Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Spatio-temporal small area surveillance of the COVID-19 pandemic

Miguel A. Martinez-Beneito\textsuperscript{a,b,*}, Jorge Mateu\textsuperscript{c}, Paloma Botella-Rocamora\textsuperscript{d,b}

\textsuperscript{a} Department of Statistics and Operations Research, University of Valencia, Burjassot (Valencia), Spain
\textsuperscript{b} Unitat Mixta de recerca en mètodes estadístics per a dades biomèdiques i sanitàries, UV-FISABIO, Spain
\textsuperscript{c} Department of Mathematics, University Jaume I of Castellon, Castelló de la Plana, Spain
\textsuperscript{d} Subdirecció General de Epidemiologia, Vigilancia de la Salud i Sanidad Ambiental, Conselleria de Sanitat Universal i Salut Pública, Valencia, Spain

\section*{ARTICLE INFO}

\textbf{Article history:}
Received 11 April 2021
Received in revised form 21 October 2021
Accepted 21 October 2021
Available online 8 November 2021

\textbf{Keywords:}
COVID-19
Disease mapping
Instantaneous reproduction number
Spatio-temporal modelling

\section*{ABSTRACT}

The emergence of COVID-19 requires new effective tools for epidemiological surveillance. Spatio-temporal disease mapping models, which allow dealing with small units of analysis, are a priority in this context. These models provide geographically detailed and temporally updated overviews of the current state of the pandemic, making public health interventions more effective. These models also allow estimating epidemiological indicators highly demanded for COVID-19 surveillance, such as the instantaneous reproduction number $R_t$, even for small areas. In this paper, we propose a new spatio-temporal spline model particularly suited for COVID-19 surveillance, which allows estimating and monitoring $R_t$ for small areas. We illustrate our proposal on the study of the disease pandemic in two Spanish regions. As a result, we show how tourism flows have shaped the spatial distribution of the disease in these regions. In these case studies, we also develop new epidemiological tools to be used by regional public health services for small area surveillance.

© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

\textsuperscript{*} Corresponding author at: Department of Statistics and Operations Research, University of Valencia, Burjassot (Valencia), Spain.
\textit{E-mail addresses:} miguel.a.martinez@uv.es (M.A. Martinez-Beneito), mateu@mat.uji.es (J. Mateu), botella_pal@gva.es (P. Botella-Rocamora).

https://doi.org/10.1016/j.spasta.2021.100551
2211-6753/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Introduction

COVID-19 has emerged as a novel health problem, posing new challenges and pressure to health systems, in particular to public health services. Therefore, new tools and research are needed to solve the problems arising from this pandemic. On one hand, biomedical research is needed to develop new vaccines and new treatments that could clinically help handling the disease. On the other hand, new public health surveillance tools are also required to timely address the pandemic. Namely, it has been even stated: “countries should not ease restrictions until they have robust systems in place to closely monitor the infection situation” (Han et al., 2020). In this scenario, the development of effective epidemiological surveillance tools that allow focusing public health actions on particular locations and moments, is a priority. These tools are crucial in the current setting, where the public health resources demanded by the pandemic are frequently higher than those available.

A compulsory requirement of effective epidemiological surveillance systems is dealing with small enough study units, both in space and time. Surveillance systems working with either large spatial units or long time periods will be neither specific nor updated enough as to guide useful epidemiological actions. Nevertheless, handling small study units either in time or space poses statistical challenges, known in the literature as small areas estimation problems (Rao, 2003; Martinez-Beneito and Botella-Rocamora), that should be solved if reliable estimates are pursued. The small area and disease mapping literature could be useful for this goal, in particular the literature on spatio-temporal disease mapping models (Cressie and Wikle, 2011; Martinez-Beneito and Botella-Rocamora).

Most epidemiological surveillance systems use incidence rates to monitor the evolution of disease outbreaks. A vast literature can be found on smoothing rates in small areas, in particular standardised rates (Besag et al., 1991; Leroux et al., 1999). The COVID-19 pandemic has focused substantial attention also on the instantaneous reproduction number $R_t$ (Nishiura and Chowell, 2009). However, although $R_t$ estimates have been previously derived and studied for several areas of a single region of study (Smirnova et al., 2020; Islam et al., 2021), no spatial estimation proposal, suitable for small areas studies, has already been made for this parameter. For each period $t$, $R_t$ is interpreted as a temporally dynamic version of the basic reproduction number of the disease $R_0$ (Milligan and Barrett, 2014). The basic reproduction number $R_0$ is defined as the number of people that would be infected, on average, by a single infected person in a completely susceptible population. For $R_0 > 1$ the pandemic describes an explosive evolution, while for $R_0 < 1$ it tends to fade out. As made clear in this COVID-19 pandemic, where mitigation actions have been undertaken at different moments, there is not a unique static value for $R_0$ inherent to the disease (Wilasang et al., 2020; Fang et al., 2020). In contrast, $R_t$ shows substantial temporal variability and this is a key quantity that allows monitoring the spread of the disease at each moment of the study period, showing if the epidemiological control measures undertaken are enough, or not, to stop that spread. While surveillance systems frequently monitor the temporal variability of $R_t$, its geographical variability is usually ignored, at least at a small area level. As a consequence, specific epidemiological measures cannot be taken wherever they are required according to this criterion.

Estimating $R_t$ can be sometimes problematic, for example, in Cori et al. (2013) the scarce daily observed counts made them consider weeks as study units. Specifically, for estimating $R_t$ on day $t$ all cases observed during the week before that day were all gathered. As a consequence, they used consecutive overlapping time windows for their analysis (one for each $R_t$), ignoring the temporal dependence that this could induce, what could impact, for example, on the confidence interval widths for the different $R_t$’s. Moreover, the weekly basis of the data makes $R_t$ estimates less sensible to the latest incidence changes that data could show. This makes the interest of $R_t$ decrease as an epidemiological surveillance tool. This modelling approach is implemented in the EpiEstim (Cori, 2020) R package, used by many public health institutions during the COVID-19 pandemic (Tebé et al., 2020; Centro Nacional de Epidemiología, 2020).

Beyond data problems arising, for example, from inaccuracies in reporting locations of cases when small areas are considered, estimation problems in $R_t$ of the kind above are mainly due to the low number of observed cases per time unit when those areas are small. However, these problems
would be worsened if many small spatial units were considered, instead of a single large unit, since those units would split the observed daily cases into smaller quantities. In that case, a small area modelling framework would be required to yield sensible incidence estimates. Otherwise, those estimates would just reproduce the high amount of noise that such small numbers frequently show.

In this paper, we propose a small area spatio-temporal model specifically devised for COVID-19 incidence data. With our proposal, we seek to develop advanced surveillance tools that allow monitoring COVID-19 for small areas. In practice, this would allow implementing specific epidemiological control measures where and when they are required according to detailed and specific information. Additionally, our spatio-temporal model will help visualising the dynamics of the disease and how their different control measures shape its spatial distribution.

In addition, we apply the mentioned spatio-temporal modelling approach to two real COVID-19 data sets. These data correspond to COVID-19 surveillance in two Spanish regions with a high spatial and temporal disaggregation. Specifically, we have daily incident cases for detailed divisions of both regions, with less than 10,000 people on average per spatial units, and a substantially lower population in many particular cases. The application of our model to these data sets illustrates its impact on practical terms.

This paper is organised as follows. Section 2 introduces a novel spatio-temporal modelling proposal for COVID-19 surveillance. Section 3 illustrates the application of that model to the analysis of the COVID-19 pandemics in two Spanish regions. This section shows the real use in practice of our model and the corresponding results. Finally, Section 4 discusses some aspects of the model introduced and its general use for COVID-19 surveillance.

2. Methodology

We consider the following general setting found for many small areas COVID-19 surveillance systems. Let us assume that we have a region of study divided into \( I \) (small) spatial units observed during \( J \) consecutive days. We say that spatial units are small if they produce unreliable statistical estimates as a consequence of their small population (Rao, 2003). Let \( O_{ij} \) denote the observed number of incident COVID-19 cases for spatial unit \( i \) (\( i = 1, \ldots, I \)) and day \( j \) (\( j = 1, \ldots, J \)). Given an underlying process \( \lambda \), we consider the observed cases follow independent Poisson distributions \( O_{ij} \sim \text{Poisson}(\lambda_{ij}) \), and model the underlying process as

\[
\log(\lambda_{ij}) = \log(P_i) + \gamma_{\text{DoW}(j)} + (\beta X)_{ij} + \epsilon_{ij},
\]

where:

- \( P_i \) is the population living in spatial unit \( i \) during the study period. \( P_i \) is assumed constant in time. If age-structured information was available, \( P_i \) could be substituted by (age standardised) expected cases in order to remove age effects on incidence.
- \( \gamma \) is a vector, of length 7, modelling the cyclic effect of week days on incidence rates. The term \( \text{DoW}(j) \) denotes a function of \( j \) that returns the day of the week, as a value between 1 and 7.
- \( X \) is a \( K \times J \) design matrix, where each column corresponds to a different day and each row corresponds to a function of a basis, which models the time trends of the incidence rates for the different spatial units. \( X_{ij} \) stands for the value of the \( k \)th function in the basis for the \( j \)th day. This matrix is fixed by design and no inference is made on \( X \).
- \( \beta \) is an \( I \times K \) matrix with the coefficients of the basis functions for each spatial unit. The variability between different rows of \( \beta \) allows the incidence rates of the spatial units to follow different trends.
- \( \epsilon \) is an \( I \times J \) matrix of independent random effects. This unstructured term is intended to model the overdispersion of the Poisson process beyond the variability induced by the rest of terms in the model. COVID-19 incident cases, for each spatial unit, tend to cluster in some particular days due to either the outbreak nature of these cases that makes them be grouped or to some other artefacts, such as the sampling of several suspicious cases around the first index case of an outbreak, which will be often tested in the same day. These artefacts may produce abnormally high numbers of cases for some particular days and locations that we try to control with this term.
The spatio-temporal variability of the process is modelled mainly by the product $\beta X$. This product relies on an appropriate basis of functions, collected in $X$, modelling the temporal trends of the disease. Once these functions are chosen, the coefficients $\beta$ combine them to describe a different time trend for each spatial unit. The number of functions $K$ in the basis may be different from the number of days $J$ of the study period, in particular it could be substantially lower. This makes the time trend for each spatial unit depend on just a few parameters $K$, lower than the number of observations on that spatial unit $J$, making a low-dimensional fit of time trends, with evident computational advantages.

Although the model above is stated for a general basis of functions $X$, in our case, we find the use of splines particularly convenient for $X$. In particular, we propose using a basis of natural cubic B-splines. Besides its computational convenience, B-splines have also compact support, thus these functions fit the local performance of the incidence curve during a particular period of days, making the interpretation of the coefficients $\beta$ straightforward. Moreover, natural splines provide a particularly sensible fit of the incidence rates at the extremes of the study period, specifically at the end of that period. This is particularly convenient for surveillance as, in that case, attention is particularly put on those precise days of the study period.

Our proposal models the time trends of the incidence rates using splines. In addition, spatial dependence is induced on those rates through the matrix of spline coefficients $\beta$. Specifically, we model those cells as $\beta_{ik} = \mu_k + \beta_{ik}'$, where $\mu_k$ stands for the mean value, throughout the entire region, of the coefficients corresponding to the $k$th spline in the basis. The vector $\mu = (\mu_1, \ldots, \mu_K)$ accounts for variations in the overall time trend during the study period. Additionally, $\beta^\prime$ models the spatial variability of the different local incidence rates. This term allows each spatial unit to follow a different time trend. In particular, we model their columns as Gaussian Markov random fields. Specifically, we chose the Leroux et al. (1999) CAR distribution for inducing spatial dependence on the incidence rates, i.e., for each column $k$ we assume the following set of conditional distributions

$$\beta_{ik}^* | \beta_{(-i)k}^* \sim N\left( \frac{\rho_k}{1 - \rho_k + \rho_k n_i} \sum_{i' \neq i} \beta_{i'k}^* \frac{(\sigma_{\beta}^2)_k}{1 - \rho_k + \rho_k n_i}, \frac{(\sigma_{\beta}^2)_k}{1 - \rho_k + \rho_k n_i} \right),$$

where $\beta_{(-i)k}^*$ denotes all the components of the column $\beta_{k}^*$, excepting its $i$th cell, and $i' \sim i$ denotes all spatial units $i'$ adjacent to spatial unit $i$ and $n_i$ denotes its number of neighbours. Parameters $\rho_k$ take values in the interval $[0, 1]$, and they may reproduce either a heterogeneous process for $\rho_k = 0$, an intrinsic CAR distribution for $\rho_k = 1$ or intermediate process, with stronger or weaker spatial dependence, for $\rho_k \in (0, 1]$. Considering different $\rho_k$ and $(\sigma_{\beta})_k$ parameters for each element in the basis of functions allows reproducing different spatial strengths and variabilities, respectively, for any moment of the study period. Therefore, this reproduces a non-separable spatio-temporal correlation structure (Torres-Avilés and Martinez-Beneito, 2015). These assumptions seem reasonable when the pandemic could show different spatial features at their different phases or waves.

Beyond the long-term temporal variability modelled by $\beta X$, the term $\gamma$ models the cyclic variability within weeks, in particular the different incidence rates that weekends have compared to weekdays. We consider the DoW function equal to 1 for the first day of study and it will increase, day by day, until 7, and this pattern will be subsequently repeated. The first element of $\gamma$, $\gamma_1$, is fixed to 0 and the rest of elements model the departures of the incidence rates of the rest of week days with respect to the first. Improper uniform prior distributions on the whole real line are assigned to $\gamma_2, \ldots, \gamma_7$.

Finally, regarding the cells of the matrix of random effects $\epsilon$, they are modelled as independent $N(0, \sigma_\epsilon^2)$ variables. In this case, $\sigma_\epsilon^2$ is also estimated within the model and a Uniform(0, $c$) distribution, for a large enough (non informative) value $c$, is assigned to the standard deviation $\sigma_\epsilon$.

2.1. Small area estimation of the instantaneous reproduction number

Beyond incidence rates, our proposal provides small area estimates of the instantaneous reproduction number $R_t$ (which is of particular interest for epidemiological surveillance) for all days $t$
and for small area divisions of the region of interest. Specifically, we estimate $R_{it}$, the instantaneous reproduction number for each (small) spatial unit $i$ and each day $t$ of the study period.

In a non-spatial setting, leaving aside the influence of the prior distribution used for the different $R_t$'s, the popular R package EpiEstim estimates those $R_t$'s as $O_t/(\sum_{s=1}^{S} O_{t-s} w_s)$, where $O_t$ denotes the incident cases for the entire region at day $t$. In this expression $w_s$ denotes the infectivity function that quantifies, for each lag time $s$, the probability of observing an $s$-day difference between two related cases, one primary and one secondary case of the disease, and $S$ is an upper bound for the subindex $s$. The incident cases $O_t$ could be quite low for some days of the period of analysis, even for the entire study region, what could make those $R_t$'s noisy and unreliable. As a potential solution, EpiEstim authors propose using overlapping time windows of size $\tau$ instead of single days, and therefore estimate $R_t$ as

$$
\left( \sum_{k=t-\tau+1}^{t} O_k \right) / \left( \sum_{k=t-\tau+1}^{t} \sum_{s=1}^{S} O_{k-s} w_s \right).
$$

This illustrates how $R_t$ estimation poses particular problems when small counts are observed. This will be frequent when working with small sized spatial units, making the use of specific methods needed if reproduction numbers were calculated for small areas. According to the disease mapping literature, if $O_t \sim \text{Poisson}(E_t \lambda_t)$ are the observed counts of some disease at the $t$th spatial unit, and $E_t$ the corresponding expected counts, the Standardised Incidence Ratio (SIR) for that unit is defined as $\text{SIR}_t = O_t/E_t$. This risk estimate shows similar small area problems to those of $R_t$'s, arising from the small values that $O_t$ and $E_t$ take. These problems are normally solved by considering a smoothed version of those SIRs defined as $\text{SIR}_t = \lambda_t/E_t$, where $\lambda_t$ is estimated within a model, which typically considers spatial dependence on the different $\lambda_t$'s. We follow now a similar approach for deriving smoothed $R_t$'s that could solve the small area problems of the corresponding unsmoothed indicator.

In parallel to the EpiEstim approach of Cori et al. (2013), and following the disease mapping smoothing ideas mentioned, we can estimate $R_{it}$ for the small area $i$ and day $t (= S + 1, \ldots, J)$ as $R_{it} = \lambda_{it} / (\sum_{s=1}^{S} \lambda_{i(t-s)} w_s)$, where $\lambda_{it}$ corresponds now to the expected incident cases for that spatial unit and day according to our model. We could set $\lambda_{it}$ equal to $\exp(\log(P_i) + \gamma_{\text{Dow}(t)} + (\beta X)_{it} + \epsilon_{it})$, or simply $\lambda_{it} = \exp(\gamma_{\text{Dow}(t)} + (\beta X)_{it} + \epsilon_{it})$ since the population term will cancel out in the numerator and denominator of $R_{it}$. Nevertheless, we could also remove the cyclic term in the latest expression, that is $\lambda_{it} = \exp((\beta X)_{it} + \epsilon_{it})$ since, in this manner, we will filter out the cyclic effect that we could find in $R_{it}$ as a consequence of that same artefact on the raw incidence figures and that we will not typically want to reproduce. Moreover, if $\lambda_{it}$ was simply defined as $\exp((\beta X)_{it})$, we would obtain a parsimonious version of $R_{it}$ where the particular data collecting features that could be influencing $O_t$ are filtered out and just the smooth spatio-temporal trend captured by the spline process would be kept. Thus, a parsimonious smoothed instantaneous reproduction number $R_{it}$ suitable for small areas could be obtained as

$$
R_{it} = \left( \exp((\beta X)_{it}) \right) / \left( \sum_{s=1}^{S} \exp((\beta X)_{i(t-s)} w_s) \right).
$$

If MCMC inference is performed on the spatio-temporal model, this allows calculating the different smoothed $R_t$'s at every MCMC step and, therefore, draw confidence bands for this parameter for every spatial unit $i$. This enables to calculate also $P(R_{it} > 1)$ for each $i$ and $t$, and therefore assess the probability of the epidemic to be out of control at every location and moment of the study period. This quantity would be of evident interest for epidemiologists.

3. Spatio-temporal modelling of COVID-19 in two Spanish regions

We apply now our proposal to the study of the COVID-19 pandemic in 2 of the 17 Spanish regions, Castilla y Leon (CyL) and Comunidad Valenciana (CV), with 2.3 and 5 million inhabitants, respectively.
The CyL COVID-19 data are publicly available. The study period for this region covers the interval from March 6th to October 14th, 2020, when those data were downloaded. CyL COVID-19 data are published at the health zone level, an administrative division corresponding to the geographical units corresponding to the primary care centres of this region. CyL is divided into 247 health zones, with populations ranging from 441 to 37509 inhabitants. The data set contains daily observed COVID-19 counts for each CyL health zone. The data set used for this analysis is supplied as supplementary material to the paper to make reproducible this part of the analysis. This data set and the rest of supplementary material are also available.

CV data have been supplied, under request, by the Health Administration of this region. As for CyL, the data set contains daily observed COVID-19 counts for the 542 municipalities of this region. Municipalities in CV are, in administrative terms, towns whose populations range from 17 to 794288 inhabitants in 2019. The median population per municipality is 1412 inhabitants. In this case, the study period covers again the interval from March 6th to October 14th, 2020.

MCMC Inference was performed by using WinBUGS. Additionally, the R package pbugs (Vergara-Hernández and Martinez-Beneito) was used for running in parallel 5 MCMC chains for each run. 2000 MCMC draws were first simulated as burn in, followed of 5000 additional draws per chain. Finally, 1 out of each 15 draws were saved, yielding a posterior sample of 1000 draws. Convergence was assessed by means of the Brooks–Gelman–Rubin statistic and the number of effective simulations (Gelman et al., 2014), which are provided by the pbugs package. Full details on the model implementation, and the rest of results shown below, can be found as a R Markdown document at the supplementary material.

For the modelling of the long-range time trend we chose a natural B-spline basis with one node every 2 weeks and 2 additional nodes at the borders of the study period. Thus, for 223 days of study, the basis contains 17 spline functions scattered around the study period. These 17 functions yield enough flexibility for modelling the rates time trends and reduce the computational burden of the model compared to a weekly basis, which would contain 33 splines modelling the time trends of the rates, and therefore 33 spatial processes within the model. Regarding the infectivity function \( w_s \), required for calculating the instantaneous reproduction numbers, we assume a shifted Gamma distribution of mean 4.7 and standard deviation 2.9. These values are suitable for COVID-19 incidence according to Nishiura et al. (2020). The maximum temporal lag for calculating \( R_t \) has been fixed to \( S = 25 \) since for that value the distribution assumed for \( w_s \) yields \( \sum_{i=1}^{25} w_s > 0.9999 \), that is, COVID-19 cases are hardly contagious after those 25 days.

We separately run the spatio-temporal model for these two data sets. For both, we considered a single spatial parameter \( \rho \) instead of different parameters \( \rho_k \) for each element in the basis, since we did not find evidence of needing different spatial parameters. Specifically, we run an additional model with different \( \rho_k \)'s per basis function and the posterior confidence intervals of these parameters substantially overlapped. In our model, with a common \( \rho \), we obtained a posterior mean for \( \rho \) of 0.51 (95% credible interval: [0.43,0.60]) for CyL, while for CV that posterior mean was 0.45 ([0.36,0.57]), so we found a mild spatial dependence for both data sets. In contrast, we found evidence of needing different variance parameters for the spatial processes of each of the basis functions. These different variances allow the spatial distribution of the disease to show different variabilities for each moment of the study period. Thus, considering different \( \sigma_k \)'s allows the spatio-temporal process to be heteroscedastic in time. Specifically, for CyL the spatial standard deviations \( \sigma_k \) range from 0.88 (95% credible interval [0.65,1.14]) to 2.14 ([1.76,2.56]), while for CV, they range from 1.26 ([0.193]) to 3.29 ([2.76,3.90]).

Regarding the overdispersion term \( \epsilon \), its standard deviation has a posterior mean of 0.56 (95% credible interval [0.55,0.57]) for CyL and of 0.77 ([0.75,0.79]) for CV. This shows how the sampling artefacts, which make incident cases cluster in some particular days and locations, are stronger in CV than for CyL. Regarding the cyclic weekly term \( \gamma \), we found a difference of 0.94 units between the day with higher and lower incidence for CyL, and of 1.16 in CV. The 95% credible intervals for both days, for both regions, were clearly disjoint, highlighting the need of this term within the model.

---

1. https://analisis.datosabiertos.jcyl.es/pages/coronavirus/
2. https://www.uv.es/mamtnez/ETCovid.html
Fig. 1. Overall time trends for the observed daily cases for both regions, Comunidad Valenciana and Castilla y León in this order (top row). Estimated $R_t$ evolution for both regions (bottom left plot), and daily rates per 100,000 inhabitants for 4 municipalities in CV (bottom right plot). Vertical lines correspond to the starting day of each month of the period of study. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 1 shows some results of the spatio-temporal model for both regions. The upper row shows, for both regions, the posterior mean for the daily observed cases, with the corresponding 95% credible bands. We first note the very different scales for both regions, even though the CV population is more than twice the CyL population. The blue line in these plots corresponds to the time trends, which take into account both the long-range time trend (splines) and cyclic weekly terms. The effect of the cyclic term is evident. On the other hand, the black line reproduces just the time trend corresponding to the spline term. These curves filter out the cyclic pattern that in general we will not be interested to reproduce. Both spline curves seem to capture the time trend of the disease, filtering the artefacts that could make the observed cases for one particular day deviate below or above that trend. From now on, the rest of results shown will be based only in the spline component of the time trend in order to filter out the cyclic effect that we will typically want to remove.

The bottom left plot of Fig. 1 shows the daily evolution of $R_t$ for both regions. This plot shows the overall time trend for $R_t$, for the entire regions, and their corresponding 95% credible bands. Note the effect of the different number of observed cases for each region, which makes the bands for CV to be substantially wider than for CyL. The daily $R_t$’s in this plot have been calculated by means of the methodology described in the previous section for each municipality and averaging all the $R_t$’s throughout the entire region. These averages are weighted according to the different populations of the spatial units. These curves have a high epidemiological value since they allow knowing if the
pandemic is either increasing ($R_t > 1$) or decreasing ($R_t < 1$) at any moment of the study period. Moreover, if wanted, we could also assess the probability of any of these two states. These curves show important differences in the pandemic diffusion for both regions, for example the different starting times of the second wave of the pandemic, which started way sooner for CV than for CyL.

The bottom right plot of Fig. 1 shows the rates (per 100,000 people) time trends for 4 CV municipalities. Rates, in contrast to the fitted observed cases, allow comparisons between regions regardless of their population sizes. These municipalities correspond to Valencia and Alicante, the two most populated cities in CV and placed 166 kilometres away. Additionally, we also show the rates for Massalfassar and Aigues, the two least populated neighbouring municipalities of Valencia and Alicante, respectively. The rate curve for Massalfassar resembles that of Valencia, similarly to Aigues and Alicante since, for example, they show concurrent peak epidemics. This is a consequence of the spatial processes in our model. Although these curves are similar, we can see evident departures within both Massalfassar-Valencia and Aigues-Alicante. For example, the first wave in Aigues showed far lower rates than for Alicante, besides these cities are neighbours and one of them has a population 300 times higher than the other, what could make the time trend of Alicante dominate that of Aigues. Something similar happens for the second wave of Valencia and Massalfassar.

Fig. 2 shows, once again for Valencia, Alicante, Massalfassar and Aigues, the posterior mean for the daily observed cases (left-hand-side of the figure), with the corresponding 95% credible bands. The right-hand-side of this figure shows the time trend of $R_t$ for these municipalities. The plots of the observed cases for Valencia and Alicante show a good fit, thus the biweekly spline basis used seems enough for describing the temporal variability of the observed cases in these large cities. On the other hand, for Massalfassar and Aigues, the observed cases are quite low due to their small size. For these municipalities, a time trend is fitted that merges the information on their observed cases and that provided by their neighbours. Thus, the time trend of Massalfassar resembles that of Valencia, with a long right tail at the second wave of the pandemic, larger than that of Aigues, which resembles the time trend of Alicante. Interestingly, Aigues does not have any observed case at the first wave of the pandemic, what explains the discrepancy at this wave between Alicante and Aigues shown in Fig. 1.

Regarding the municipal $R_t$ time trends, the longer widths of the Massalfassar and Aigues 95% credible intervals stand out compared to those of Valencia and Alicante. For example, during the second wave in Massalfassar and Aigues, the credible band for these days usually contain the value 1. In contrast, for Valencia and Alicante, most of the July and August days showed $R_t$’s significantly higher than 1. This is due to the different populations, and therefore daily observed cases, of these municipalities. In addition, once again the $R_t$ curves of neighbouring municipalities resemble each other. See for example the curves of Alicante and Aigues that describe similar plateaus and peaks from May to mid-September. This is, evidently, a consequence of the spatial dependence induced in the model since the observed cases in Aigues are not enough to describe such a detailed time trend.

In addition to smoothed spatio-temporal epidemiological indicators, as the case of $R_t$, the spatio-temporal model proposed yields also surveillance tools of high epidemiological value. For example, Fig. 3 shows, for each region, a correlation plot for the coefficients of the spline basis functions. The row and column names in the plot correspond to the day showing the peak of the corresponding spline function. It can be seen how, for both regions, all coefficients before July show high correlations, regardless of the moment of the corresponding peak, while this setting changes with the starting of July. This suggests that the geographical pattern of the pandemic was steady during its first wave, when confinement measures were taken. However, at the end of that wave, coinciding with the lifting of mobility restrictions (active from March 14th to June 21st), the spatial pattern of the disease changed. In the case of CV, this change was abrupt, the spatial patterns from July hardly show any correlation with the spatial patterns before that month, but milder for CyL. These different patterns may be explained by the large touristic sector of CV, whose coastline attracts millions of tourists each summer. This sector attracted thousands of tourists coinciding with the starting of the summer and the lifting of restrictions, which could make its spatial pattern change completely, in contrast to CyL.
Fig. 2. Time trend of the observed daily cases for Valencia, Alicante, Massalfassar and Aigues, in this order, (left column) and time trend of $R_t$ for these same municipalities (right column). Vertical lines correspond to the starting day of each month of the period of study.

Fig. 3 shows some additional differences between CV and CyL spatial pattern evolutions. For example, in CV, we see from July a dynamic evolution of the pandemic since the spatial patterns of the spline functions in this period show hardly any correlation between consecutive spline functions. Thus, the pandemic in CV during this period moves around randomly, in contrast to the first phase of the pandemic when population was confined. In contrast, for CyL correlations show overlapping diagonal squares, which point out a less haphazard evolution of the disease. These plots could be further interpreted with the aid of some additional results. For example, the final figure of the accompanying supplementary material shows how the first wave of the pandemic in CyL was mainly urban, while the second wave, at least its summer period, showed higher rates in the rural areas. This suggests again that tourism has possibly taken an important effect in the second pandemic wave also in this region since rural zones receive most of the tourism in this zone of Spain, mainly during the summer period.

Fig. 4 illustrates some of the small area analysis possibilities of our spatio-temporal model. As already discussed, both rates and $R_t$’s are the main goals of most COVID-19 epidemiological analyses. These two indicators measure, respectively, the current state of the pandemic and its time trend, a kind of derivative of the incidence curve. Therefore, these two quantities are complementary indicators of the pandemic. Joint summaries of these two indicators have been proposed as synthetic
Fig. 3. Correlation plots for the coefficients of the basis functions for Comunidad Valencia (left) and Castilla y León (right). Names of rows and columns correspond to the day (month/day format) showing the peak of the corresponding spline function.

Fig. 4. Risk levels according to smoothed weekly incidence rates and $R_t$’s (left) and its geographical representation (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

measures of the pandemic “risk”. The joint representation of these two indicators in a single plot has been proposed and used in an applied epidemiological context, but always for large areas (López Codina, 2020), due to the statistical problems arising when small areas are considered. Nevertheless, our spatio-temporal model would make it possible to derive tools of that kind for small area problems, as shown in Fig. 4.

https://biocomsc.upc.edu/en/Covid-19/Risk%20Diagrams
The left-hand side of Fig. 4 shows a joint plot of the incidence rate for the last 7 days, jointly with the $R_t$ for each CV municipality on the latest day of the study period. Results could also be reproduced for previous days yielding therefore a dynamic view of the municipal risks. Incidence rates for the last 7 days have been defined as 7 times the incidence rate for the latest day of study to represent the most updated information on the disease, according to the available data. Colour cuts in this plot have been set as the different official risk levels for each of these indicators set by the Spanish government (Consejo Interterritorial and Sistema Nacional de Salud, 2020), these are $\{10, 25, 75, 125\}$ for the incidence rates and $\{1, 1.1, 1.5, 2\}$ for the $R_t$'s. These cuts could be alternatively defined as a function of sample percentiles, for example, if this was found more convenient. Colours in Fig. 4 have been chosen to represent different levels of increasing risk (higher as we move towards the upper-right side of the plot), varying from greens (low risks) to reds (high risks). This plot shows how the highest rate corresponds to a municipality with around 1000 incident cases per 100,000 people and the highest $R_t$ is above 2.5 for one municipality.

The right-hand side of Fig. 4 shows the geographical mapping of the risks shown in the left-hand side of this figure. Municipalities are coloured in the right plot with the same risk level colours than in the left side plot. This choropleth map shows an evident spatial pattern, highlighting several regions of particularly high or low risks. This spatial pattern arises as a consequence of the spatio-temporal dependence in the model, since otherwise municipalities would show a more noisy pattern. This map would enable CV epidemiological authorities to focus on those municipalities with higher risks.

4. Conclusions

This paper introduces small area spatio-temporal analyses for COVID-19 epidemiological surveillance. Although the model introduced has been applied to Spanish data, it could be also applied to any other region with similar available data. Only population and daily incident cases data are required to run this model, which does not seem strong requirements.

We would like to mention an epidemiological limitation of the analysis conducted in this paper. It is widely known that COVID-19 incidence data are not comparable for both epidemic waves, at least in Spain. The lack of diagnostic tests of the disease during the first pandemic wave makes the available data for that period underestimate the height of that wave, making both waves non-comparable in absolute terms. Nevertheless, the joint analysis of the entire epidemic period (both waves) has allowed performing an overall analysis of the spatial distribution of the disease during all that period. In principle, the spatial distribution of the disease would not depend on the lack of diagnostic tests of the disease since this was something general for all CV and CyL. Nevertheless, direct comparisons of the strengths of both epidemic waves should be, and have been, avoided.

One interesting aspect of our proposal is its computational performance. The dimensions of the data sets analysed for this paper are large, with $120,286 (= 223 \cdot 542)$ and $55,081 (= 223 \cdot 247)$ spatio-temporally correlated observations. Moreover, we have used a heteroscedastic spatio-temporal dependence structure for our analyses, which makes the underlying process more flexible but computationally more demanding. This has been solved by the dimensionality reduction achieved with the use of the spline basis, which has reduced the number of spatial patterns to be fitted from 223 to just 17. This makes the fit substantially faster than spatio-temporal models with one spatial process per time unit (Martinez-Beneito et al., 2008). Nevertheless, the computing times required for fitting both data sets have been high (16.2 and 7.9 h), respectively. These computing times could be possibly improved using other MCMC sampling engines, such as Nimble or STAN. Nevertheless, for the usual case of being surveillance the main purpose of the study, in particular surveillance of the latest day of analysis, we would not need to fit such a long study period. Most of the natural B-spline basis functions have compact support so, for the latest day of study, increasing the period of analyses does not yield any improvement. The first splines in the basis will not take any effect on the final part of the study period. Thus, reducing the days of analysis to the latest 56 days would reduce the spline basis to just 5 functions, cutting down computing time (for CV this reduces computing time to just 1.16 h), and producing an equivalent fit for the latest day of study.

Most current COVID-19 epidemiological surveillance systems use incident rates for the last 7 or even 14 days. Using 14-day rates may seem justified as a proxy to disease prevalence,
assuming that the disease lasts around those days. Similarly, 7-day rates seem also a reasonable option for removing the weekly cyclic character that the data could show. Moreover, these two rates consider several days, increasing therefore the number of observed cases used to compute each rate and \( R_t \), alleviating potential small area problems. Nevertheless, despite these appealing features, epidemiological surveillance with 7 or 14-day rates does not seem an optimal choice in terms of timeliness, since these rates take into account substantial outdated information. Effective surveillance systems need data as updated as possible and that feature is lost with the temporal aggregation of these time periods. In this sense, our model avoids those temporal aggregations, making COVID-19 surveillance more effective.

Spatio-temporal analyses of COVID-19 pandemics provide valuable insight on the disease dynamics. As illustrated, these analyses show the effect of tourist activities on the spread and spatial distribution of the disease. Moreover, we have developed enhance surveillance tools to monitor and control the advance of the disease. Nevertheless, these analyses are useful only if small areas are used. If large areas were analysed instead, results and conclusions would be much less detailed and powerful, what would reduce their impact. Therefore, the development of small area analysis tools seems necessary. This work is just a step in that direction.

References

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Stat. Math. 43, 1–21. http://dx.doi.org/10.12831/800116466.

Centro Nacional de Epidemiología, Instituto de Salud Carlos III, https://cneovid.isciii.es/covid19/, Accessed on April 11th, 2021.

Consejo Interterritorial, Sistema Nacional de Salud, Coordinated response actions for the transmission control of covid-19 [in spanish], https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/sanidad14/Documents/2020/221020_ActuacionesrespuestaCOVID.pdf.

Cori, A., 2020. Epiestim: Estimate time varying reproduction numbers from epidemic curves, R package version 2.2-3. URL https://CRAN.R-project.org/package=EpiEstim.

Cori, A., Ferguson, N., Fraser, C., Cauchemez, S., 2013. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am. J. Epidemiol. 178 (9), 1505–1512. http://dx.doi.org/10.1093/aje/kwt133, (Sep. 2013).

Cressie, N., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. John Wiley & Sons.

Fang, Y., Nie, Y., Penny, M., 2020. Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis. J. Med. Virol. 92 (6), 645–659. http://dx.doi.org/10.1002/jmv.25750, (mar 2020).

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2014. Bayesian Data Analysis, third ed. Chapman & Hall/CRC, Boca Raton.

Han, E., Tan, M., Turk, E., Sridhar, D., Leung, G., Shibuya, K., Asgari, N., Oh, J., García-Basteiro, A., Hanefeld, J., Cook, A., Hsu, L., Teo, Y., Heymann, D., Clark, H., McKee, M., Legido-Quigley, H., 2020. Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe. Lancet http://dx.doi.org/10.1016/s0140-6736(20)32007-9.

Islam, A., Sayeed, M., Rahman, M., Zamil, S., Abedin, J., Saha, O., Hassan, M., 2021. Assessment of basic reproduction number \((r_0)\), spatial and temporal epidemiologic determinants, and genetic characterization of sars-cov-2 in Bangladesh. Infect. Genet. Evol. 92, 104884. http://dx.doi.org/10.1016/j.meegid.2021.104884, (2021). URL https://www.sciencedirect.com/science/article/pii/S1567134821001817.

Leroux, B.G., Lei, X., Breslow, N., 1999. 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. Springer, Berlin. ISBN New York.

López Codina, D., 2020. Tools for early detection and risk assessment. In: COVID-19 & Response Strategy, ISGlobal. (2020), URL https://www.isglobal.org/en/-/como-hacer-frente-a-los-nuevos-brotes-de-la-covid-19-.

Martinez-Beneito, M., Botella-Rocamora, P., Disease Mapping from foundations to multidimensional modelling. Martinez-Beneito, M.A., López-Quílez, A., Botella-Rocamora, P., 2008. An autoregressive approach to spatio-temporal disease mapping. Stat. Med. 27, 2874–2889.

Milligan, C., Barrett, A., 2014. Vaccinology: An Essential Guide. John Wiley & Sons, (2014), URL https://www.ebook.de/de/product/23507913/vaccinology.html.

Nishiura, H., Chowell, G., 2009. The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In: Mathematical and Statistical Estimation Approaches in Epidemiology. Springer, Netherlands, pp. 103–121. http://dx.doi.org/10.1007/978-90-481-2313-1_5, (2009).

Nishiura, H., Linton, N.M., Akhmetzhanov, A.R., 2020. Serial interval of novel coronavirus (COVID-19) infections. International Journal of Infectious Diseases 93, 284–286. http://dx.doi.org/10.1016/j.ijid.2020.02.060, (Apr. 2020).

Rao, J.N.K., 2003. Small Area Estimation. John Wiley & Sons.

Smirnova, A., Sterrett, N., Mujica, O.J., Munayco, C., Suárez, L., Viboud, C., Chowell, G., 2020. Spatial dynamics and the basic reproduction number of the 1991–1997 cholera epidemic in peru. PLOS Negl. Trop. Dis. 14 (7), 1–22. http://dx.doi.org/10.1371/journal.pntd.0008045, (07 2020).
Tebé, C., Valls, J., Satorra, P., Tobías, A., 2020. COVID19-world: a shiny application to perform comprehensive country-specific data visualization for SARS-CoV-2 epidemic. BMC Med. Res. Methodol. 20 (1), http://dx.doi.org/10.1186/s12874-020-01121-9.

Torres-Avilés, F., Martinez-Beneito, M.A., 2015. STANOVA: A smooth-ANOVA-based model for spatio-temporal disease mapping. Stochastic Environmental Research and Risk Assessment 29 (1), 131–141. http://dx.doi.org/10.1007/s00477-014-0888-1.

Vergara-Hernández, C., Martinez-Beneito, M.A., pbugs, https://github.com/fisabio/pbugs.

Wilasang, C., Sararat, C., Jitsuk, N., Yolai, N., Thammawijaya, P., Auewarakul, P., Modchang, C., 2020. Reduction in effective reproduction number of COVID-19 is higher in countries employing active case detection with prompt isolation. J. Travel Med. 27 (5), http://dx.doi.org/10.1093/jtm/taaa095.