900)
Distance band	WAIC = 2938.6	Mean	0.241	1.071	25.000
	CPO = 1820.9	SD	(0.157)	(0.046)	(2.151)
		95% CI	(-0.067,0.550)	(0.980,1.161)	(20.985,29.432)
$W \circ D$ single_house and	WAIC = 2945.3	Mean	-0.811	0.762	22.535
Contiguity of order 1	CPO = 1905.7	SD	(0.121)	(0.036)	(1.943)
		95% CI	(-1.048,-0.573)	(0.692,0.832)	(18.909,26.536)
$W \circ D$ single_house and	WAIC = 2940.4	Mean	0.216	1.062	28.129
Distance band	CPO = 1806.3	SD	(0.144)	(0.042)	(2.445)
		95% CI	(-0.066,0.498)	(0.979,1.145)	(23.571,33.170)
$W \circ D$ budgetmeters and	WAIC = 2946.3	Mean	-0.780	0.773	23.622
Contiguity of order 1	CPO = 1928.9	SD	(0.118)	(0.035)	(2.045)
		95% CI	(-1.012,-0.549)	(0.702,0.839)	(19.809,27.838)
$W \circ D$ budgetmeters and	WAIC = 2938.9	Mean	0.268	1.076	27.501
Distance band	CPO = 1812.3	SD	(0.148)	(0.043)	(2.380)
		95% CI	(-0.023,0.558)	(0.991,1.162)	(23.058,32.402)
Mobility	WAIC = 2960.6	Mean	-1.766	0.482	14.842
	CPO = 1915.7	SD	(0.115)	(0.034)	(1.275)
		95% CI	(-1.993,-1.541)	(0.416,0.548)	(12.460,17.466)

The information criteria values obtained (i.e., WAIC) for the fitting of these models indicate that the best fit for the models accounting only for contiguity or distance amongst municipalities is for the distance band spatial weights (WAIC = 2938.4). With regard to the predictive accuracy measure (i.e., CPO), the best fitting model is the one using the contiguity of order one criterion (CPO = 1807.4). As for the models taking into

account the similarity in socio-economic status, the combination of `single_house` or `budgetmeters` and distance bands are the best fitting models (WAIC = 2940.3 and CPO = 1811.3, and WAIC = 2938.7 and CPO = 1813.9, respectively).

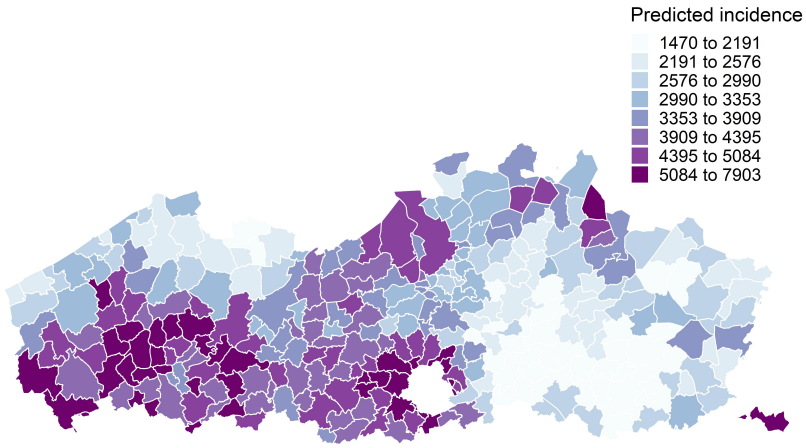
For the model considering the mobility matrix, we can conclude that, according to the information and the predictive criteria, this model did not provide a good fit for the dataset under study. Unlike the simulation study, where the spatial pattern was explicitly constructed based on the mobility matrix and the models accurately identified this structure, in this case, the mobility structure does not seem to be the underlying structure driving the spatial pattern of incidence rates in the dataset under study.

Similar results are observed in Table 5, where the fitting of these models appears to be very similar, according to the WAIC and CPO values, to the ones reported in Table 4. Here, the models with the smallest values were the ones using the distance band weights matrix (WAIC = 2938.6 and CPO = 1820.9) and similarity matrix of the distance band and `single_house` or `budgetmeters` (WAIC = 2940.4 and CPO = 1806.3, and WAIC = 2938.9 and CPO = 1812.3, respectively). In these weights matrices, larger weights are specified for municipalities that lie within the distance band and have similar values of these variables. Therefore, the fitting of these models suggests that this structure could be properly explaining the underlying spatial dependence, assuming that the variables considered represent the socio-economic or demographic characteristics of the population in these municipalities.

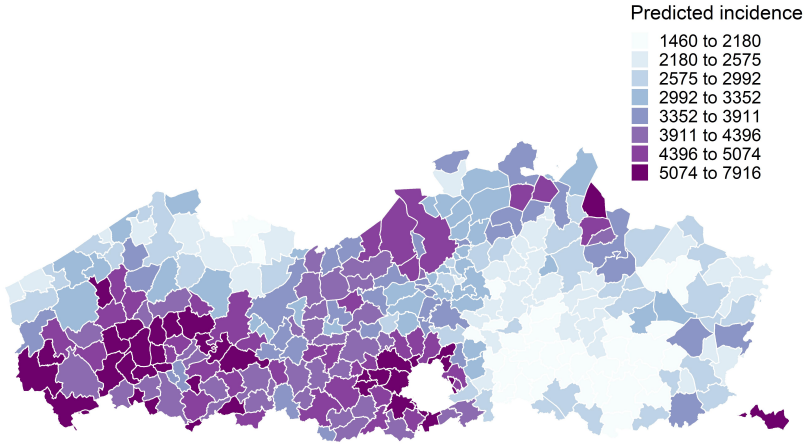
Regarding the spatial autoregressive parameter ρ, here the spatial lag is also significant for all the fitted models, indicating that the spatial autocorrelation is being properly captured. However, we believe it is relevant to mention that, while this finding provides some clear evidence that the autoregressive structure of the model is appropriate for capturing the spatial autocorrelation present in the dataset under study, it does not provide any argument in favor of this being the best or the correct model.

Moreover, the interpretation of parameter ρ can be useful in order to quantify how much the spatial structure considered can influence the resemblance of the incidence rate in a municipality to the geometric mean of the incidence rates of its neighbours. In the models where the distance band matrix was used, the parameter ρ has posterior mean approximately equal to 1, and, thus, in this setting, we find that the rate in a municipality is close to the geometric mean of the rates in the municipalities within the distance band. For the models where the specified weights matrix was either the exponential or the inverse distance, the estimated values of ρ was approximately equal to two, suggesting that the rate in a municipality is the square of the geometric mean of the rates of its neighbours. For the remaining models, this parameter's estimated value was smaller than one. For example, in the model with the mobility matrix, it was $\hat{\rho} = 0.4823$, suggesting that, for this connectivity structure, the rate in a municipality is approximately the squared root of the geometric mean of the rates of its neighbours.

Figure 2 includes the maps of the predicted incidence obtained after fitting the geometric mean spatial conditional normal Poisson models using the spatial weights matrix following the distance band criterion and the similarity spatial matrix combining the dif-



(a) *Predicted incidence obtained from the model using the spatial matrix following the distance band criterion, fitted to the COVID-19 data in Flanders.*



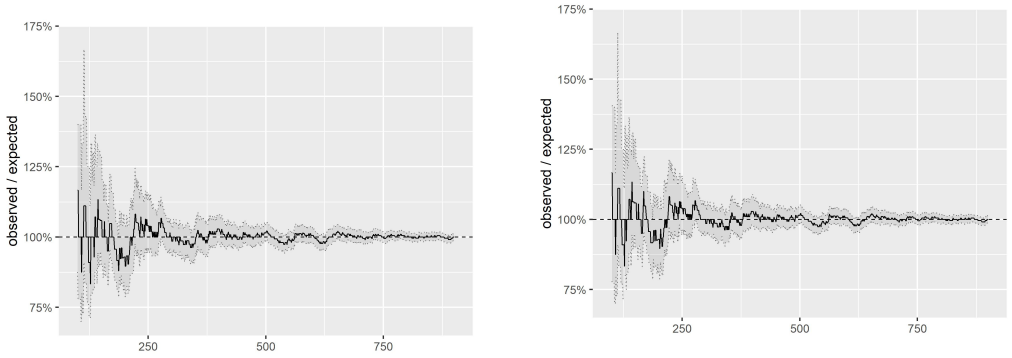
(b) *Predicted incidence obtained from the model using the similarity spatial matrix combining the differences in the variable budgetmeters and the distance bands criterion, fitted to the COVID-19 data in Flanders.*

Figure 2. *Predicted incidence obtained from some of the geometric mean spatial conditional normal Poisson models considered, fitted to the COVID-19 data in Flanders.*

ferences in the variable `budgetmeters` and the distance bands criterion, which were considered as the best fitting ones. Similar maps obtained for some of the other fitted models have been included in Figure S2 in the supplementary material. If we compare these maps with the observed incidence map shown in Figure 1, we can see that, in general, the predictions are quite accurate, as they are very similar to the observed incidence. In addition, when compared to each other, we note that the predictions obtained differ only for a small number of municipalities. In addition, scatterplots of the observed versus the predicted rates, obtained from the fitting of these models are included in Figure S3 in the supplementary material, where it can be seen that the fitted models show high accuracy in the prediction of the incidence rates.

We can also check the distributional assumptions in the fitted models, which is a Poisson distribution, where the overdispersion is accommodated by means of the inclusion of a random effect in the regression for the mean. This can be achieved by using the `distribution_check` function from the R package inlatools (Onkelinx, 2019). Here, simulations are drawn from the model and the empirical cumulative distribution function (eCDF) is computed for the observed response and for the simulated data, so that they can be compared.

Figure 3 includes the plots which illustrate these comparison results for two of the fitted models. In each figure, the black line is the result of dividing the eCDF of the observed data by the median of the eCDF's of the simulated datasets, and the grey bands represent the 95% credible intervals of the simulated data. In addition, the dotted horizontal line placed at 100% indicates where the ratio of the eCDF's is equal to one. If the eCDF is inside the credible intervals, which is the case for all of the models fitted here, the assumed distribution in the model seems to be a plausible one. Moreover, given that the eCDF is quite close to the reference line, these results suggest that the data is well modelled with this distribution.



(a) *Geometric mean model where the spatial matrix follows the distance band criterion.*

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

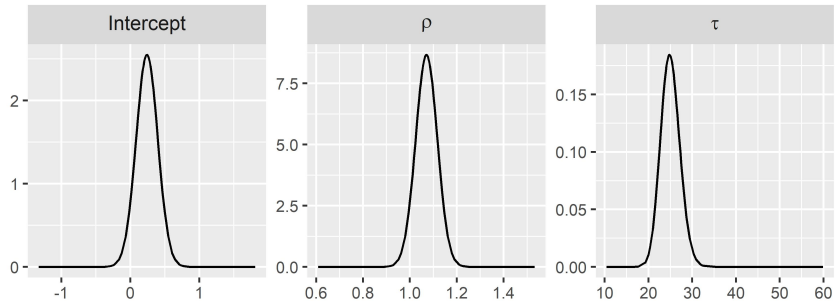
Figure 3. *Distribution check for some of the fitted models.*

Additionally, Figure 4 shows the marginal posterior distribution of the parameters estimated from some of the fitted models, where it can be verified that the normality assumption holds. Distribution checks and the posterior marginals for the estimated parameters in the geometric mean models corresponding to the spatial weights matrix following the contiguity of order one criterion and the mobility matrix have been included in Figures S4 and S5 in the supplementary material.

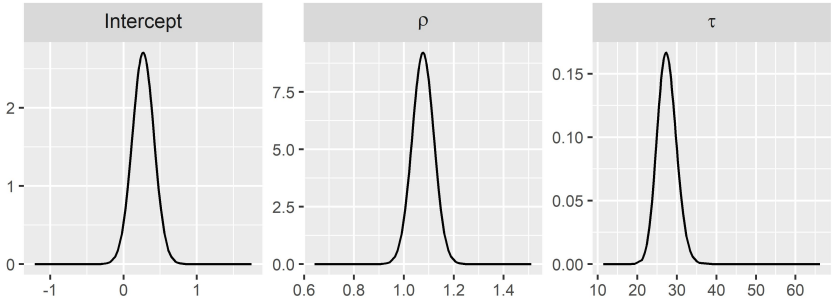
After examining the results obtained in this section, we could conclude that, on the one hand, with the proposed model we present an appealing interpretation of the spatial parameter, given by the geometric mean of the incidence rates. We have shown how this interpretation can change for the different fitted models, indicating how much

the spatial structure considered explains the spatial autocorrelation by means of the geometric mean of the rates in the neighbouring municipalities. On the other hand, by examining different weights matrices, we can have a better idea of the underlying spatial dependence structure of the data. When the similarity matrices based on the distance band were used, the information criteria values were similar to the model considering the traditional distance band matrix. Therefore, taking into account that they provide similar predictions and similar fit, we believe that, for the specific data set considered, this weights matrix could represent a proper choice for modelling the spatial underlying structure of the data.

Finally, we would like to briefly mention that the computation time needed for fitting the models in this section is of approximately one second for each one of the fitted models.



(a) *Geometric mean model where the spatial matrix follows the distance band criterion.*



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

Figure 4. *Posterior densities from the parameters estimated for some of the fitted models.*

4.3. Comparison to the BYM2 and Leroux spatial models

In this section, we will fit the BYM2 and Leroux spatial models to the COVID-19 data in Flanders. Additional details about these models have been included in Section 8 in the

supplementary material. We should stress here that one of our main goals in this work is to present the geometric mean proposal as a new extension of the spatial conditional Poisson model in Cepeda-Cuervo et al. (2018). In these models, the interpretation of the spatial parameters is different from that of the BYM2 or the Leroux models. Furthermore, the spatial conditional and the geometric mean models offer the possibility of specifying any weights matrix in a straightforward way, as it is used for computing a spatial lag. In our view, this feature makes these models more appealing for investigating different spatial structures, which is another one of our goals in this work. In the case of the BYM2 or the Leroux models, this is not straightforward due to its limitations, where the assumed spatial structure needs to be symmetric, which is not the case, for example, for the mobility matrix we have employed before. Although it is known that any matrix can be symmetrized, this would include carrying out a previous process, which is not required when fitting our proposed models.

Nevertheless, we believe it can be useful to compare the performance of the proposed methods with that of the BYM2 and Leroux models, often employed in disease mapping applications. Therefore, in order to specify the BYM2, we consider the model in equation (S3) in the supplementary material, where, in order to specify the penalized complexity priors and following Simpson et al. (2017), for the precision parameter τ_s we assume that $\text{Prob}(1/\sqrt{\tau_s} > 0.2/31) = 0.01$ and, for the mixing parameter ϕ_s, $\text{Prob}(\phi_s < 0.5) = 2/3$. Additionally, for the Leroux model, we consider the formulation in equation (S4) in the supplementary material. In this case, the prior for the precision parameter τ_u is a noninformative Gamma distribution (i.e., $\tau_u \sim G(1 \times 10^{-4}, 1 \times 10^{-4})$) and the prior for the spatial parameter ϕ_u is a uniform distribution over the unit interval ($\phi_u \sim U(0,1)$) (Lee, 2013). For the intercept, we assume a noninformative normal prior distribution (i.e., $\beta \sim N(0, 1 \times 10^5)$). Note that, in the BYM2 and Leroux models, the spatial weights matrix is defined based on contiguity of order one. The results obtained after fitting these models to the COVID-19 data in Flanders are included in Table 6 and Table 7.

Table 6. *Results obtained after fitting the BYM2 model to the COVID-19 incidence data in Flanders.*

	Mean	SD	95% CI
$\hat{\beta}$	-3.3924	0.004	(-3.400,-3.385)
$\hat{\tau_s}$	12.057	1.131	(9.885,14.318)
$\hat{\phi_s}$	0.976	0.020	(0.923,0.998)
WAIC = 2932.9 CPO = 1802.6			

Results reported in the previous section indicate that the smallest WAIC resulted for the model using the distance band criterion (WAIC = 2938.6) and the smallest CPO was obtained for the geometric mean model using the similarity matrix of the distance band and `budgetmeter` (CPO = 1806.3). As for the BYM2 model, we can see that the WAIC and CPO values obtained are slightly lower than those obtained for the pre-

Table 7. *Results obtained after fitting the Leroux spatial model to the COVID-19 incidence data in Flanders.*

	Mean	SD	95% CI
$\hat{\beta}$	-3.189	0.258	(-3.590,-2.544)
$\hat{\tau_u}$	6.907	0.603	(5.777,8.150)
$\hat{\phi_u}$	0.989	0.017	(0.942,0.999)
WAIC = 2954.8 CPO = 1958.7			

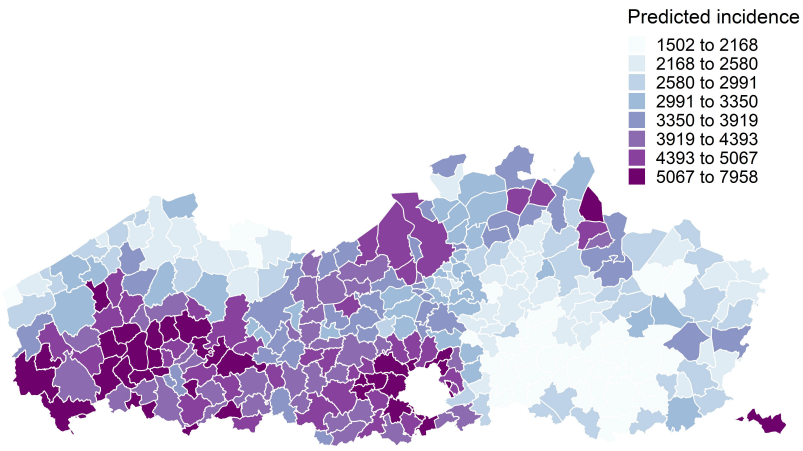
vious models. In contrast, the Leroux model yielded larger information criteria values when compared to both the spatial conditional and BYM2 models, suggesting a weaker goodness of fit to the data under study.

In addition, the value obtained for the mixing parameter in the BYM2 model, $\hat{\phi}_s = 0.976$, suggests that more than 97% of the variability in the data is being explained by the spatially structured effect. Similarly, the estimated spatial parameter in the Leroux model, $\hat{\phi}_u = 0.989$, indicates that most of the variability in the data is explained by the spatial component, which is consistent with the BYM2 estimate of ϕ_s.

Regarding the predictive accuracy of these models, Figure 5 includes the maps of the predicted incidence obtained from their fitting, where we can see that the predictions are very accurate when compared to the map of the observed incidence in Figure 1, and also very similar to the ones obtained in the previous section for our proposed methods (see Figure 2). The scatterplots of the observed versus the predicted incidence rates are also included in Figure S3(i) in the supplementary material, showing some issues in some of the municipalities.

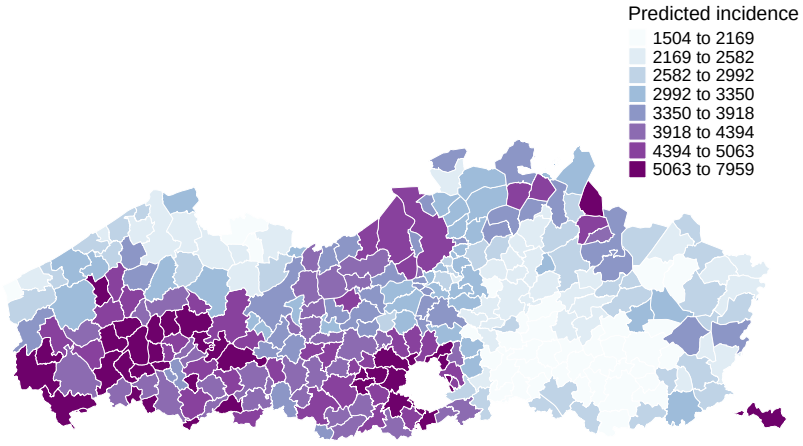
The computation time needed to fit the BYM2 model was of approximately ten seconds, while the Leroux model required five seconds. In contrast, for the spatial conditional and geometric mean models fitted in Section 4.2, the runtime was of about one second for each one of the fitted models. Moreover, in the simulation study carried out in Section 3, we fitted 250 models for each one of the 12 scenarios. When fitting the Leroux and the BYM2 models to these datasets, the runtime increased by a factor of five and ten, respectively, when compared to the geometric mean model. In other words, fitting the Leroux model for the entire simulation study would require more than 4 hours, while the BYM2 model would take about 8 hours, compared to the 50 minutes required for the geometric mean model. Therefore, in our view, this is an important advantage worth mentioning for our proposed models over the BYM2 and the Leroux models, which are commonly used in this area of research.

Despite the fact that the information criteria values favoured the BYM2 model and that its predictive accuracy is similar to the one from the geometric mean model, we restate our goal here of presenting the geometric mean proposal, which can be viewed as an alternative to the Leroux, BYM and BYM2 models, and to investigate the weights matrices which best reflect the spatial underlying process.



(a) *Predicted incidence obtained from the BYM2 model, fitted to the COVID-19 data in Flanders.*

Figure 5. *Predicted incidence obtained from the BYM2 and Leroux models, fitted to the COVID-19 data in Flanders.*



(b) *Predicted incidence obtained from the Leroux model, fitted to the COVID-19 data in Flanders.*

Figure 5. *Predicted incidence obtained from the BYM2 and Leroux models, fitted to the COVID-19 data in Flanders (Continued).*

There are situations where the spatial conditional models may offer a better fit than the Leroux, BYM and BYM2 models, or viceversa. We believe that the choice of the model to fit should depend on the specific objective of the study. For example, Morales-Otero and Núñez-Antón (2021) reported that, given by the information criteria values obtained, the spatial conditional and the BYM and BYM2 models offered a very similar fitting to the infant mortality data they studied. In addition, in Morales-Otero, Gómez-

Rubio and Núñez-Antón (2022), the spatial conditional models were employed in order to illustrate a new fitting approach in INLA.

5. Discussion

In this work, we have studied the geographical spread of COVID-19 cases in the municipalities of Flanders in Belgium during the period going from September 2020 to January 2021. In order to be able to fit these data, we have considered the Bayesian spatial conditional model proposals (Cepeda-Cuervo et al., 2018), which assume the incidence of cases in a municipality is conditional on the incidence of cases in neighbouring municipalities. These models offer a great flexibility and also the possibility that considering different weights matrices can be done in a direct and very simple way.

We have proposed a geometric mean spatial conditional model, where the logarithm of the rates is employed for computing the spatial lag component. This model offers an interpretation of the spatial parameter ρ based on the geometric mean, representing how the incidence rate in one municipality resembles the geometric mean of the rates in its neighbours. For the spatial weights matrix used in these models, we have proposed alternative specifications based on a combination of the similarity of a certain variable in the different locations and the distance between these municipalities. In addition, we have also considered the connectivity structure given by the mobility of individuals among the municipalities under study.

In order to further assess the performance of the proposed methods when the correlation among the different municipalities under study is given by a connectivity pattern, such as, for example, the mobility matrix, we have carried out a simulation study where we induce correlation in the response variable based on this structure. In this study, we have been able to appropriately verify that the models are able to identify the correct spatial structure for most of the cases under study.

In the application to the COVID-19 data in Flanders, we have compared these proposed models with the ones in Cepeda-Cuervo et al. (2018) finding that our proposal provides a similar fit, but offers a particular and straightforward interpretation within the context of the specific dataset under analysis. We have fitted these models by using different definitions for the weights matrices employed to compute the spatial lag, such as the classical ways of accounting for spatial autocorrelation based on contiguity and distance, as well as the similarities weights matrices we proposed as alternatives. In addition, we have also studied the use of the mobility matrix in modelling the COVID-19 incidence data in Flanders, which is given by the proportion of time individuals from one municipality spent in a different one.

In order to provide a comparison of the proposed models with other commonly used models employed in disease mapping applications, we have also fitted the BYM2 and Leroux spatial models to the dataset under study. Results indicate that the BYM2 model provides a similar fit based on information criteria and demonstrates a comparable predictive accuracy to that of our proposed model. However, it may be the case that the

dataset under study may not be the best example to fully justify the need for the geometric mean spatial conditional model. Moreover, we believe it is important to clarify that this study initially began as an investigation into whether the mobility connectivity structure could explain the spatial pattern of COVID-19 incidence across municipalities in Flanders. Addressing this question required flexible spatial models, such as the spatial conditional models proposed by Morales-Otero and Núñez-Antón (2021), which motivated the use of this approach in our analysis. Subsequently, we developed the geometric mean spatial conditional model, adjusting the primary focus of this work to introducing it as a flexible alternative for capturing spatial dependencies and making it possible to specify different spatial structures.

The BYM2 model is well established in this area of research but, at the same time, we also believe that our model provides interpretational advantages, computational simplicity, and the flexibility to easily test for different spatial structures. For example, the computation time required to estimate a geometric mean model is ten and five times shorter, respectively, than the one needed for the BYM2 or Leroux models. We consider this to be a significant advantage of our proposed model, particularly when researchers need to perform simulation studies, such as the one presented in Section 3, where a large number of models must be efficiently fitted. Nonetheless, we recognize that further applications are necessary to fully evaluate the benefits and limitations of the geometric mean model in different contexts, and we intend to continue exploring this model proposal structure in our future research.

In any case, overall results suggest a strong spatial correlation in the dataset under study, which is best explained by the distance band spatial weights matrix. This implies that, for the data under study, the underlying spatial process is well explained and modelled by this spatial structure.

Finally, we believe it it worth mentioning that, in this work, we focus on the analysis of the data corresponding to the time period of the COVID-19 second wave in Flanders, and we have tried to characterize the overall spatial pattern believed to be present in this wave. Our main interest for this specific application does not include transmission. However, for future research we are also interested in performing comparison with the spatial pattern of additional COVID-19 waves in the area under study, by being able to propose a spatio-temporal approach, which is out of the scope of this paper. We have already developed spatio-temporal extensions of the spatial conditional models and specific proposals are in the process of being finalized, so that they are part of a different manuscript to be later submitted for possible publication. Moreover, one of our objectives is to be able to apply these proposals to the comparisons of the different waves in the dataset we have analysed here.

Acknowledgements

This research has been partially funded by Ministerio de Ciencia e Innovación (MCIN, Spain), Agencia Estatal de Investigación (AEI/10.13039/501100011033/) and Fondo Eu-

ropeo de Desarrollo Regional (FEDER) "Una manera de hacer Europa" under the I+D+i research grant PID2020-112951GB-I00 and by the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group) under research grant IT-1508-22.

Supplementary material

Supplementary material has been included in an accompanying document. It includes heatmaps of the weights matrices considered here, scatterplots of the observed versus the predicted rates, additional details about the BYM2 and the Leroux models, among other information that can be useful to the readers.

Code availability

There is an R script available with an example of how to fit the proposed geometric mean spatial conditional model, using a simulated dataset. The link is: https://github.com/mabelmo/Spatial-Autoregressive-Geometric-Mean-Model-.git.

References

Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics* 27(3), 247–267.

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

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, 1–20.

Carroll, R., Lawson, A.B., Faes, C., Kirby, R.S., Aregay, M., and Watjou, K. (2015). Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. *Spatial and Spatio-temporal Epidemiology* 11-15, 45–54.

Case, A., Hines, J.R. Jr., and Rosen, H.S. (1993). Budget spillovers and fiscal policy interdependence: Evidence from the States. *Journal of Public Economics* 52(3), 285–307.

Cepeda-Cuervo, E., Córdoba, M., and Núñez-Antón, V. (2018). Conditional overdispersed models: Application to count area data. *Statistical Methods in Medical Research* 27, 2964–2988.

D'Angelo, N., Abbruzzo, A., and Adelfio, G. (2021). Spatio-temporal spread pattern of COVID-19 in Italy. *Mathematics* 9(19), 2454.

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., and Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International Journal of Health Geographics* 6(54).

Ejigu, B.A. and Wencheko, E. (2020). Introducing covariate dependent weighting matrices in fitting autoregressive models and measuring spatio-environmental autocorrelation. *Spatial Statistics* 38, 100454.

Fritz, C., De Nicola, G., Rave, M., Weigert, M., Khazaei, Y., Berger, U., Küchenhoff, H., and Kauermann, G. (2022). Statistical modelling of COVID-19 data: Putting generalized additive models to work. *Statistical Modelling* 0(0), 1471082X221124628.

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* 5(3), 187–199.

Jackson, M.C. and Sellers, K.F. (2008). Simulating discrete spatially correlated Poisson data on a lattice. *International Journal of Pure and Applied Mathematics* 46(1), 137–154.

Johnson, D.P., Ravi, N., and Braneon, C.V. (2021). Spatiotemporal Associations Between Social Vulnerability, Environmental Measurements, and COVID-19 in the Conterminous United States. *GeoHealth* 5(8), e2021GH000423.

Knorr-Held, L. and Richardson, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Applied Statistics* 52(2), 169–183.

Konstantinoudis, G., Padellini, T., Bennett, J., Davies, B., Ezzati, M., and Blangiardo, M. (2021). Long-term exposure to air pollution and COVID-19 mortality in England: A hierarchical spatial analysis. *Environment International* 146, 106316.

Konstantinoudis, G., Cameletti, M., Gómez-Rubio, V., León Gómez, I., Pirani, M., Baio, G., Larrauri, A., Riou, J., Egger, M., Vineis, P., and Blangiardo, M. (2022). Regional excess mortality during the 2020 COVID-19 pandemic in five European countries. *Nature Communications* 13(482).

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.

Morales-Otero, M., Gómez-Rubio, V., and Núñez-Antón, V. (2022). Fitting double hierarchical models with the integrated nested Laplace approximation. *Statistics and Computing* 32(62)

Natalia, Y.A., Faes, C., Neyens, T., and Molenberghs, G. (2022). The COVID-19 wave in Belgium during the Fall of 2020 and its association with higher education. *PLOS ONE* 17(2), e0264516.

Onkelinx, T. (2019). The inlatools package, https://inlatools.netlify.app/. Last accessed 09 August 2022.

Pettit, L.I. (1990). The Conditional Predictive Ordinate for the Normal Distribution. *Journal of the Royal Statistical Society - Series B* 52(1), 175–184.

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.

Sahu, S.K. and Böhning, D. (2022). Bayesian spatio-temporal joint disease mapping of Covid-19 cases and deaths in local authorities of England. *Spatial Statistics* 49, 100519.

Simpson, D., 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.

Slater, J.J., Brown, P.E., Rosenthal, J.S., and Mateu, J. (2022). Capturing spatial dependence of COVID-19 case counts with cellphone mobility data. *Spatial Statistics* 49, 100540.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11(116), 3571–3594.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* 41(3-4), 434–449.

Zeger, S.L. and Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics* 44(4), 1019–1031.