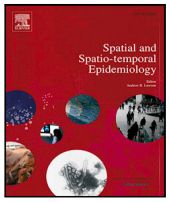
Spatial and Spatio-temporal Epidemiology 51 (2024) 100681

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste



Original research

Similarity- and neighbourhood-based dynamic models for infection data: Uncovering the complexities of the COVID-19 infection risks

Helena Baptista [a,*], Jorge M. Mendes [b], Ying C. MacNab [c]

[a] NOVA Information Management School (NOVA IMS), Campus de Campolide, Lisboa, 1099-032, Portugal
[b] Comprehensive Health Research Centre (CHRC), NOVA Medical School, Universidade NOVA de Lisboa, Lisboa, Portugal
[c] Epidemiology and Biostatistics, School of Population and Public Health, University of British Columbia, Vancouver, Canada

ARTICLE INFO

Keywords:
COVID-19
Gaussian Markov random field
Similarity-based Gaussian Markov random fields
Adaptive modelling
Forecasting

ABSTRACT

Understanding spatial and temporal risk dependencies and correlation is crucial when studying infectious diseases which spread out in consecutive waves. By analysing weekly COVID-19 case data collected from the disease's first reported case on March 3, 2020, to April 22, 2021, in 278 municipalities in Mainland Portugal, we demonstrate that the complexity of infection risks varies based on the outbreak's severity, suggesting that a single model definition is insufficient to explain the multifaceted underlying phenomena. This study employs a dynamic, conditionally specified Gaussian Markov random field model with a novel approach to characterise COVID-19 infection risk dependencies through the similarity of areal-level covariates within a Bayesian hierarchical model framework that accounts for each identifiable wave. The results indicate that the neighbourhood-based conditional autoregressive model, which is static and based on an adjacency-based neighbourhood matrix, do not necessarily captures the disease's complex spatial–temporal nature. Furthermore, the best-fitting dynamic model may not necessarily be the best predicting model in certain situations, which can lead to inadequate resource allocation in epidemic situations. Accurate forecasting can help inform decisions regarding difficult-to-measure impacts, potentially saving lives. Implementing the proposed novel approach would have produced information that would have been overwhelmingly critical to the respective authorities in protecting those in more unfavourable economic or other conditions.

1. Introduction

Much time has passed since December 2019, when the Chinese authorities identified a new coronavirus strain, SARS-CoV-2, and the world was thrown into turmoil due to its rapid global spread. Many scientific papers on this topic have been published in multiple disciplines, including medicine, mathematics, social sciences and geographical statistics.

A review of published scientific papers evaluating 63 articles using spatial analysis to treat this problem was made available, as early as May 2020 (Franch-Pardo et al., 2020). The same type of search delivered more than 5000 entries at the time of writing this paper.

On top of the possible geographically related complexities of the COVID-19 (henceforth, COVID) spreading mechanisms, the high discrepancy found in incidence by age group and other demographics, cannot be overlooked while discoursing on the subject. Publications early in 2020 showed that Karaye and Horney (2020), Zhang and Schwartz (2020), Paez et al. (2021) (i) minority status and language, household composition and transportation, and housing and disability predict COVID infection; (ii) positive correlations exist between COVID incidence and mortality rates and socio-economic factors including population density, proportions of elderly residents, poverty, and percentage population tested; (iii) higher incidence is associated with higher Gross Domestic Product (GDP) per capita and the presence of mass transit systems; in contrast, population density and percentage of older adults displayed negative associations with incidences of COVID.

Historically, the main goal of disease mapping is to determine the underlying disease risk and has been dedicated to non-infectious and infectious diseases alike, whose incidence distribution is spatially (and sometimes temporally) structured. When applied to rare and non-communicable chronic diseases such as heart disease and cancer (Besag et al., 1991; Best et al., 1999), the spatial and temporal correlations are classically due to the structure of unknown environmental cofactors. In the context of a contagious disease, like COVID, the outcome of a primary case can, in addition, generate secondary occurrences of the pathology in a close spatial and temporal neighbourhood. In this case, extra-dependencies originate in contagion besides the correlations due

* Corresponding author.
E-mail address: mhbaptista@novaims.unl.pt (H. Baptista).

https://doi.org/10.1016/j.sste.2024.100681
Received 21 January 2024; Received in revised form 23 May 2024; Accepted 18 August 2024
Available online 4 September 2024

to external factors. Depending on the spatial and temporal scales, the secondary infection cases (resulting from direct contamination due to people interactions) may occur in the same area and during the same period; thus, local overdispersion may appear. The secondary cases may also occur in the neighbouring regions and/or during the following periods; therefore, infectious diseases generally show higher spatial and temporal dependencies, which stresses the importance of using adequate statistical methodologies to account for those correlations. Spatio-temporal conditional autoregressive processes (CAR-AR) with both real and simulated data have produced smoothed risk maps that are epidemiologically relevant and interpretable (Coly et al., 2021). In a recent paper, MacNab (2022) conducted a deeply comprehensive reflection on the past and present of Bayesian disease mapping and probed adaptive models specifically. Eight different adaptive CAR models were used to create comparable and consistent posterior estimates of relative risks for the COVID county-level aggregates of daily infection cases for eighty-seven counties of Minnesota, USA.

On top of posterior estimates of relative risks, Bayesian spatio-temporal disease mapping models have been formulated for short-term forecasting of areal-level COVID infection or mortality (Sahu and Böhning, 2022; Stewart et al., 2022; MacNab, 2023). Specifically, (non-adaptive) CAR-AR (e.g., the CARBayesST) models have been developed for joint modelling of COVID infection and mortality and short-term mortality forecasts in local authorities of England (Sahu and Böhning, 2022). To illustrate near real-time monitoring and forecasting for COVID situational awareness, Stewart et al. (2022) present spatial–temporal (ST) models for the counties of USA. Adaptive Gaussian Markov random field ST models for short-term forecasts of COVID infection in counties of Minnesota (USA) have also been proposed (MacNab, 2023). It is worth mentioning that work has been produced more recently (Lawson, 2023) to evaluate the predictive capability of Bayesian Susceptible–Infected–Removed models for short-term areal-level infection forecasting.

The dataset used here is a collection of all positive cases of COVID in Portugal, for the period between March 3, 2020 and April 22, 2021 (Mendes et al., 2022). The dataset includes the patient's municipality and the diagnosis date. The model proposed in this work is an adaptive Bayesian model with a Gaussian Markov random field (GMRF), which is the golden standard prior distribution applied to a set of random effects to model the spatio-temporal correlations inherent in most ecological data (MacNab, 2022). The GMRF prior includes, in its definition, an adjacency matrix for a given lattice and its neighbourhood system. Our adaptive model transforms the adjacency matrix to include information other than geographical closeness.

Examples of published research dedicated to modelling the COVID spreading patterns using other than neighbourhood information include a study for Italy (Mingione et al., 2022), where two different specifications for the adjacency matrix in the CAR-AR model were utilised. The first matrix, referred to as W1, specified a neighbourhood structure based on proximity flows and the availability of direct train, flights, and ferry connections. For instance, this matrix could lead to distant regions being neighbours due to frequent internal flight connections. The original matrix was a weighted measure of commuters' flow and was not symmetric since exchanges may have different magnitudes in the two directions. The matrix in this application was dichotomised as a means of symmetrising it, e.g., $w_{ij} = 1$ if a positive flow existed in at least one of the two directions, and is equal to zero otherwise. The second adjacency matrix, referred to as W2, was the most typically adopted network defined by regions' mutual geographical position. It was considered a first-order structure, where only pairs of regions sharing at least one land border were considered neighbours. Another example is a study for Spain (Slater et al., 2021), which used information on the number of trips between regions as edge weights, and concluded that the number of people moving between regions explained the variation in COVID case counts better than physical proximity data. Still another study (Sahu and Böhning, 2022) proposed a two-stage hierarchical Bayesian model as a joint bivariate model for the number of cases and deaths observed weekly in England's different local authority administrative regions. An adaptive model was proposed for the weekly COVID death rates as part of the joint bivariate model. The adaptive model detected possible step changes in death rates in neighbouring areas. The joint model was also used to evaluate the effects of several socio-economic and environmental covariates on the rates of cases and deaths. Including these covariates points to the presence of a north-south divide in both the case and death rates.

The (Sahu and Böhning, 2022) paper is also among the earlier works of using spatiotemporal disease mapping models for small area COVID-19 forecasting. The traditional (Knorr-Held, 2000) spatiotemporal model, named the Anova model (i.e., the ST.CARanova() model in the CarBayesST R package (Lee, 2020)), was used for the England COVID-19 mortality forecasting. More recently, MacNab (2023) presents a comprehensive development of spatiotemporal models for small area COVID-19 infection risk prediction and infection forecasting. In MacNab (2023), two model constructions for a comprehensive development of spatiotemporal models are presented and illustrated. One is a generalised spatiotemporal autoregressive-moving-average (STARMA) model construction for spatiotemporal extensions of the time series ARMA models. The other is the convolution construction for formulation of multidimensional convolution models. For both model constructions, a rich variety of adaptive parameterisations are introduced for flexible characterisation of spatial and spatiotemporal dependency, variability, and discontinuity.

Lawson (2023) concludes that, from a predictive point of view, it is clear that spatio-temporal models applied to county-level COVID data within the US vary in how well they fit over time and how well they predict future events. A fundamental result of the study is that the predictive capability of models varies over time and using the same model could lead to poor predictive performance. Another example is a study conducted for a small region in Spain, for patients treated in the hospital over 29 weeks, using not only contiguity and distance matrices but also some socio-demographic variables (Briz-Redón et al., 2021), the population density, the average household income in Euro and the proportion of the population aged 65 years and over. These three socio-demographic covariates were used to compute a "socio-demographic distance" between each pair of areas under analysis. In this case, the authors concluded that testing multiple sensible specifications of the neighbourhood matrix is highly advisable; otherwise, an unsuitable choice of this matrix can lead to poor models in terms of explanatory and forecasting capability.

In this paper, we contribute and extended the recent literature to areal-level COVID infection risk inference and forecasting: (1) we illustrate, for the first time, the use of the neighbourhood-based adaptive CARBayesST model for both infection risk inference and infection forecasting; (2) we extend the CARBayesST model formulation to develop a new similarity-based Gaussian Markov random field ST model for infection risk inference and infection forecasting; and (3) we illustrate and contrast the above-mentioned neighbourhood-based (adaptive) CARBayesST and similarity-based Gaussian Markov random field ST models, for their complementary roles of uncovering the complexities of COVID infection risks and for infection forecasting. In this study, the Knorr-Held (2000) model (i.e. the above-mentioned ST.CARanova() model in CARBayesST package) was also fit for areal level risk prediction and forecasting. To conserve space, results of the ST.CARanova() models are presented and discussed in the Appendix C.

The remainder of the paper is organised as follows. Data and the study motivation are described in the following section. The details of the applied spatio-temporal adaptive and non-adaptive models used are provided next. The following section shows the relevant results obtained, and the paper concludes with enlightening future developments. Extra information on the models used and other results obtained are provided in the Appendix.

2. Data and study motivation

The study region is mainland Portugal, which has been partitioned into $k = 278$ local municipalities. Data were obtained from the General Directorate of Health (DGS, Direcção Geral da Saúde, in Portuguese) (General Directorate of Health - NHS[Internet], 2021), which used to released it to authorised researchers periodically. Data are available on a day/anonymised patient level, between March 3, 2020 and April 22, 2021. The end date is intentionally after the highest wave whose spike occurred on January 27, 2021 has ended and before an high proportion of the population was vaccinated (at that point, only 8% of the population was fully vaccinated, according to the Portuguese authorities), as the vaccine would create an entirely different scenario. The data was further cleaned to include only complete (date, municipality, age and sex) new cases of COVID, totalling 776 977 cases. New cases were summed at the week/municipality level. From now on, Y_{kt} is the count of new COVID cases in municipality k in week t.

The rationale for calculating a weekly sum per municipality is an attempt to maximise granularity while accounting for the "seasonality" of testing and the erratic daily fluctuations due to late reporting. Weekend days have fewer reported cases simply due to testing and reporting levels.

Municipalities have different population sizes and demographic structures, and indirect standardisation is implemented to account for the fact that areas with larger populations are likely to exhibit more new cases. Specifically, the population in municipality k in 2019, obtained from Statistics Portugal (INE, Instituto Nacional de Estatística, in Portuguese) (Portugal Statistics, 2021), is split into $V = 18$ strata based on sex and age, and n_{kv} denotes the number of people in strata v in municipality k. These strata specific population sizes n_{kv} are multiplied by national strata-specific incidence rates $\gamma_{tv} = newcases_{tv}/population_v$, and the results are summed over strata to give the expected number of new COVID cases in municipality k during period t (as defined above). Mathematically these expected counts are computed by $E_{kt} = \sum_{v=1}^{18} n_{kv}\gamma_{tv}$, and represent the number of new cases expected if national age and specific COVID incidence rate are applied to municipality k during period t.

More important than the number of new cases is the standardised incidence rate (SIR), which is computed by dividing the observed number of cases (Y) by the expected number of cases (E). If a SIR is equal to 1 the observed and expected number of cases is equal, representing an average risk area relative to the entire study region. Similarly a SIR of 0.9/1.1 corresponds to a 10% decreased/increased risk compared to the national average. The mean SIR per municipality, for the cumulative period, is shown in Fig. 1. These rates are obtained simply as the ratio of the sum of the observed and expected number of COVID cases over the 403 days.

Cartographic datasets with the administrative boundaries of Portuguese mainland municipalities were obtained from the General Directorate of the Territory (DGT, Direcção Geral do Território, in Portuguese) (CAOP, 2020).

It seems clear that the two major metropolises of the country, Lisboa and Porto, were the most affected areas, jointly with some very specific regions of the countryside, both in the northern and in the southern parts of the country. This phenomenon is mostly in line with the previously mentioned local overdispersion expected in a contagious disease.

Auxiliary data referring to 2019 (time-invariant), collected from INE are (Portugal Statistics, 2021) :

- municipal population density, based on the total municipality population and the municipality area in square kilometres (henceforth, density).
- municipal proportion of the population above 60 years of age (henceforth, above 60).

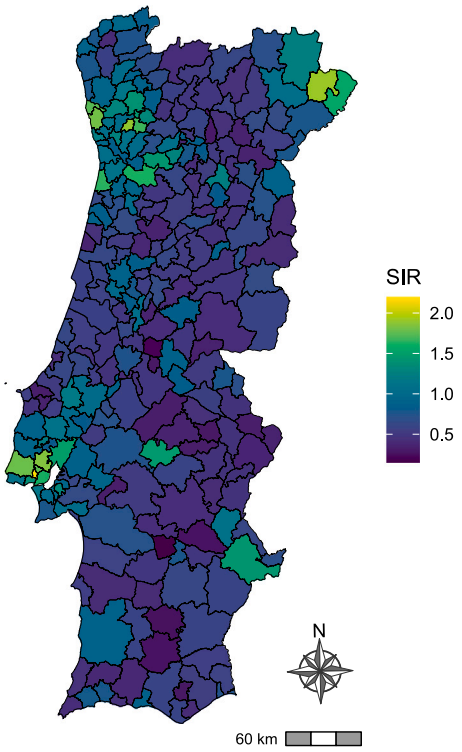


Fig. 1. Mean SIR for each municipality, over the whole period.

- municipal Gini coefficient of the declared raw income net of income tax paid per family, an indicator of inequality of income distribution (takes the zero value when all have equal income and 100 when all income is given to one individual) (henceforth, Gini).
- municipal proportion per one thousand actively working people of *social integration income* receivers. It refers to a special program with the goal of providing income for basic needs to all (henceforth, social program).
- municipal absolute number of actively working people (henceforth, active people).
- municipal elderly index, the quotient between the number of people above 65 years of age and those below 15 years of age (henceforth, elderly index).

The study motivation is twofold. First, identifying high-risk areas and patterns of disease incidence; second predicting the future behaviour of the disease spread. When achieved, both goals are critical for public health planning activities, such as priority setting for allocating funds and localised disease prevention or intervention. Our study shows that the traditional spatio-temporal approaches are insufficient to achieve both goals, properly. Therefore, a dynamic modelling approach is needed to accomplish those two goals, identifying high-risk areas to treat the already infected population better and forecasting the high-risk areas for better prevention of disease spread. The auxiliary data above is crucial to achieving both goals.

3. Spatio-temporal models

From an epidemiological perspective, there is no strict definition for what is considered an epidemic wave (or phase) or not. However the word wave implies a natural pattern of peaks and valleys, suggesting that future disease outbreaks are possible even during a lull. In our case, retrospectively studying the data gives us the advantage of setting each period's start and end dates to investigate. In reality, we want to mimic

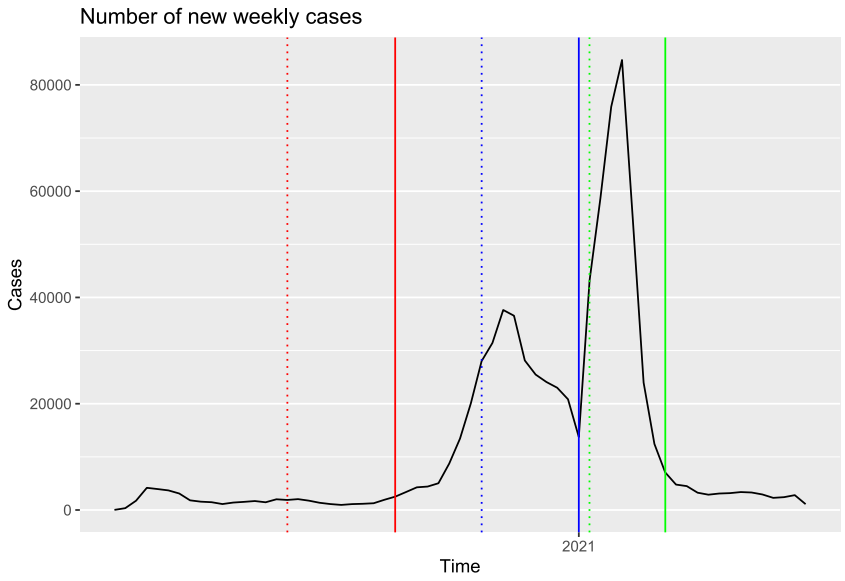


Fig. 2. COVID new cases, the dotted lines determine the beginning of the period, and the solid lines the end of the period. Red lines correspond to the low incidence period, the blue lines to the high incidence period and the green lines to the highest period, as defined for the study.

the analysis that could have been done to help decide on the actions that should/would need to be taken to prevent significant life losses, which means forecasting the future number of cases.

Taking advantage of that, in the Bayesian setting, missing values in the response variable are handled by computing their predictive distribution, and this is possible because the data-generating process is specified in the model likelihood. To measure the forecast accuracy we used the root mean squared error, which is mathematically computed by

$$RMSE = \sqrt{\frac{1}{2k}\sum_{t=1}^{2}\sum_{k=1}^{278}(y_{tk} - y_{tk}^{*})^2}$$

where y_{tk}^{*} is the predicted value for the municipality k on week t. As we will create a two-week forecast, the RMSEs are calculated for the two weeks of forecasts (combined), thus $\frac{1}{2k}$. Further to the RMSE, we also measure bias and coverage for the forecast. Bias is defined as the difference between the forecast and the actual values and is extremely important in this case, because a positive bias, can lead to a shortage of available medical personnel and/or medicines to treat the disease cases. Mathematically, bias is computed as

$$Bias = \frac{1}{2k}\sum_{t=1}^{2}\sum_{k=1}^{278}(y_{tk} - y_{tk}^{*}).$$

The coverage rate is defined as the proportion of municipalities for which the actual value is contained in the 95% credible interval calculated for the forecast.

With that goal in mind, we illustrate three different periods, a period with very low incidence from June 22 to August 16, 2020 (weeks 26 to 33), in a total of 8 weeks, a period of high incidence, from October 26, 2020 to January 3, 2021 (weeks 44 to 53), in a total of 10 weeks, and the period with highest incidence, from January 4 to February 28, 2021 (weeks 1 to 8) in a total of 8 weeks. See in Fig. 2 the time series plot of the number of cases and the periods' delimitation on time.

3.1. Model definition

Working with rare events such as COVID, the observed case counts may be viewed as Poisson events and can be modelled as a log-linear model, given by:

$$Y_{kt} \sim Poisson(E_{kt}\theta_{kt}) \quad \text{for} \quad k = 1, \ldots, K, \quad t = 1, \ldots, N,$$

$$ln(\theta_{kt}) = \beta_0 + \boldsymbol{\beta}_p \mathbf{X}_p + \psi_{kt}, \tag{1}$$

where θ_{kt} is the unknown relative risk, the E_{kt} are defined above, the ψ_{kt} is the random effect, one for each municipality k and time period t, and $\mathbf{X}$ are the above mentioned auxiliary data. The regression parameters β are to be estimated by the model. Given the well-know ecological bias (Wakefield, 2007), and in light with the main aim of this work, which is not assess the effect of those covariates nor to purpose interventions to impact the disease spread, the estimates of the regression parameters are only illustrated in Appendix A and in Table A.12.

The applied model is a spatially autocorrelated first-order autoregressive process (given the fact that, per selected period, the data contains only 8, 10, and 8 data points, respectively for each period), $\boldsymbol{\psi}_t = \rho_T \boldsymbol{\psi}_{t-1} + \boldsymbol{\epsilon}_t$, where $\boldsymbol{\psi}_t = (\psi_{1t}, \ldots, \psi_{kt})$ denotes the vector of random effects for all areal units at time t, and the vector of errors $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{kt})$ is modelled as a spatially correlated process (see below in Section 4.2 for the assessment of both spatial and temporal autocorrelation).

The covariance structure of $\boldsymbol{\epsilon}_t$ controls the spatial autocorrelation, and is given by $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W}, \rho_S)^{-1})$, where τ^2 is the process variance and (ρ_S, ρ_T) respectively control the levels of spatial and temporal autocorrelation, with values of 0 corresponding to independence while a value of 1 corresponds to strong autocorrelation. The spatial correlation is induced by the matrix $\mathbf{W}$, which denotes the pairwise interactions between each pair of municipalities (w_{kj}), and will be developed in detail in Section 3.2 below.

The precision matrix $\mathbf{Q}(\mathbf{W}, \rho_S)$ corresponds to the CAR prior proposed by Leroux et al. (2000). Leroux CAR (LCAR) is one of the most used priors for disease mapping and is a two-parameter and full rank GMRF. The two-parameters are (ρ_S, σ) respectively, the spatial and non-spatial parameters. ρ_S is a spatial dependence smoothing parameter and σ is a Gaussian scale parameter. The LCAR has gained popularity due in part to the analytic result that when $\rho_S = 0$, the LCAR reduces to independent and identical Gaussian priors for the log relative risks, and when $\rho_S = 1$ that reveals a strong autocorrelation. The precision matrix is equivalent to:

$$\epsilon_{kt} | \boldsymbol{\epsilon}_{-kt}, \mathbf{W} \sim N\left(\frac{\rho_S \sum_{j=1}^{K} w_{kj}\epsilon_{jt}}{\rho_S \sum_{j=1}^{K} w_{kj} + 1 - \rho_S}, \frac{\tau^2}{\rho_S \sum_{j=1}^{K} w_{kj} + 1 - \rho_S} \right)$$

Models are fitted in a Bayesian setting using a Markov chain Monte Carlo (McMC) simulation, combining Gibbs sampling and Metropolis–Hastings steps. Software to implement the model in R is available in the CARBayesST package (Lee, 2020).

The priors used in the model are defined as follows: for the regression parameters $\boldsymbol{\beta}$ a $N \sim (0, 100000)$ is used, for the process variance τ^2 an $Inverse - Gamma \sim (1, 0.01)$ is used, and for the autocorrelation parameters (ρ_S, ρ_T) a $Uniform \sim (0, 1)$ is used.

Inference is based on 6000 McMC samples generated from three independent Markov chains each. Each chain was burnt in for 200 000 samples by which convergence was assessed to have been reached, and then run for a further 2000 000 samples, which were thinned by 1000 to reduce their autocorrelation significantly. Convergence was visually evaluated using traceplots and numerically assessed using the Gelman–Rubin diagnostic (Gelman et al., 2013), see the Appendix for some examples.

3.2. Relationship matrices

We used four different definitions of the relationship matrix.

The first one is the traditional first-order symmetric contiguity matrix **W**, where if $w_{kj} = 1$ means that municipality k shares a common physical border with municipality j, and otherwise if $w_{kj} = 0$. Thus if $w_{kj} = 1$ then the random effects $(\epsilon_{kt}, \epsilon_{jt})$ are modelled as spatially correlated, while if $w_{kj} = 0$ then $(\epsilon_{kt}, \epsilon_{jt})$ are assumed to be conditionally independent.

The second matrix type is built in a three-step process and is repeated for six times, each one for each variable mentioned in the auxiliary data ($p = 1, \ldots, 6$). Each variable is used to create one similarity matrix **S**. The process goes as follows:

1. this is the "distance step" (Baptista et al., 2016) where for every municipality pair w_{kj}, the absolute gap on the p variable, is calculated between region k and region j,

   $$p_{kj} = |p_k - p_j|,$$

2. this is the "mean distance step", where the mean distance is calculated for every municipality:

   $$\tilde{p_k} = \frac{1}{K} \sum_{j=1}^{K} p_{kj}$$

3. this is the final step, where a symmetric matrix **S**, is defined by:

   $$s_{kj} = \begin{cases} 1, & \text{if } p_{kj} \leq \tilde{p_k} \\ 0, & \text{otherwise,} \end{cases}$$

This symmetric matrix can lead to spatially distant regions being neighbours; for example, two regions with high population density, Lisboa and Porto, while physically separated by more than 300 km, will be neighbours in the density matrix case.

The third matrix type is the independent case. That was achieved by using a matrix, called **I** and defined by:

$$i_{kj} = \begin{cases} 1, & \text{if } k = j \\ 0, & \text{otherwise.} \end{cases}$$

By doing this, the model runs under an independent Gaussian prior with a precision matrix equal to $\tau(1 - \rho_S)I_n$ because when $D_w = I_n$ (where D_w is diagonal with $D_{kk} = \tau_i^2$) then $D_w - I_n = 0$, a matrix with all elements equal to zero, the resulting posterior variance is $\sigma = \frac{1}{\sqrt{\tau(1-\rho_S)}}$.

Finally, an adaptive model proposed elsewhere (Rushworth et al., 2017) gives the fourth matrix type. This specific model allows spatially varying random effects to be correlated (inducing smoothness) or conditionally independent (no smoothing), which is achieved by modelling the adjacency elements in **W**, i.e. $\mathbf{w}^+ = w_{kj}|j \sim k$, (when k is a physical neighbour of j in the sense of the W matrix) as random variables on the unit interval, rather than being set equal to 1. The remaining elements of **W** corresponding to non-adjacent areal units remain set at 0. Note that **W** is the same matrix used in the spatial neighbourhood-based model. This model is implemented in the CarBayesST R package (Lee, 2020), and the associated variance priors used are the package defaults.

Table 1
Measures of fit of the model for the low period.

Model	DIC	p.d	WAIC	p.w	LMPL	loglik
1. neighbourhood	**5336**	697	5323	522	−2849	−1971
2. density	5359	762	5310	540	−2884	−1918
3. above 60	5355	758	5309	539	−2867	−1920
4. gini	5364	762	5315	540	−2883	−1920
5. social program	5355	763	**5302**	537	−2859	−1915
6. active people	5362	760	5316	540	−2863	−1921
7. elderly index	5390	787	5314	538	−2870	−1908
8. independent	5964	423	6702	853	−3167	−2559
9. adaptive	**5306**	676	**5295**	508	−2817	−1977

Notes: DIC: Deviance Information Criterion.
p.d: DIC - Estimated effective number of parameters.
WAIC: Watanabe–Akaike Information Criterion (Vehtari et al., 2016).
p.w: WAIC - Estimated effective number of parameters.
LMPL: The Log Marginal Predictive Likelihood.
loglik: loglikelihood.

3.3. Likelihood model

The likelihood model used is the one defined in Eq. (1). All nine models run, include covariates, however when a similarity matrix is used, the **S** matrix, the vector of covariates **X** is not p but is $p-1$, as the variable used to compute the matrix is not included. This way prevents the information contained in the covariate from being used twice.

Nine models were run for each of the three different periods, one for matrix types, 1, 3 and 4, and six models using the matrix type 2.

4. Results

Given our two-fold study motivation we used the Watanabe–Akaike Information Criterion (WAIC) to select the best model for short term risk predictions, and the Deviance Information Criteria (DIC) for in-sample risk prediction and smoothing (Gelman et al., 2014). The WAIC has the desirable property of averaging over the posterior distribution rather than conditioning on a point estimate which is particularly helpful for models with hierarchical and mixture structures in which the number of parameters increases with sample size and where point estimates often do not make sense. Together with leave-one-out cross-validation, WAIC is the best method for estimating point-wise out-of-sample prediction accuracy from a fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values; in other words, it is best for short-term risk predictions.

4.1. Risk smoothing

Tables 1–3 contain the fit measures for the spatio-temporal models for the three periods considered.

Overall, it can be said that the adaptive models outperformed (in terms of DIC and WAIC) the non-adaptive models in all three periods. The adaptive model 9 shows the lowest DIC and WAIC for the low period, the second lowest DIC and WAIC for the high period, and the second lowest DIC in the highest period. The social program similarity model (model 3) in the low period and the density-based similarity model (model 2) in the highest period, show the second lowest WAIC.

Table 4 contains the selected (lowest DIC) model's posterior parameters medians and credible intervals for the periods considered. Given the priors considered (see above in the model definition chapter), the high spatial and temporal autocorrelations in the Low period model are of special note.

Fig. 3 shows the posterior exceedance probabilities that the risk in the week with an higher incidence in each of the three periods is greater than 1, as provided by the selected model for each period.

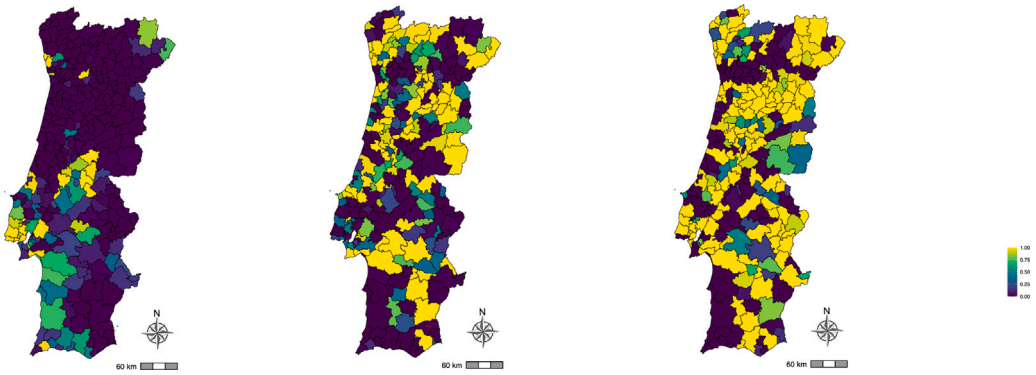


Fig. 3. Posterior exceedance probabilities, for the low period (left), high period (centre), and highest period (right).

Table 2
Measures of fit of the model for the high period.

Model	DIC	p.d	WAIC	p.w	LMPL	loglik
1. neighbourhood	19 507	2106	19 278	1381	−10 473	−7647
2. density	19 557	2186	19 254	1382	−10 474	−7593
3. above 60	19 543	2178	19 235	1373	−10 432	−7593
4. gini	19 544	2174	19 243	1375	−10 442	−7598
5. social program	19 538	2176	19 232	1373	−10 440	−7594
6. active people	19 561	2182	19 261	1382	−10 451	−7599
7. elderly index	19 566	2183	19 266	1383	−10 441	−7600
8. independent	**18 587**	1624	**18 211**	944	−9 576	−7669
9. adaptive	**19 420**	2093	**19 213**	1387	−10 388	−7617

Notes: DIC: Deviance Information Criterion.
p.d: DIC - Estimated effective number of parameters.
WAIC: Watanabe–Akaike Information Criterion (Vehtari et al., 2016).
p.w: WAIC - Estimated effective number of parameters.
LMPL: The Log Marginal Predictive Likelihood.
loglik: loglikelihood.

Table 3
Measures of fit of the model for the highest period.

Model	DIC	p.d	WAIC	p.w	LMPL	loglik
1. neighbourhood	16 878	1725	16 674	1116	−8997	−6715
2. density	16 863	1766	**16 588**	1093	−8944	−6666
3. above 60	16 869	1749	16 628	1105	−8981	−6685
4. gini	16 875	1752	16 633	1107	−8988	−6685
5. social program	16 878	1757	16 629	1105	−8979	−6682
6. active people	16 871	1760	16 611	1100	−9002	−6675
7. elderly index	16 867	1759	16 611	1101	−8982	−6675
8. independent	**16 124**	1337	**15 809**	769	−8325	−6725
9. adaptive	**16 791**	1686	16 610	1105	−8944	−6709

Notes: DIC: Deviance Information Criterion.
p.d: DIC - Estimated effective number of parameters.
WAIC: Watanabe–Akaike Information Criterion (Vehtari et al., 2016).
p.w: WAIC - Estimated effective number of parameters.
LMPL: The Log Marginal Predictive Likelihood.
loglik: loglikelihood.

Table 4
Posterior parameters medians and credible intervals, for the Low period (adaptive model); and the high and highest periods (independent prior).

Parameter	Median	2.5%	97.5%
Low period			
τ^2	1.36	1.10	1.68
ρ_S	0.86	0.74	0.93
ρ_{1T}	0.79	0.73	0.85
High period			
τ^2	0.05	0.01	0.26
ρ_{1T}	0.84	0.80	0.88
Highest period			
τ^2	0.03	0.00	0.15
ρ_{1T}	0.82	0.78	0.86

4.2. Short-term forecast

Using the high-level period, we illustrate the possible forecasting process. We used the four different types of models to calculate the forecast. Using model 1, model 5, model 8 and model 9; we predicted two weeks within the wave, week 52 and week 53 of 2020, and two weeks right after the end of the wave, week 1 and week 2 of 2021. However, while showing good smoothing properties, the adaptive model shows inferior forecasting properties, as it does not converge when provided with no values in the response variable on the mentioned weeks, invalidating this model's forecasting process.

Before starting the forecast process, the spatial autocorrelation of the model's residuals was tested. Moran's I statistic is the most commonly used measure of spatial correlation (Moran, 1950). For model 1, the test shows no residual spatial autocorrelation, with a Moran's I value of −0.15 and a $p-$ value of 0.99, for model 5 and model 8 the test shows residual spatial autocorrelation, as expected, with a Moran's I value of 0.26 and 0.25 respectively and a $p-$ value < 0.001. The temporal autocorrelation is taken care by the inclusion of the autoregressive term, in the random effects. For more complex models in terms of temporal autocorrelation see Appendix C.

It needs to be noted that, as seen in "Model definition" section, the temporal autoregressive model is of order one, which will imply some possible degradation on the second-week forecast, as a predicted value will be used instead of an actually verified one. However, in the real world, data availability will not be immediate, and planning activities need time, which cannot be accommodated in one week only.

The overall country prediction is not very useful for making resources allocations decisions. Therefore, we created six possible segments, three based on the number of cases observed by municipality and three based on the municipal proportion per one thousand actively working people of *social integration income* receivers. Segments were created on quantiles; the top 75% quantile, the lower 25% quantile and the second and third quantiles, represent the middle 50% of the segments.

Tables 5 to 10 are a compilation of the results obtained from all perspectives. Those tables, namely, one for within the wave (Tables 5 to 7) and one for outside the wave (Tables 8 to 10) forecasts, include the result of the three forecast quality measures: RMSE, Bias and Coverage. On top of that, those measures are provided at the municipality segment group level.

As shown in Tables 5 to 7, as expected, the overall RMSE is lower for model 8, followed by model 5. It is the highest for model 1, for the within the wave forecast. This aspect remains consistent in the highest types of municipalities, either when segmented by its quantiles in terms of the number of cases or by its quantiles in terms of the municipal proportion per one thousand actively working people of *social integration income* receivers. The mid and lowest quantiles have minimal differences between the three models. This outcome is an helpful result because the municipalities with the most needed attention are those

Table 5
RMSE for the predictions within the wave (High period).

RMSE per group of	Model 1. neighbourhood	Model 5. social income	Model 8. independent	Mean cases	Mean social income
Overall	42	38	32	61	26
High quantile cases	80	73	59	182	25
Mid quantile cases	15	15	16	28	24
Low quantile cases	11	12	13	4	31
High quantile social income	56	47	40	69	49
Mid quantile social income	41	38	31	57	23
Low quantile social income	25	27	27	61	11

Note: Values have been rounded into integers as is not possible to have non-integers cases.

Table 6
Bias for the predictions within the wave (High period).

Bias per group of	Model 1. neighbourhood	Model 5. social income	Model 8. independent
Overall	−5	−6	−4
High quantile cases	−18	−22	−13
Mid quantile cases	1	0	1
Low quantile cases	−5	−5	−5
High quantile social income	−16	−14	−11
Mid quantile social income	−4	−4	−2
Low quantile social income	3	−3	−1

Note: Values have been rounded into integers as is not possible to have non-integers cases.

Table 7
Coverage for the predictions within the wave (High period).

Coverage per group of	Model 1. neighbourhood	Model 5. social income	Model 8. independent
Overall	0.87	0.87	0.83
Above	34	35	52
Below	40	39	45
High quantile cases	0.95	0.96	0.96
Mid quantile cases	0.90	0.89	0.83
Low quantile cases	0.70	0.71	0.66
High quantile social income	0.85	0.86	0.81
Mid quantile social income	0.83	0.83	0.79
Low quantile social income	0.95	0.95	0.91

Note: Values have been rounded into integers as is not possible to have non-integers cases.
Coverage above: Number of actual observations above the upper value in the 95% credible interval.
Coverage below: Number of actual observations below the lower value in the 95% credible interval.

with an high number of cases and those with an high proportion of economically deprived people. However, model 5 is the only one with no positive bias cases, which is crucial during a pandemic. Following the pattern, coverage results continue to show better adequacy for model 5 versus model 8 and model 1.

As can be depicted in Tables 8 to 10, the forecast for the two weeks outside the wave, predictions for model 5 show a RMSE equal to that provided by model 1, while model 8 suffers a significant loss in performance with the highest RMSE. As expected, all three model's forecasts become degraded in accuracy when forecasting outside the wave. However, for those municipalities with the highest quantile of the municipal proportion per one thousand actively working people of *social integration income* receivers, model 5 shows a lower RMSE, which would have permitted a more efficient resource allocation, at the right moment. Again, model 5 is the best-performing one, with only one case of positive bias. Predicting lower values of disease cases, can create a sense of false recovery. Again, the coverage analysis goes in the same direction, with better coverage for model 5 in the areas with more cases and more municipal proportion per one thousand actively working people of *social integration income* receivers.

Therefore, given the results, model 5 seems to be the model that produce the best-balanced forecast, allowing an exemplary deployment of resources. This model forecast is almost unbiased for the period after the wave, unlike the two other models' forecast, which could lead to a situation of missing resources.

5. Conclusions

Once again, it is confirmed that COVID infection spreading patterns are more complex than just spatial and mobility closeness related. In fact, for the Low period, the adaptive model (model 9) closely followed by the model with the proportion of people receiving *social integration income* (model 5), included in the similarity matrix, are the two models delivering the lowest WAIC. This measure is a proxy of the proportion of the population with more unsatisfied needs and the population suffering the most during a pandemic. That information would have been critical to the authorities to create timely measures to protect those in more unfavourable economic and other conditions and prevent further spreading on that (and on other) segments of society.

The highest period is consecutive to the high period, resulting from a poorly controlled pandemic. In both high periods the WAIC reached by the independent model is the highest. The highest period is unique, as almost all municipalities were in the "danger zone" regarding the number of daily new cases, leading the country to an overall confinement period of more than three months, which could have been avoided if preventive measures (like non pharmaceutical interventions) had been taken ahead of time, using the forecasting power shown by the similarity model.

Of particular note is the fact that all but the similarity matrix models assume that if two areas do not share a border, they will be modelled as conditionally independent. While physical closeness is an important factor in infection spread, other types of "closeness" can also play important roles in disease mapping and forecasting. Close work between statisticians and epidemiologists, can uncover those types of "closeness" and can play a major role in characterising disease risks. In fact, looking at the NTD (neglected tropical diseases) (NTD, 2023) website, it is also the opinion of those experts, that mathematical modelling for infectious diseases and public health works best through collaborations with epidemiologists, policymakers and field experts.

Furthermore, in this work we decide to create a discrete matrix, with areas being either "close" (represented by one) and "not close" (represented by zero) of each other, while a different option could have been taken by measuring that "closeness" in a continuum, like it was done on a previous study (Baptista et al., 2016). The former option can possibly result on information loss, and a comparison between the two options is part of a larger on-going research.

The present work illustrates once again (Baptista et al., 2016) that Bayesian disease mapping models gain significantly in explanatory and forecasting power by including extra information, according to the specific knowledge of the epidemiologists, in order to fit the right

Table 8
RMSE for the predictions after the wave (High period).

RMSE per group of	Model 1. neighbourhood	Model 5. social income	Model 8. independent	Mean cases	Mean social income
Overall	80	81	94	179	26
High quantile cases	149	148	175	512	24
Mid quantile cases	42	43	45	91	25
Low quantile cases	16	14	16	19	30
High quantile social income	120	116	142	199	49
Mid quantile social income	66	66	78	178	23
Low quantile social income	55	64	52	162	11

Note: Values have been rounded into integers as is not possible to have non-integers cases.

Table 9
Bias for the predictions within the wave (High period).

Bias per group of	Model 1. neighbourhood	Model 5. social income	Model 8. independent
Overall	6	−1	12
High quantile cases	20	−7	49
Mid quantile cases	6	5	2
Low quantile cases	−7	−5	−5
High quantile social income	−3	0	16
Mid quantile social income	13	3	15
Low quantile social income	3	−8	1

Note: Values have been rounded into integers as is not possible to have non-integers cases.

Table 10
Coverage for the predictions after the wave (High period).

Coverage per group of	Model 1. neighbourhood	Model 5. social income	Model 8. independent
Overall	0.94	0.94	0.92
Above	23	23	35
Below	8	8	11
High quantile cases	1.00	1.00	1.00
Mid quantile cases	0.96	0.97	0.95
Low quantile cases	0.84	0.83	0.75
High quantile social income	0.91	0.92	0.87
Mid quantile social income	0.95	0.95	0.93
Low quantile social income	0.97	0.96	0.94

Note: Values have been rounded into integers as is not possible to have non-integers cases.
Coverage above: Number of actual observations above the upper value in the 95% credible interval.
Coverage below: Number of actual observations below the lower value in the 95% credible interval.

suitable model to the problem at hand. "One model to rule them all" does not exist, what exists is a plethora of models that will prove to be useful at different stages of pandemics and probably at different stages in all other diseases when modelled in a spatio-temporal setting. As authorities become more aware of the power of data and data analysis, collecting useful and important data will require the willingness of all parties, which should also be more accessible by the researchers and the public.

Further work may consider using a Negative Binomial model to account for overdispersion caused by the contagious nature of the disease. On top of that, while the LCAR is commonly used, the proper CAR (pCAR) has its own advantages, and it may be more suited for these type of problems because of its two parameters, the spatial and the non-spatial, play separate and different roles, one regulates spatial dependency, the other controls non-spatial variance. With its rich options for multivariate generalisation, the pCAR, and some of its adaptive and multivariate extensions, have theoretical and practical appeals for modelling and interpreting spatial dependencies (MacNab, 2022).

Finally, work in progress is being conducted to include simulations both on the risk smoothing and in the forecasting tasks at hand. In a time of big data and fast decisions, testing beyond doubt the advantages and disadvantages of the different models is of paramount importance, and while our proposal shows better accuracy in forecasting, more information on the operational characteristics of the method are needed.

CRediT authorship contribution statement

Helena Baptista: Writing – original draft, Methodology, Formal analysis. **Jorge M. Mendes:** Writing – review & editing, Methodology. **Ying C. MacNab:** Writing – review & editing, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), Portugal, under the project UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020) - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Appendix A. Spatio-temporal model for the whole period

A Poisson log-linear model, without any covariate, for our data to assess the spatio-temporal autocorrelation existence in the data, is given by

$$Y_{kt} \sim Poisson(E_{kt}\theta_{kt}) \quad \text{for} \quad k = 1, \ldots, K, \;\; t = 1, \ldots, N,$$
$$ln(\theta_{kt}) = \beta_0 + e_{kt}, \tag{A.1}$$

where θ_{kt} is the disease risk in municipality k during period t relative to the expected counts E_{kt}, and is in the same scale as the SIR. e_{kt} is the error term, which would ideally be independent and identically distributed with mean zero and constant variance. As the population size is large and the incidence of the disease is low, the Poisson distribution is particularly suitable to model such data.

This model was fitted using maximum likelihood, and the t is defined as the cumulative figure of new cases for every 14 consecutive days. Then two values were selected by calendar civil month (in an attempt to closely match the data released to the media by the authorities during the pandemics). From the residuals (e_{kt}), we selected the last data period to measure the spatial autocorrelation, and the Lisbon municipality, one of the largest municipalities (in terms of population), to measure the temporal autocorrelation. Moran's I statistic is the most commonly used measure of spatial correlation (Moran, 1950). In this case, the test shows residual spatial autocorrelation, with a

Moran's I value of 0.067 and a $p-$ value of 0.03. The residual temporal autocorrelation test shows a statistically significant positive value at lag one and at lag two. As already stated, spatial correlation can be due to the spatial structure of covariates. Therefore, a Poisson log-linear model, with covariates, for our data is given by

$$\begin{aligned} Y_{kt} &\sim Poisson(E_{kt}\theta_{kt}) \qquad \text{for} \qquad k = 1, \dots, K, \;\; t = 1, \dots, N, \\ ln(\theta_{kt}) &= \beta_0 + \boldsymbol{\beta}_p \mathbf{X}_p + e_{kt}, \end{aligned} \tag{A.2}$$

$\boldsymbol{\beta}$ are coefficients to be estimated by the model, and $\mathbf{X}_p$ are the variables mentioned in Section 2.

While all covariates seem related to the disease's number of cases (its coefficients are statistically different from zero, for a 95% confidence level), that is not enough to eliminate spatial correlation. We again selected the last data period residuals, to measure the remaining spatial autocorrelation. Moran's I value is now 0.063 with a $p-$ value of 0.04, indicating that some spatial autocorrelation remains. On top of that, all regression coefficients are near one, with minimal standard errors, also indicating the need to include a set of random effects (to account for risk correlations in space and time).

The model used to handle that spatial and temporal correlation for our data is given by

$$\begin{aligned} Y_{kt} &\sim Poisson(E_{kt}\theta_{kt}) \qquad \text{for} \qquad k = 1, \dots, K, \;\; t = 1, \dots, N, \\ ln(\theta_{kt}) &= \beta_0 + \psi_{kt} \end{aligned} \tag{A.3}$$

and by

$$\begin{aligned} Y_{kt} &\sim Poisson(E_{kt}\theta_{kt}) \qquad \text{for} \qquad k = 1, \dots, K, \;\; t = 1, \dots, N, \\ ln(\theta_{kt}) &= \beta_0 + \boldsymbol{\beta}_p \mathbf{X}_p + \psi_{kt} \end{aligned} \tag{A.4}$$

where ψ_{kt} is the random effect for municipality k and time period t. The model used is a spatially autocorrelated second-order autoregressive process, $\boldsymbol{\psi}_t = \rho_1 \boldsymbol{\psi}_{t-1} + \rho_2 \boldsymbol{\psi}_{t-2} + \boldsymbol{\epsilon}_t$, where $\boldsymbol{\psi}_t = (\psi_{1t}, \dots, \psi_{kt})$ denotes the vector of random effects for all areal units at time t, and the vector of errors $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \dots, \epsilon_{kt})$ is modelled as spatially correlated. The covariance structure of $\boldsymbol{\epsilon}_t$ controls the spatial autocorrelation, and is given by $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W}, \rho_S)^{-1})$, where τ^2 is the process variance. The spatial correlation is induced by the neighbourhood matrix $\mathbf{W}$, which denotes whether each pair of municipalities is close. If $w_{ij} = 1$ it means that municipality i shares a common border with municipality j, and is $w_{ij} = 0$ otherwise. Thus if $w_{ij} = 1$ then the random effects $(\epsilon_{it}, \epsilon_{jt})$ are modelled as spatially correlated, while if $w_{ij} = 0$ then $(\epsilon_{it}, \epsilon_{jt})$ are assumed to be conditionally independent. The precision matrix $\mathbf{Q}(\mathbf{W}, \rho_S)$ corresponds to the CAR prior proposed by Leroux et al. (2000).

Both models are fitted in a Bayesian setting using a Markov chain Monte Carlo (McMC) simulation, combining Gibbs sampling and Metropolis–Hastings steps. Software to implement the model in R is available in the CARBayesST package (Lee, 2020).

The priors used in the model are defined as follows: for the regression parameters $\mathbf{X}$ a $N \sim (0, 100000)$ is used and for the process variance τ^2 an $Inverse - Gamma \sim (1, 0.01)$ is used.

Inference for both models (Eqs. (A.3) and (A.4)) is based on 6000 McMC samples generated from three independent Markov chains each. Each chain was burnt in for 200 000 samples by which convergence was assessed to have been reached, and then run for a further 2 000 000 samples, which were thinned by 1000 to reduce their autocorrelation significantly. For model (A.3), convergence was visually assessed using traceplots (see Fig. A.4 for β_0) and numerically assessed using the Gelman–Rubin diagnostic, and none of the values of the statistic were above 1.1, which is suggested as a convergence criterion (Gelman et al., 2013). For model (A.4), convergence was visually assessed using traceplots and numerically assessed using the Gelman–Rubin diagnostic. Four of the regressor's upper credible interval values of the statistic were above 1.1, which is suggested to be poor convergence (Gelman et al., 2013).

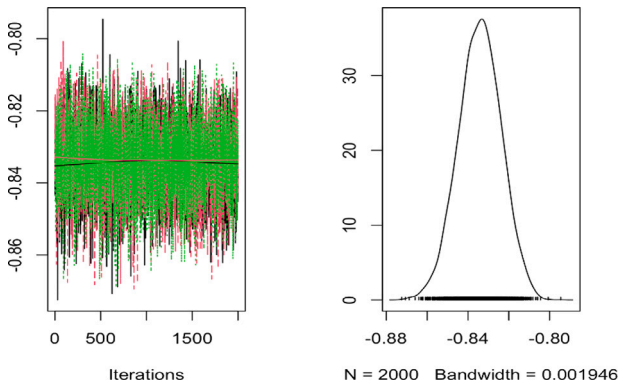


Fig. A.4. Traceplots of the McMC samples from each chain.

Table A.11
Posterior parameters medians and credible intervals, for model (A.3).

Parameter	Median	2.5%	97.5%
τ^2	1.18	1.10	1.25
ρ_S	0.80	0.75	0.85
ρ_{1T}	0.78	0.74	0.82
ρ_{2T}	−0.15	−0.19	−0.11

Table A.12
Posterior parameters medians and credible intervals, for model (A.4).

Parameter	Median	2.5%	97.5%
τ^2	1.14	1.07	1.21
ρ_S	0.76	0.70	0.82
ρ_{1T}	0.78	0.74	0.82
ρ_{2T}	−0.16	−0.20	−0.12
Density	1.11	1.01	1.20
Above 60	0.90	0.77	1.08
Gini	0.99	0.93	1.05
Social program	1.04	0.97	1.11
Active people	1.15	1.04	1.26
Elderly index	1.02	0.86	1.19

For model (A.3), the posterior medians of the process variance, the spatial and the temporal autocorrelations parameters $(\tau^2, \rho_S, \rho_{1T}, \rho_{2T})$ and 95 per cent credible intervals are shown in Table A.11.

The high value of $\rho_S = 0.80$ is clear, reflecting the strong spatial correlation ($\rho_S = 0$ corresponding to independence, while $\rho_S = 1$ corresponds to strong spatial autocorrelation) (Leroux et al., 2000).

For model (A.4), over and above the statistics already presented for model (A.3), the regressor's median coefficients are shown in Table A.12. Despite the inclusion of the covariates, the spatial autocorrelation remains as before.

Despite the poor convergence signals for model (A.4), we can say that after correcting for the spatial dependence all but two covariates' effects remain statistically different from zero. The importance of population density and the proportion of active people for the disease's risk seem evident. The relative risk of each covariate is given by the $exp(\beta_p)$ (Lee, 2020). For the risk factors: population density, above 65 years of age, gini, social program and elderly index, the posterior medians relative risk are close to 1, and are not statistically related with the disease's risk as the 95% credible interval contains the null risk of 1. In contrast, the population density and the proportion of active people are significantly related to the disease risk, with a 95% credible interval that is wholly above 1. The posterior median relative risk is 1.11 for density, suggesting that higher-density areas will have an increased risk of the disease, the same rationale for an increase in the proportion of active people (1.15). Including the time-invariant variables explains a tiny portion of the risk variation, as can also be seen by the slight drop in the process variance τ^2 from 1.18 to 1.14.

Table B.13
Measures of fit of the model for the Low, High and Highest periods for the model with the Mahalanobis distance.

Model	DIC	p.d	WAIC	p.w	LMPL	loglik
Low	5 386	786	5 312	538	−2 879	−1908
High	19 566	2188	19 251	1377	−10 453	−7595
Highest	16 864	1758	16 602	1097	−8 971	−6674

Notes: DIC: Deviance Information Criterion.
p.d: DIC - Estimated effective number of parameters.
WAIC: Watanabe–Akaike Information Criterion (Vehtari et al., 2016).
p.w: WAIC - Estimated effective number of parameters.
LMPL: The Log Marginal Predictive Likelihood.
loglik: loglikelihood.

Overall, it can be said that modelling such a long period of time only reinforces what is known regarding the nature of the phenomenon, a highly spatial and temporally autocorrelated process, and the inclusion of possible extra information does not provide much clarification on the spread of the disease.

Appendix B. A model including all covariates

It can be argued that the potential for multiple covariates to drive correlations in disease prevalence among regions should be considered, and therefore a model was run taking all the information and using the multivariate version of the statistical distance, the Mahalanobis distance between the areas defined as:

$$p_{ij} = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})},$$

where $\mathbf{x}_i' = (x_{1i}, x_{2i}, \ldots, x_{pi})$, $\mathbf{x}_j' = (x_{1j}, x_{2j}, \ldots, x_{pj})$, $\bar{\mathbf{x}}' = (\bar{x_1}, \bar{x_2}, \ldots, \bar{x_p})$ and $\mathbf{S}^{-1}$ is the inverse of the sample covariance matrix of the seven variables.

Results for that model show no improvement relative to the models already presented in Section 4.1. Table B.13 shows the model results.

Appendix C. A model including spatio-temporal interactions

It can be argued that our method may not truly improve forecast accuracy compared to traditional space–time models, such as those including space–time random effects considering precision structures (Knorr-Held, 2000), where authors create a model with spatio-temporal interactions, assuming that disease variations cannot be separated into temporal and spatial effects.

This type of model, given the interactions require more data points than the model shown before, and therefore, when we tried to run the model for the Low period, given its only 8 data points per area, it does not converge. For the two other periods, High and Highest, the model converge and the results are on Table C.14. On top, for the High period (the one we are using to forecast), we used two different matrices, the regular one, called **W**, and the second one, the **S**, build with the information of Social Program, the measure used on model 5. This model is implemented in the CarBayesST R package (Lee, 2020), using the function *ST.CARanova* and the associated variance priors used are the package defaults.

As it can be seen in Table C.14, none of the models achieve the lowest DIC and or the lowest WAIC, for the respective periods. For the High period it comes after the model 8 and model 9, and at par with model 5, for the model with the **W** matrix, while in the highest period it is the worse performing model in terms of DIC and WAIC.

In terms of forecasting for the period High, we use the model with the matrix **W**, as this was the best performing model. Results show a global RMSE for the within the period forecasting of 81, and for the two weeks after the wave of 596, both largely above other model's RMSE, showing that more complexity is not beneficial in this case.

Table C.14
Measures of fit of the model for the High and Highest periods for the model with spatio-temporal interactions.

Model	DIC	p.d	WAIC	p.w	LMPL	loglik
High with **W**	19 656	2362	19 229	1411	−10 700	−7466
High with **S**	19 669	2370	19 241	1416	−10 703	−7465
Highest with **W**	17 043	1881	16 755	1161	−9 259	−6641

Notes: DIC: Deviance Information Criterion.
p.d: DIC - Estimated effective number of parameters.
WAIC: Watanabe–Akaike Information Criterion (Vehtari et al., 2016).
p.w: WAIC - Estimated effective number of parameters.
LMPL: The Log Marginal Predictive Likelihood.
loglik: loglikelihood.

References

Baptista, H., Mendes, J.M., MacNab, Y.C., Xavier, M., de Almeida, J.M.C., 2016. A Guassian random field model for similarity-based smoothing in Bayesian disease mapping. Stat. Methods Med. Res. 25 (4), 1166–1184. http://dx.doi.org/10.1177/0962280216660407.

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

Best, N., Arnold, R., Thomas, A., Waller, L., Conlon, E., 1999. Bayesiann models for spatially correlated disease and exposure data. In: Bernardo, J., Berger, J., Dawid, A., Smith, A. (Eds.), Bayesian Statistics 6. Oxford Science Publications, Oxford, pp. 131–147.

Briz-Redón, Á., Iftimi, A., Correcher, J.F., De Andrés, J., Lozano, M., Romero-García, C., 2021. A comparison of multiple neighborhood matrix specifications for spatio-temporal model fitting: a case study on COVID-19 data. Stoch. Environ. Res. Risk Assess. 0123456789, 271–282. http://dx.doi.org/10.1007/s00477-021-02077-y.

CAOP, C.A.O.d.P., 2020. DGT. URL: https://www.dgterritorio.gov.pt.

Coly, S., Garrido, M., Abrial, D., Yao, A.F., 2021. Bayesian hierarchical models for disease mapping applied to contagious pathologies. PLoS ONE 16 (1 January), 1–28. http://dx.doi.org/10.1371/journal.pone.0222898.

Franch-Pardo, I., Napoletano, B.M., Rosete-Verges, F., Billa, L., 2020. Spatial analysis and GIS in the study of COVID-19. A review. Sci. Total Environ. 739, http://dx.doi.org/10.1016/j.scitotenv.2020.140033.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2013. Bayesian Data Analysis, third ed. Chapman and Hall.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Stat. Comput. 24 (6), 997–1016. http://dx.doi.org/10.1007/s11222-013-9416-2, arXiv:1307.5928.

General Directorate of Health - NHS[Internet], 2021. DGS. URL: https://www.sns.gov.pt/.

Karaye, I.M., Horney, J.A., 2020. The impact of social vulnerability on COVID-19 in the U.S.: An analysis of spatially varying relationships. Am. J. Prev. Med. 59 (3), 317–325. http://dx.doi.org/10.1016/j.amepre.2020.06.006.

Knorr-Held, L., 2000. Bayesian modelling of inseparable space-time variation in disease risk. Stat. Med. 19 (17–18), 2555–2567.

Lawson, A.B., 2023. Evaluation of predictive capability of Bayesian spatio-temporal models for Covid-19 spread. BMC Med. Res. Methodol. 23 (1), 182. http://dx.doi.org/10.1186/s12874-023-01997-3.

Lee, D., 2020. A tutorial on spatio-temporal disease risk modelling in R using Markov chain Monte Carlo simulation and the CARBayesST package. Spat. Spatio-Temporal Epidemiol. 34, 100353. http://dx.doi.org/10.1016/j.sste.2020.100353.

Leroux, B.G., Lei, X., Breslow, N., 2000. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In: Halloran, M.E., Berry, D. (Eds.), Statistical Models in Epidemiology, the Environment, and Clinical Trials. In: The IMA Volumes in Mathematics and its Applications, vol. 116, Springer New York, New York, NY, pp. 179–191.

MacNab, Y.C., 2022. Bayesian disease mapping: Past, present, and future. Spat. Stat. http://dx.doi.org/10.1016/j.spasta.2022.100593.

MacNab, Y.C., 2023. Adaptive Gaussian Markov random field spatiotemporal models for infectious disease mapping and forecasting. Spat. Stat. 53, http://dx.doi.org/10.1016/j.spasta.2023.100726.

Mendes, J.M., Baptista, H., Oliveira, A., Jardim, B., de Castro Neto, M., 2022. Beyond comorbidities, sex and age have no effect on COVID-19 health care demand. Sci. Rep. 12 (1), 1–12. http://dx.doi.org/10.1038/s41598-022-11376-5.

Mingione, M., Alaimo Di Loro, P., Farcomeni, A., Divino, F., Lovison, G., Maruotti, A., Lasinio, G.J., 2022. Spatio-temporal modelling of COVID-19 incident cases using Richards' curve: An application to the Italian regions. Spat. Stat. 49 (June), http://dx.doi.org/10.1016/j.spasta.2021.100544.

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 1 (2), 17–23.

NTD, 2023. NTD, Modelling Consortium. URL: https://www.ntdmodelling.org.

Paez, A., Lopez, F.A., Menezes, T., Cavalcanti, R., Pitta, M.G.d.R., 2021. A spatio-temporal analysis of the environmental correlates of COVID-19 incidence in Spain. Geogr. Anal. 53 (3), 397–421. http://dx.doi.org/10.1111/gean.12241.

Portugal Statistics, 2021. INE. URL: https://ine.pt/.

Rushworth, A., Lee, D., Sarran, C., 2017. An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. J. R. Stat. Soc. Ser. C. Appl. Stat. 66 (1), 141–157. http://dx.doi.org/10.1111/rssc.12155, arXiv:1411.0924.

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

Slater, J.J., Brown, P.E., Rosenthal, J.S., Mateu, J., 2021. Capturing spatial dependence of COVID-19 case counts with cellphone mobility data. Spat. Stat. (xxxx), 100540. http://dx.doi.org/10.1016/j.spasta.2021.100540.

Stewart, R., Erwin, S., Piburn, J., Nagle, N., Kaufman, J., Peluso, A., Christian, J.B., Grant, J., Sorokine, A., Bhaduri, B., 2022. Near real time monitoring and forecasting for COVID-19 situational awareness. Appl. Geogr. 146 (July), 102759. http://dx.doi.org/10.1016/j.apgeog.2022.102759.

Vehtari, A., Gelman, A., Gabry, J., 2016. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. (September), 1–20. http://dx.doi.org/10.1007/s11222-016-9696-4, arXiv:1507.04544.

Wakefield, J., 2007. Disease mapping and spatial regression with count data. Biostatistics (Oxf. Engl.) 8 (2), 158–183.

Zhang, C.H., Schwartz, G.G., 2020. Spatial disparities in coronavirus incidence and mortality in the United States: An ecological analysis as of May 2020. J. Rural Health 36 (3), 433–445. http://dx.doi.org/10.1111/jrh.12476.