medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Spatial prediction of air pollution levels using a hierarchical Bayesian spatiotemporal model in Catalonia, Spain

Marc Saez[1,2], Maria A. Barceló[1,2]

[1] Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Girona, Spain
[2] CIBER of Epidemiology and Public Health (CIBERESP), Madrid, Spain

Corresponding author:

Prof. Dr. Marc Saez, PhD, CStat
Research Group on Statistics, Econometrics and Health (GRECS)
and CIBER of Epidemiology and Public Health (CIBERESP)
University of Girona
Carrer de la Universitat de Girona 10, Campus de Montilivi
17003 Girona, Spain
Tel 34-972-418338, Fax 34-972-418032
http://www.udg.edu/grecs.htm e-mail: marc.saez@udg.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Authorship

MS had the original idea for the paper and designed the study. The bibliographic search and the writing of the introduction were carried out by MS and MAB. The methods and statistical analysis were chosen and performed by MS. MAB created the tables and figures. MS and MAB wrote the results and the discussion. The writing and final editing was done by all authors. MS and MAB reviewed and approved the manuscript.

Highlights

We show a hierarchical Bayesian spatiotemporal model.

Our model provides predictions of both long-term and short-term exposure.

The computational cost is low.

The model only needs a minimum number of stations being distributed throughout the territory.

The other requirement of our model is that the spatial and temporal dimensions are either independent or separable.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Our objective in this work was to present a hierarchical Bayesian spatiotemporal model that allowed us to make spatial predictions of air pollution levels in an effective way and with very few computational costs.

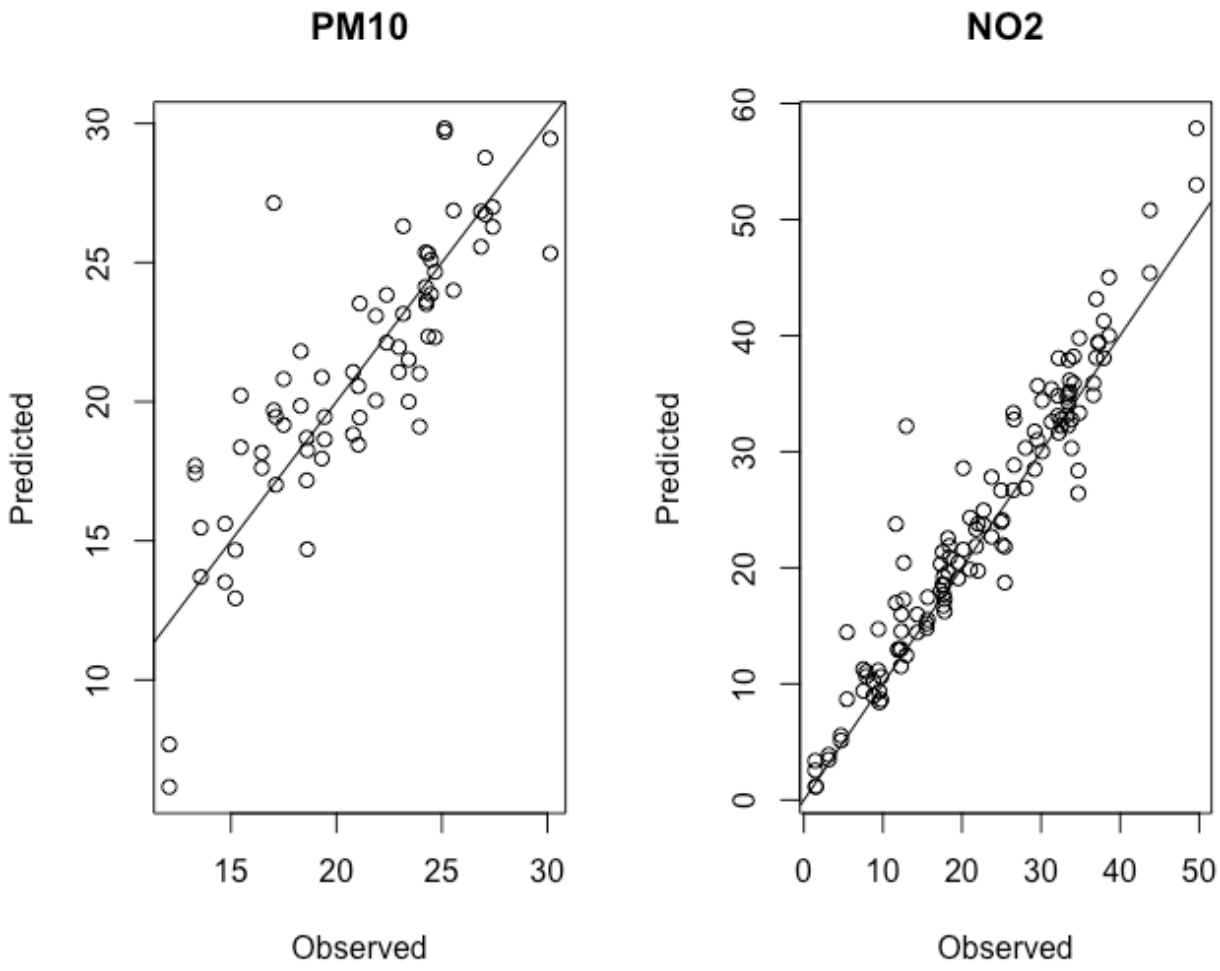
We specified a hierarchical spatiotemporal model, using the Stochastic Partial Differential Equations of the integrated nested Laplace approximations approximation. This approach allowed us to spatially predict, in the territory of Catalonia (Spain), the levels of the four pollutants for which there is the most evidence of an adverse health effect.

Our model allowed us to make fairly accurate spatial predictions of both long-term and short-term exposure to air pollutants, with a low computational cost. The only requirements of the method we propose are the minimum number of stations distributed throughout the territory where the predictions are to be made, and that the spatial and temporal dimensions are either independent or separable.

Key words: Spatial predictions, Hierarchical Bayesian spatiotemporal model, Stochastic Partial Differential Equations (SPDE), Integrated Nested Laplace Approximations (INLA).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Graphical abstract



medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Software and Data availability

We used open data with free access using these sources.

Air pollutants

Departament de Territori i Sostenibilitat, Generalitat de Catalunya [Available at: https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-autom-t/tasf-thgu, last accessed on March 14, 2021].

Meteorological variables

METEOCAT, Generalitat de Catalunya. Meteorological data from XEMA [Available at: https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Dades-meteorol-giques-de-la-XEMA/nzvn-apee, last accessed on March 14, 2021].

AEMET. AEMET Open Data [in Spanish] [Available at: http://www.aemet.es/es/datos_abiertos/AEMET_OpenData, last accessed on March 14, 2021].

Digitized cartography of the ABS

Departament de Salut. Cartography [Available at: https://salutweb.gencat.cat/ca/el_departament/estadistiques_sanitaries/cartografia/, accessed on March 14, 2021].

Code will be available at www.researchprojects.es

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1.- Introduction

In studies assessing the health effects of exposure to air pollution, there is the problem of how to estimate that exposure. Air pollution monitoring station locations do not usually coincide with where the majority of the subjects exposed to such pollution are found. In fact, the air pollution monitoring stations are not often distributed homogeneously in the territory under study, and it is quite usual that large areas, even some densely populated one, do not have any stations at all.

Many studies use the measurements observed in the geographical region of the study to estimate, by means of point estimators, exposure levels for that entire region. The estimators most widely used are the inverse-distance weighted average and the arithmetic mean of the values of the air pollutant observed in several monitor stations, although sometimes the values of the pollutants observed in the nearest monitoring station are also used as estimators. The problem, as Wannemuehler *et al.* (2009) pointed out, is that when air pollution levels exhibit spatial variation across the study region, using these point estimators leads to a bias, as a consequence of ignoring the spatial structure (i.e., spatial dependence) of the data. Furthermore, when that biased estimated level is related to a health variable, this leads to an underestimation of the health effect of interest (Wannemuehler *et al.*, 2009).

There are numerous studies that propose models to estimate the levels of air pollutants, explicitly incorporating both spatial and temporal dependence (Cameletti *et al.*, 2011, 2013; Pirani *et al.*, 2013; Shaddick *et al.*, 2013; Liang *et*

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

al., 2015, 2016; Calculli *et al.*, 2015; Cheam *et al.*, 2017; Mukhopadhyay and Sahu, 2018; Chen *et al.*, 2018; Clifford *et al.,* 2019; Nicolis *et al.,* 2019; Wan *et al.,* 2021) (to refer to only some of those that have appeared in the last ten years). However, we must point out that very few studies attempt to predict air pollution levels in locations where there is no monitoring station (ie, spatial prediction) (Cameletti et al., 2011, 2013; Pirani et al., 2013; Shaddick et al., 2013; Mukhopadhyay and Sahu, 2018; Nicolis et al., 2019), or, having them, to perform out-of-sample temporal predictions (Wan et al., 2021).

The spatial domain of these studies ranges from cities (Santiago de Chile - Nicolis *et al.,* 2019-; Beijing - Wan *et al.,* 2021-) to countries (EU-15 countries - Shaddick *et al.*, 2013-), passing through metropolitan areas (Greater London - Pirani *et al.*, 2013-) and regions (Po valley, northern Italy - Cameletti *et al.*, 2011, 2013-; England and Wales - Mukhopadhyay and Sahu, 2018-). The pollutants that are predicted in these studies are coarse particles, PM_{10}, those with a diameter of 10 micrometres (μm) or less (Cameletti *et al.*, 2011, 2013; Pirani *et al.*, 2013; Mukhopadhyay and Sahu, 2018), fine particles, $PM_{2.5}$, those with a diameter of 2.5 μm or less (Mukhopadhyay and Sahu, 2018; Nicolis *et al.,* 2019; Wan *et al.,* 2021), nitrogen dioxide, NO_2 (Shaddick *et al.*, 2013; Mukhopadhyay and Sahu, 2018) and ozone, O_3 (Mukhopadhyay and Sahu, 2018). Regarding the frequency at which pollutants are observed, daily data (Cameletti *et al.*, 2011, 2013; Pirani *et al.*, 2013; Mukhopadhyay and Sahu, 2018) dominate, although hourly data (Nicolis *et al.,* 2019; Wan *et al.,* 2021) and annual data (Shaddick *et al.*, 2013) are also used.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The models used in most of these articles, in addition to incorporating spatial and temporal dependencies, include explanatory variables among which appear, in decreasing order of the number of studies, meteorological variables (Cameletti *et al.*, 2011, 2013; Pirani *et al.*, 2013; Shaddick *et al.*, 2013; Nicolis *et al.,* 2019; Wan *et al.,* 2021), other pollutants different from the one predicted (Cameletti *et al.*, 2011, 2013), topographical variables (altitude – Cameletti *et al.*, 2013; Wan *et al.,* 2021- and distances to sea and roads - Shaddick *et al.*, 2013 -, and to mountains - Wan *et al.,* 2021-), site types (Pirani *et al.*, 2013; Mukhopadhyay and Sahu, 2018), and land use variables (Shaddick *et al.*, 2013).

With one exception (Wan *et al.,* 2021), the studies use a Bayesian approach since it is the one that best allows the uncertainty of complex space-time data to be incorporated. Most of the studies that use the Bayesian approach perform the inference using the Monte Carlo Markov Chain (MCMC) (Cameletti *et al.*, 2011; Pirani *et al.*, 2013; Shaddick *et al.*, 2013; Mukhopadhyay and Sahu, 2018; Nicolis *et al.,* 2019). Only one uses the Stochastic Partial Differential Equations (SPDE) representation of the INLA approximation (Cameletti *et al.*, 2013). Using MCMC implies a high computational model complexity that, in some cases, prevents the practical application of the methods proposed by these studies. As an exception, it is worth mentioning Nicolis *et al.* (2019), who use the *spTimer* package (Bakar and Sahu, 2015). This package, which uses MCMC, allows large space-time data sets to be handled with fast computation and very good data processing capacity. The INLA approach is much more computationally effective than MCMC, producing accurate approximations to posterior distributions, even for very complex models (Lindgren and Rue, 2015).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

These few studies that provide methods for spatial prediction use a relatively large number of monitoring stations. In this study we intend to present an equally effective model that allows the use of information from a small number of monitoring stations. Furthermore, we intend to make spatial predictions with a computational cost much lower than existing methods.

Specifically, our objective in this work was to present a hierarchical Bayesian spatiotemporal model that allowed us to make spatial predictions of air pollution levels in an effective way and with very few computational costs. In this work, we used the SPDE representation of the INLA approximation to spatially predict, in the territory of Catalonia (Spain), the levels of the four pollutants for which there is the most evidence of an adverse health effect: PM_{10}, NO_2, O_3 and $PM_{2.5}$. We performed the spatial predictions at a point level (defined by its UTM coordinates), allowing them to be aggregated later in any spatial unit required. We were especially interested in the long-term exposure to air pollutants. That is, by living in a certain area an individual is exposed to a mix of pollutants that have lasting effects on their health. We also considered the performance of our method to spatially predict short-term exposure to air pollutants, which has more temporary effects on health.

2.- Methods

2.1.- Data

We obtained information on the hourly levels of air pollution for 2011-2020 from the 143 monitoring stations from the Catalan Network for Pollution Control and

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Prevention (XVPCA) (open data) (*Departament de Territori i Sostenibilitat, Generalitat de Catalunya*, 2021), located throughout Catalonia (Figure 1), and that were active during that period. The pollutants we were interested in for making spatial predictions were PM_{10}, NO_2, O_3 and $PM_{2.5}$ (all of them expressed as $\mu m/m^3$) (air pollutants of interest, hereinafter). However, the monitoring stations also measured other pollutants: nitrogen monoxide (NO), sulphur dioxide (SO_2), carbon monoxide (CO), benzene (C_6H_6), hydrogen sulphide (H_2S), dichloride (Cl_2), and heavy metals (mercury, arsenic, nickel, cadmium and lead). We have used these other pollutants as covariates.

Not all pollutants of interest were measured at all the monitoring stations. Thus, during the entire period 2011-2020, PM_{10} was measured at 122 stations, NO_2 at 77 stations, O_3 at 62 stations and $PM_{2.5}$ at 42 stations. As can be seen in Figure 2, most of the monitoring stations were located in the city of Barcelona and in its metropolitan area. In the rest of the territory, the stations were located in cities (especially those that measure NO_2 and $PM_{2.5}$) and, in the case of O_3, also in rural areas. On the other hand, in 2020 (which we used to spatially predict short-term exposure), the number of air pollution monitoring stations dropped considerably, from 143 to 78. In particular, those stations that measured particles dropped dramatically ($PM_{2.5}$ from 42 to 3, 92.88% less; PM_{10} from 122 to 36, 70.49% less). The number of stations that measured O_3 went from 62 to 50 (19.35% less stations) and NO_2 from 77 to 67 stations (12.99% less) (Table 1).

As we said, our primary interest was in spatially predicting long-term exposure to air pollutants. In this case, we used the monthly averages, after obtaining the

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

daily averages from the hourly data, from January 2011 to December 2019. To make the spatial predictions of the short-term exposure, we used the daily averages from January 1, 2020 to November 29, 2020.

We carried out the spatial predictions at a point level, with the centroids being Basic Health Areas (ABS, for its acronym in Catalan from here on). Catalan health planning defines an ABS as the elementary territorial unit through which primary health care services are organized (*Atenció Primària Girona. Institut Català de la Salut*, 2021). The ABSs are either made up of neighbourhoods or districts in urban areas or by one or more municipalities in rural areas. Their delimitation is determined by geographical, demographic, social and epidemiological factors and, in particular, based on the accessibility the population has to services and the efficiency in the organization of health resources [18]. Catalonia is divided into 376 ABSs, with a population between 371 and 72,321 inhabitants (mean 20,266 inhabitants, standard deviation 13,391, median 18,457 inhabitants, first quartile -Q1- 10,554, third quartile -Q3- 27,529). The population density was in the range of 0.31-34,590.58 inhabitants/km^2 (mean 3,486.36, standard deviation 6,719.23, median 309.18, Q1 44.83, Q3 3,752.54). In Catalonia, 769 of the 947 of the municipalities belong to a single ABS. Of the 178 remaining, 46 were divided into more than one ABS, 37 of them into a maximum of five ABSs, eight between six and 14 ABSs and one, (the city of Barcelona) into 67 ABSs (Idescat, 2021).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Less than a third of ABSs have at least one air pollution monitoring station (105 from a total of 376). An ABS has five monitoring stations, six ABSs have three stations, 22 ABSs have two stations and the remaining 76 have only one station.

As covariates, we included the altitude of the air pollution monitoring station (in m) and the area of the ABS (in km^2). The altitude (as well as other information related to the monitoring station, such as its latitude and longitude) were obtained from the *Departament de Territori i Sostenibilitat* (2021). We transformed the geographic coordinates (latitude and longitude) to UTM coordinates (in km) using the R package *rgdal* (Bivand *et al.*, 2021). The areas of the ABS, as well as the UTM coordinates of their centroids, were calculated using QGIS (version 2.18) from the digitized cartography of the ABS (information of 2018) (open data) (*Departament de Salut*, 2021).

It is known that, at least in the short term, exposure to air pollution is correlated with various meteorological variables. For this reason, in the case of spatial prediction of short-term exposure, we also included several meteorological variables as covariates. Most of them, such as temperature (in ºC), relative humidity (in %), wind speed at 10m (in m/s) and atmospheric pressure (hPA) influence the dispersion of the pollutant; although some also influencing its formation, for instance, global solar radiation (W/m^2) (O_3 is a secondary pollutant, formed when the two atoms that make up oxygen gas dissociate under the action of light solar). The sources of the data were the stations of the Network of Automatic Meteorological Stations (XEMA) of the Meteorological Service of

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Catalonia (METEOCAT) (open data). We also used the daily data from the State Meteorological Agency's (AEMET) automatic stations.

Albeit not as much as the air pollutants monitoring stations, the meteorological stations are also dispersed throughout the territory. Catalonia has 217 meteorological stations, 188 belonging to METEOCAT and 29 to AEMET. All of them measured all the meteorological variables every day. We used the same model (explained in this work) for short-term exposure to carry out the spatial prediction of the daily values of the meteorological variables at the ABS level, for the year 2020. Further details concerning this, can be found in Ribas *et al.* (2021).

2.2.- Model specification

We specified a hierarchical spatiotemporal model as follows:

At the top of the hierarchy:

$$Z(s_i, t) = Y(s_i, t) + \varepsilon(s_i, t) \qquad \{1\}$$

where i denoted the air pollution monitoring station where the pollutant was observed; t was the time unit; s_i was the location of the station; $Y(.,.)$ the spatiotemporal process, the realization of which corresponded to the pollutant measurements (at station i and time unit t); and $\varepsilon(.,.)$ was the measurement error defined by a Gaussian white-noise process (i.e., spatially and temporally uncorrelated) (σ_ϵ^2 was the nugget effect).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

At the next level, we specified the following measurement equation:

$$y(s_i, t) = \mu(s_i, t) + \eta(s_i, t) \qquad \{2\}$$

where $y(.,.)$, is the realization of the spatiotemporal process; $\mu(.,.)$ denoted the large-scale component, depending on the covariates; and $\eta(.,.)$ was a spatiotemporal process.

The spatiotemporal process was an independent in time Gaussian field (GF) with zero mean and a Matérn covariance function:

$$Cov\big(\eta(s_i, t), \eta(s_i', t)\big) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \, (\kappa\|s_i - s_i'\|)^{\nu} \, \mathrm{K}_{\nu} \, (\kappa\|s_i - s_i'\|) \qquad \{3\}$$

where K_{ν} is the modified Bessel function of the second type and order $\nu > 0$. ν is a parameter controlling the smoothness of the GF, σ^2 is the variance and $\kappa > 0$, is a scaling parameter related to the range, ρ, the distance to which the spatial correlation becomes small. We used $\rho = \sqrt{8\,\nu}/\kappa$, where ρ corresponded to the distance where the spatial correlation is close to 0.1 for each ν (Lindgren *et al.*, 2011). $\kappa = 2\phi\sqrt{\nu}$, where ϕ is a parameter controlling the rate of decay of the spatial correlation as the distance $\|s_i - s_i'\|$ increases.

Due to its computational problems, we chose to represent the GF as a Gaussian Markov Random Field (GMRF) (Rue *et al.,* 2009). GMRFs are defined by a precision matrix with a sparse structure allowing inference to be performed in a computationally effective way. We linked the GF and GMRF through the

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Stochastic Partial Differential Equations (SPDE) approach (Lindgren *et al.*, 2011). The SPDE allowed us to find a GMRF, with local neighbourhood and sparse precision matrix (instead of spatiotemporal covariance function and the dense covariance matrix of a GF, respectively), that best represented the Matérn field. Further details can be found in Lindgren *et al.* (2011) and in Cameletti *et al.* (2013).

We specified the large-scale component, $\mu(.,.)$, as a generalized linear mixed model (GLMM) with response from the Gaussian family. Specifically, for each of the pollutants of interest (PM_{10}, NO_2, O_3 and $PM_{2.5}$) we specified two GLMMs: one for long-term exposure and the other for short-term exposure.

Long-term exposure:

$$\mu_{i,t} = \beta_0 + \sum_{j=1}^{14} \beta_j pollutant_{j,it} + \beta_{15} altitude_i + \beta_{16} area_i + sd_y_{i,year} + \eta_i + \tau_{month}$$

Short-term exposure:

$$\begin{aligned}\mu_{i,t} = \beta_0 + \sum_{j=1}^{14} \beta_j pollutant_{j,it} \\ + \beta_{15} temperature_{it} + \beta_{16} relative\ humidity_{it} \\ + \beta_{17} wind\ speed_{it} + \beta_{18} atmospheric\ pressure_{it} \\ + \beta_{19} solar\ radiation_{it} + \beta_{20} altitude_i + \beta_{21} area_i + sd_y_{i,week} + \eta_i \\ + \tau_{day}\end{aligned}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where i denoted the air pollution monitoring station where the pollutant was observed (i=1,2,…143); t was the time unit (month in the case of long-term exposure, day in the case of short-term exposure); $\mu_{i,t} = E(y_{it})$, y_{it} denoted the air pollutants of interest, PM_{10}, NO_2, O_3 and $PM_{2.5}$; $pollutant_{j,it}$ corresponded to the pollutant j measurements at station i and time unit t. Pollutants considered were, first, the pollutants of interest other than the pollutant for which the spatial prediction was made and, second, the rest of the pollutants (i.e., NO, SO_2, CO, C_6H_6, H_2S, Cl_2, mercury, arsenic, nickel, cadmium and lead); $area_i$ was the area of the ABS i; $sd_y_{i,.}$, η_i and $\tau_.$ denoted random effects.

In the models, we included $sd_y_{i,year}$, $sd_y_{i,week}$ structured random effects, indexed on a standard deviation of the air pollutant that was being predicted, in the ABS i, during a particular year (2011 to 2018) and a particular week of 2020 (weeks 1 to 37), respectively. We chose, a random walk of order one (rw1) as the structure of the random effect. In the integrated nested Laplace approximations (INLA) approach (Rue *et al.*, 2009, 2017), the random walk of order 1 for the Gaussian vector x is constructed assuming independent increments (R INLA project, 2021a):

$$\Delta x_i = x_i - x_{i-1} \sim N(0, \sigma_x^2)$$

Following the INLA approach, when, as in our case, the random effects are indexed on a continuous variable, they can be used as smoothers to model non-linear dependency on covariates in the linear predictor.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

η_i denoted a random effect indexed on the air pollution monitoring station. This random effect was unstructured (independent and identically distributed random effects) and captured individual heterogeneity, that is to say, unobserved confounders specific to the station and invariant in time.

We also included τ_{month} and τ_{day}, structured random effects indexed on time, in order to control the temporal dependency associated to possible seasonal effects throughout the year (long-term exposure) and throughout the week (short-term exposure). In this case, a model for seasonal variation with periodicity *m* (12 for long-term exposure, seven for short-term exposure), for the random vector (x_1, x_2,…,x_n) (n>m) was obtained assuming that the sums were independent Gaussian with a precision τ. The density for x is derived from the n-m+1 increments (R INLA project, 2021b):

$$\tau^{\frac{n-m+1}{2}} e^{-\frac{\tau}{2}\sum(x_i+x_{i+1}+\ldots+x_{i+m-1})^2}$$

2.3.- Inference

Inferences for GMRFs were made following a Bayesian perspective, using the INLA approach (Rue *et al.*, 2009, 2017).

We started from the SPDE representation, which uses a finite element representation to define the Matérn field as a linear combination of basis functions defined on a triangulation of the domain (mesh, hereinafter). This consists of subdividing the domain into a set of non-intersecting triangles meeting in, at most, a common edge or corner (Lindgren *et al.,* 2011; Cameletti *et al.,* 2013).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Then, instead of projecting the subsequent mean of the random field onto mesh nodes to target locations where we do not have observed data, we performed the spatial prediction of the random field jointly with the parameter estimation process. For this, we projected the mesh into those locations with no air pollutants observed and then we jointly computed the posterior means at all the locations (with observed and unobserved air pollutants measurements) (Krainski *et al.,* 2020).

We separately estimated each year (long-term exposure) and each week (short-term exposure) and then merged every year and every week.

We used priors that penalize complexity (called PC priors). These priors are robust in the sense that they do not have an impact on the results, and furthermore, they have an epidemiological interpretation (Simpson *et al.,* 2017).

All analyses were carried out using the free software R (version 4.0.3), through the INLA package (Rue *et al.*, 2009, 2017; R INLA project, 2021c). The maps were represented using the *leaflet* package (Cheng *et al.,* 2019).

2.4.- Measures of predictive performance

The predictive performance of each model was assessed by cross-validation, considering a training set (2011 to 2018 for long-term exposure, weeks 1 to 36 - January 1 to September 8, 2020 -, for short-term exposure) and a test set (2019

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

for long-term exposure and weeks 37 - September 9 - to 48 - November 29, 2020- for short-term exposure).

The prediction accuracy was assessed by:

- Mean absolute percentage error (MAPE)

$$MAPE = \frac{1}{N}\sum_{i}\sum_{t}\left|\frac{y(s_i,t) - \hat{y}(s_i,t)}{\hat{y}(s_i,t)}\right| * 100$$

where N was the total number of available observations in the test set; $y(s_i,t)$ were the pollutant measurements (at station i and time unit t) at the test set; and $\hat{y}(s_i,t)$ were the posterior means.

- Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}\sum_{t}\left(y(s_i,t) - \hat{y}(s_i,t)\right)^2}$$

- Correlation coefficient

$$r = \frac{\sum_i \sum_t \left(y(s_i,t) - \overline{y(s_\iota,t)}\right)\left(\hat{y}(s_i,t) - \overline{\hat{y}(s_\iota,t)}\right)}{\left(\sum_i \sum_t \left(y(s_i,t) - \overline{y(s_\iota,t)}\right)^2 \sum_i \sum_t \left(\hat{y}(s_i,t) - \overline{\hat{y}(s_\iota,t)}\right)^2\right)^{\frac{1}{2}}}$$

- Actual coverage of the 95% prediction intervals

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2.5.- Sensitivity analysis

We conducted two sensitivity analyses. In the first place, we carried out a new cross-validation and, in the second place, we changed the spatiotemporal model. In both cases, we consider the spatial prediction of long-term exposure to NO_2.

As regards cross-validation, we considered, as training sets, five random samples from the monitoring stations in which NO_2 was measured during the entire period 2011-2019. Specifically, we considered random samples of, approximately, 75% of the stations (58 out of a total of 77 stations), of 70% (55 stations), of 50% (41 stations), of 45% (35 stations), and of 20% (18 stations). As a test set, we considered the rest of the stations (19, 22, 36, 42 and 59 remaining stations, respectively).

Next, we calculated the measures' prediction accuracy (explained previously).

With respect to the spatiotemporal model above, we considered an independent in time Gaussian field (GF), following Camelleti *et al.* (2013) we assumed a spatiotemporal Gaussian field that changes in time according to an autoregressive of order one (AR(1)).

Returning the measurement equation {2}:

$$y(s_i, t) = \mu(s_i, t) + \eta(s_i, t) \qquad \{2\}$$

the realization of the spatiotemporal process, $\eta(.,.)$, was specified as,

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

$$\eta(s_i, t) = \phi\eta(s_i, t-1) + \omega(s_i, t) \qquad \{4\}$$

where $|\phi| < 1$.

Here, it was $\omega(s_i, t)$ what was assumed to be zero mean Gaussian and a Matérn covariance function:

$$Cov\big(\omega(s_i, t), \omega(s_i', t)\big) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\kappa\|s_i - s_i'\|\right)^{\nu} \mathrm{K}_{\nu} \left(\kappa\|s_i - s_i'\|\right) \qquad \{5\}$$

In the addition, in the GLMM specification of the large-scale component, $\mu(.,.)$, in the linear predictor we included structured random effects indexed on year, τ_{year}, in order to capture the long-term trend.

$$\mu_{i,t} = \beta_0 + \sum_{j=1}^{14} \beta_j pollutant_{j,it} + \beta_{15} altitude_i + \beta_{16} area_i + sd_y_{i,year} + \eta_i + \tau_{month} + \tau_{year}$$

With this analysis, our objective was to compare not only the predictive performance of the model {1-2}, {4-5} with the one specified above {1-3}, but, above all, to compare the computation time in the inference of both specifications.

3.- Results

Descriptive results are shown in Table 1. Regarding long-term exposure, we observed that, with the exception of O_3, the daily averages of pollutants decreased in 2019 ($PM_{2.5}$ 22.39% less, NO_2 12.88% less and PM_{10} 8.48% less).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

In contrast, the daily average of O_3 increased by 3.97% in 2019 compared to 2011-2018. With regard to short-term exposure, the levels of NO_2 and of the PM_{10} were higher from September 9 (week 37) (23.11% and 4.87%, respectively). Conversely, the levels of O_3 and of $PM_{2.5}$ (although in this case only measured in three stations) were lower than the levels before September 9 (19.34% and 9.81%). When we excluded the lockdown (which took place in Spain from March 14 - week 11- to June 21 - week 25 -, both 2020), the variation from September 9 changed sign for PM_{10}, it was 7.67% lower, they were moderated for NO_2 (which was 3.64% higher) and O_3 (12.48% lower), while they were increased in the case of $PM_{2.5}$ (13.74% lower).

The measures of predictive performance are shown in Table 2. With the exception of $PM_{2.5,}$ the results for long-term exposure were quite good. Achieved coverages of the 95% credibility intervals for predictions were greater than 90%, correlation coefficients were greater than 0.80, and MAPEs less than 10%. Furthermore, if Table 2 is compared with Table 1, it is observed that the reduction in the variability of the spatial prediction, measured between the ratio of the RMSE and the standard deviations of the pollutants observed, was, at most, one third of the standard deviations of the pollutants during the period 2011-2018 (19.40% for NO_2, 25.71% for O_3 and 33.89% for PM_{10}), again with the exception of $PM_{2.5}$ (the RMSE in this case was 59.92% of the standard deviation in the period 2011-2018). Therefore, except for $PM_{2.5}$, our method managed to significantly reduce the variability of the spatial prediction around fairly accurate predictions. Although quite good, note that, in relative terms, the results for PM_{10} were somewhat worse than for gaseous pollutants (NO_2 and O_3).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The poor results obtained for $PM_{2.5}$ are because of its smaller sample size. Although it is true that in the period as a whole up to 42 stations measured $PM_{2.5}$, the year with the lowest number of active stations was 2018 (31 stations), with the rest of the years ranging between 33 and 35 active stations. No year fell below 40 active stations for the rest of pollutants (the year with the lowest number of stations measuring PM_{10} was 2018 with 94 stations, while the other years ranged between 100 and 107 stations; in the cases of NO_2 and O_3 it was 2015 with 59 and 44 stations, respectively, with the other years oscillating between 62 and 66, and 45 and 57, respectively).

Regarding the short-term exposure, first, predictive performance was worse when we did not exclude the lockdown period (which took place in Spain from March 14 to June 21, 2020) than when we did. In fact, note that predictive performance measures were much better for gaseous pollutants (NO_2 and O_3). The results for the coarse particles, $PM_{10,}$ were quite poor (we did not interpret the results for $PM_{2.5}$ as it was measured in only three stations). This was likely due to the lower number of stations where PM_{10} was measured (36 stations, versus 67 for NO_2 and 50 for O_3, see Table 1). The variability of the spatial prediction was reduced much less than in long-term exposure, especially for NO_2. The RMSEs were between 33.83% for O_3 and 43.49% for NO_2, of the standard deviations of the pollutants (excluding lockdown).

The results of the sensitivity analyses, when the number of stations in the training set was greater than 40 and when the spatiotemporal Gaussian field changed in

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

time according to an AR(1) (model {1-2}, {4-5}), were quite similar to the results for the spatiotemporal process independent in time Gaussian field (model {1-3}) and all the stations were included in the training set (Table 3).

When we varied the number of stations in the training set, but used every year (2011 to 2019), the predictive performance seemed to depend on the size of the sample. The more stations the training set had, the better the results, while dramatically deteriorating with a small sample size. In particular, the cut-off appears to be 40 stations. Below this, the predictive performance measures were poor.

Although the predictive performance of the spatiotemporal Gaussian field model changed in time according to an AR (1) (model {1-2}, {4-5}) was very similar to that of the spatiotemporal process independent in time Gaussian field (model {1-3}) (perhaps somewhat worse, in relative terms), the computation time was much longer. Using a 6-core Intel Core i9 (2.9 GHz 32 GB RAM), while the model inference {1-3} required on average 0.05 seconds per observation (a total of 569 seconds on average), the model {1-2}, {4-5} required 0.354 seconds (a total of 3,947 seconds), that is, seven times more computing time.

The maps of the posterior means and the posterior standard deviations for 2019 (in quintiles) of the spatiotemporal process independent in time Gaussian field (model {1-3}) for the long-term exposure of PM_{10}, NO_2 and O_3 are shown in Figures 3. We decided not to represent the posterior means for $PM_{2.5}$ because of its poor predictive performance. The spatial distributions of the subsequent

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

means of PM_{10} and NO_2 were quite similar, although in the high levels of NO_2, (fourth and fifth quintiles) there was somewhat more spatial variation. Note that, unlike PM_{10} and NO_2, the lowest levels of O_3 (first and second quintiles) occurred in the urban areas. As expected, the uncertainty, as measured by the posterior standard deviations, was, in general, higher in those areas with few (or no) monitoring stations. Note, however, that higher levels of air pollutants do not always coincide with higher standard deviations.

4.- Discussion

Our results were quite good in terms of predictive performance, at least for those pollutants that were observed in more than 40 collecting stations (PM_{10}, NO_2 and O_3 in long-term exposure and NO_2 and O_3 in short-term exposure).

The current coverage of the spatial predictions of these pollutants are in line with similar studies. Using the same model and the same data (PM_{10}), but using two different methods for the inference, Camelleti *et al.* find coverage between 0.95 and 0.97 (using MCMC) (Camelleti *et al.,* 2011) and 0.897 (using INLA SPDE) (Camelleti *et al.,* 2013). Mukhopadhyay and Sahu (2018) find coverage between 0.91 and 0.92 for the spatial predictions for O_3 (in our case, 0.89 for short-term exposure and 0.945 for long-term exposure), between 0.89 and 0.90 for PM_{10} (in our case, 0.917 for long-term exposure) and between 0.95 and 0.965 for NO_2 (in our case, 0.905 for short-term exposure and 0.963 for long-term exposure). Note: we have preferred not to comment on the results in which we found poor predictive performance. Our coverage could also be comparable to those

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

provided by Pirani *et al*. (2013) for the spatial predictions for PM_{10} (between 0.87 and 0.93), although it should be noted that these show the coverage at 90%.

The correlation coefficients between the observed levels of air pollutants and the subsequent means of the spatial predictions were higher in our case than in Camelleti *et al.* (0.863 when the inferences were made with MCMC - Camelleti *et al.,* 2011- and 0.702 when they were made with INLA SPDE - Camelleti *et al.,* 2013-, compared to 0.917 in our case), and than in Pirani *et al.* (2013) (between 0.73 and 0.78), in both cases for PM_{10}. However, they were somewhat lower than in Mukhopadhyay and Sahu (2018) (0.88-0.89 for PM_{10}, 0.92-0.94 for NO_2, and 0.93-0.94 for O_3). It should be said, nonetheless, that the number of observations in Mukhopadhyay and Sahu range between 56,625 (for PM_{10}) and 100,138 (for NO_2), while in our case we had 11,157 observations.

The reduction in the variability of the spatial prediction, can only be compared with Mukhopadhyay and Sahu (2018), since they are the only ones who show these standard deviations. In this sense, both Mukhopadhyay and Sahu and ourselves achieved a similar reduction in the variability of the spatial prediction.

Although good, the results of the predictive performance were less so for the spatial prediction of long-term exposure to PM_{10} (although it was being observed in the largest number of collecting stations, see Table 1) and for the short-term exposure for gaseous pollutants (NO_2 and O_3).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Regarding the spatial prediction of long-term exposure to PM_{10}, we believe that it is a consequence of the location of the monitoring stations. The stations that measure PM_{10}, although more abundant in urban areas, are also located in rural areas, while those that measure NO_2 are located almost exclusively in urban areas. In the city of Barcelona, while 13% of NO_2 is generated outside the municipality, it is 71% in the case of PM_{10} (Barcelona City Council, 2021; Saez *et al.,* 2020). It is not unreasonable to suppose that these figures can be extrapolated to the entire Barcelona Metropolitan Area, which comprises 41.75% of the total population of Catalonia and where the majority of PM_{10} and NO_2 monitoring stations are located. In other words, while NO_2 monitoring stations measured almost all NO_2 pollution, PM_{10} monitoring stations did not collect all PM_{10} pollution data. This could also explain why the posterior means of the PM_{10} predictions exhibited less spatial variability than the NO_2 predictions (Figures 3a and 3b).

With regard to the spatial predictions of short-term exposure, the reduction in the number of monitoring stations during 2020 could have led to a deterioration in the predictive performance. However, we believe it could also be due to the data behaviour during 2020. As a consequence of the lockdown to flatten the COVID-19 pandemic curve, mobility was greatly reduced in 2020. Specifically, mobility was reduced by 40% on average, compared to pre-COVID-19 levels, during the lockdown and did not fully recover in the September-November 2020 period (being 5 to 15% lower, depending on the area of Catalonia [26]). We are sure that this anomalous behaviour would have influenced the predictive performance of the spatial predictions of short-term exposure.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The predictive performance of our model depends on the number of stations where pollutants are measured. We have found that with less than 40 stations, probably spread throughout the territory (although not necessarily homogeneously), the predictive performance deteriorates considerably.

Our method is quite similar to that of Camelleti et al. (2013). However, as we show with the sensitivity analysis, our method, in which we perform the inference year by year (or week by week) and then merge the subsequent ones, has a much shorter computation time, in addition to somewhat better results, even for such an atypical year as 2020.

Nevertheless, we are convinced that our results might not be as good if the spatial and temporal dimensions were dependent and not separable, that is, if the spatial dependence varied over time. Fortunately, the spatial dependence of air pollutants does not vary over time. Even during 2020, although air pollution levels decreased as a consequence of the reduction in mobility, the spatial dependence was more or less similar to previous years. For spatial dependence to vary over time, major changes in infrastructures or, likewise, lasting limitations in mobility that were not homogeneous throughout the territory, have to be produced. Of course, other types of spatiotemporal data could imply other results.

5.- Conclusion

In this work, we have shown a hierarchical Bayesian spatiotemporal model that has allowed us to make fairly accurate spatial predictions with a low

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

computational cost. Our model provides predictions of both long-term and short-term exposure. The only requirements of the method that we propose lie in a minimum number of stations being distributed throughout the territory where the prediction is to be made and that the spatial and temporal dimensions are either independent or separable.

Funding

This work was partially financed by the SUPERA COVID19 Fund, from SAUN: Santander Universidades, CRUE and CSIC, and by the COVID-19 Competitive Grant Program from Pfizer Global Medical Grants. It also received funding, in the form of a free transfer of data, from the AEMET. The funding sources did not participate in the design or conduct of the study, the collection, management, analysis, or interpretation of the data, or the preparation, review, or approval of the manuscript.

Acknowledgements

This study was carried out within the 'Cohort-Real World Data' subprogram of CIBER of Epidemiology and Public Health (CIBERESP).

Ethics

Not applicable.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

AEMET. AEMET Open Data [in Spanish] [Available at: http://www.aemet.es/es/datos_abiertos/AEMET_OpenData, last accessed on February 27, 2021].

Atenció Primària Girona. Institut Català de la Salut. Basic Health Areas (ABS) [In Catalan] [Available at: http://www.icsgirona.cat/ca/contingut/primaria/370, last accessed on February 27, 2021].

Bakar KS, Sahu SK. spTimer: Spatio-temporal Bayesian modelling using R. *Journal of Statistical Software* 2015; 63(15):1–32. doi: 10.18637/jss.v063.i15.

Barcelona City Council. Barcelona Air Quality Improvement Plan (2015-2018) [in Catalan] [Available at: https://ajuntament.barcelona.cat/ecologiaurbana/ca/que-fem-i-per-que/ciutat-productiva-i-resilient/pla-de-qualitat-de-l-aire-de-bcn, last accessed on March 14, 2021].

Bivand R, Keitt T, Rowlingson B. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.5-8 [Available at https://CRAN.R-project.org/package=rgdal, last accessed on February 27, 2021].

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Calculli C, Fassò A, Finazzi F, Pollice A, Turnone A. Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. *Environmetrics.* 2015; 26(6):406-417. doi: 10.1002/env.2345.

Cameletti M, Ignaccolo R, Bande S. Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics.* 2011; 22(8): 985–996. doi: 10.1002/env.1139.

Cameletti M, Lindgren F, Simpson D, Rue H. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Adv Stat Anal.* 2013; 97(2):109–131. doi: 10.1007/s10182-012-0196-3.

Cheam ASM, Marbac M, McNicholas PD. Model-based clustering for spatiotemporal data on air quality monitoring. *Environmetrics.* 2017; 28(3):e2437. doi: 10.1002/env.2437.

Chen L, Guo B, Huang J, He J, Wang H, Zhang S, Chen SX, Assessing air-quality in Beijing-Tianjin-Hebei region: The method and mixed tales of $PM_{2.5}$ and O_3. *Atmos Environ.* 2018; 193:290-301. doi: 10.1016/j.atmosenv.2018.08.047.

Cheng J, Karambelkar B, Xie Y. leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.3. https://CRAN.R-project.org/package=leaflet, 2019.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Clifford S, Low-Choy S, Mazaheri M, Salimi F. A Bayesian spatiotemporal model of panel design data: Airborne particle number concentration in Brisbane, Australia. *Environmetrics.* 2019; 30(7):e2597. doi: 10.1002/env.2597.

Departament de Salut. Cartography [Available at: https://salutweb.gencat.cat/ca/el_departament/estadistiques_sanitaries/cartografia/, accessed on February 27, 2021].

Departament de Territori i Sostenibilitat, Generalitat de Catalunya [Available at: https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-autom-t/tasf-thgu, last accessed on February 27, 2021].

Idescat. Statistical Institute of Catalonia [Available at: https://www.idescat.cat/?lang=en, last accessed on February 27, 2021].

Krainski E, Gómez-Rubio V, Bakka H, Lenzi A, Castro-Camilo D, Simpson D, Lindgren F, Rue H. Advanced Spatial Modelling with Stochastic Partial Differential Equations Using R and INLA. London: Chapman and Hall/CRC, 2020, chapter 2.5 [Available at: https://becarioprecario.bitbucket.io/spde-gitbook/, last accessed on March 6, 2021].

Liang X, Zou T, Guo B, Li S, Zhang H, Zhang S, Huang H, Chen SX. Assessing Beijing's $PM_{2.5}$ pollution: severity, weather impact, APEC and winter heating. *Proc R Soc A.* 2015; 471(2182):20150257. doi: 10.1098/rspa.2015.0257.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Liang X, Li S, Zhang S, Huang H, Chen SX. $PM_{2.5}$ data reliability, consistency, and air quality assessment in five Chinese cities. *JCR Atmospheres* 2016; 121(17):10220-10236. doi: 10.1002/2016JC024877.

Lindgren FK, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J R Stat Soc Series B Stat Methodol*. 2011; 73(4):423-498. doi: j.1467-9868.2011.00777.x.

Lindgren F, Rue H. Bayesian spatial modelling with R-INLA. *Journal of Statistical Software* 2015; 63(19). doi:10.18637/jss.v063.i19.

METEOCAT, Generalitat de Catalunya. Meteorological data from XEMA [Available at: https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Dades-meteorol-giques-de-la-XEMA/nzvn-apee, last accessed on February 27, 2021].

Mukhopadhyay S, Sahu SK. A Bayesian spatiotemporal model to estimate long-term exposure to outdoor air pollution at coarser administrative geographies in England and Wales. *J R Stat Soc Ser A-Stat Soc.* 2018; 181(2): 465–486. doi: 10.1111/rssa.12299.

Nicolis O, Díaz M, Sahu SK, Marín JC. Bayesian spatiotemporal modeling for estimating short-term exposure to air pollution in Santiago de Chile. *Environmetrics.* 2019; 30(7):e2574. doi: 10.1002/env.2574.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Pirani M, Gulliver J, Fuller GW, Blangiardo M. Bayesian spatiotemporal modelling for the assessment of short-term exposure to particle pollution in urban areas. *Journal of Exposure Science & Environmental Epidemiology* 2013; 27:1-9. doi: 10.1038/jes.2013.85.

QGIS. Version 2.18 [Available at: https://qgis.org/downloads/, last accessed on February 27, 2021].

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/, 2021.

R INLA project 2021a. Random walk of order 1 (RW1) [Available at: https://inla.r-inla-download.org/r-inla.org/doc/latent/rw1.pdf, last accessed on March 12, 2021].

R INLA project 2021b. Model for seasonal variation [Available at: https://inla.r-inla-download.org/r-inla.org/doc/latent/seasonal.pdf, last accessed on March 12, 2021].

R INLA project, 2021c [Available at: http://www.r-inla.org/home, last accessed on March 12, 2021].

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Ribas V, Miralles F, Rey O, Rafael X, Subías P, Torrent M, Vicens JA, Saez M, Barceló MA, Ponce-de-León M, Valencia A, Arenas A, Saura P. *Big Data i Intel·ligència Artificial per a la prevenció d'epidèmies. Monitoratge i predicció per a la detecció primerenca de brots epidemics* [in Catalan] Barcelona: Generalitat de Catalunya, 2021.

Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J R Stat Soc Series B Stat Methodol*. 2009; 71:319-392. doi:j.1467-9868.2008.00700.x.

Rue H, Riebler A, Sørbye H, Illian JB, Simpson DP, Lindgren FK. Bayesian computing with INLA: A review. *Annual Reviews of Statistics and its Applications* 2017; 4(March):395-421. doi: annurev-statistics-060116-054045.

Saez M, Tobias A, Barceló MA. Effects of long-term exposure to air pollutants on the spatial spread of COVID-19 in Catalonia, Spain. *Environ Res.* 2020; 191:110177. doi: 10.1016/j.envres.2020.110177.

Shaddick G, Yan H, Vienneau D. A Bayesian hierarchical model for assessing the impact of human activity on nitrogen dioxide concentrations in Europe. *Environ Ecol Stat.* 2013; 20:553–570. doi: 10.1007/s10651-012-0234-z.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Simpson DP, Rue H, Martins TG, Riebler A, Sørbye SH. Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *Statistical Science* 2017; 32(1):1-46. doi: 10.1214/16-STS576.

Wan Y, Xu M, Huang H, Chen SX. A spatio-temporal model for the analysis and prediction of fine particulate matter concentration in Beijing. *Environmetrics* 2021; 32(1):e2648. doi: 10.1102/env.2648.

Wannemuehler K, Lyles R, Waller L, Hoekstra R, Klein M, Tolbert P. A conditional expectation approach for associating ambient air pollutant exposures with health outcomes. *Environmetrics.* 2009; 20(7):877-894. doi: 10.1002/env.978.

Table 1.- Description of the air pollutants.

	Long-term exposure			Short-term exposure			
	Number stations	**2011-2018**	**2019**	**Number stations**	**Week 1-36, 2020**	**Week 1-10 and 26-36 (excluding lockdown), 2020**	**Week 37-48, 2020**
PM_{10} (µm/m^3)	122			36			
mean (sd)		23.00 (7.79)	21.05 (6.37)		19.29 (9.85)	21.91 (10.96)	20.23 (11.34)
median [Q1-Q3]		22.32 [18.10-27.15]	20.68 [16.67-25.37]		17.83 [13.13-23.42]	20.29 [15.30-26.21]	18.08 [12.71-25.42]
NO_2 (µm/m^3)	77			67			
mean (sd)		25.00 (15.10)	21.78 (12.76)		15.49 (12.04)	18.40 (13.36)	19.07 (12.98)
median [Q1-Q3]		23.59 [12.87-35.61]	20.92 [11.59-30-61]		12.25 [6.54-21.46]	15.21 [7.88-26.46]	17.11 [8.71-28.13]
O_3 (µm/m^3)	62			50			
mean (sd)		53.64 (20.56)	55.77 (20.22)		55.23 (19.76)	50.90 (21.26)	44,55 (19.40)
median [Q1-Q3]		55.33 [37.62-68.86]	57.47 [39.60-70.15]		56.67 [43.92-68.00]	53.27 [36.58-64.79]	43.92 [30.25-58.42]
$PM_{2.5}$ (µm/m^3)	42			3			
mean (sd)		13.98 (5.12)	10.85 (2.91)		11.42 (7.54)	11.94 (8.44)	10.30 (6.58)
median [Q1-Q3]		13.33 [10.47-16.75]	10.48 [9.86-11.57		9.33 [5.86-14.51]	9.33 [5.66-15.34]	9.13 [5.45-14.02]

Number of stations	143	78

Daily averages

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 2.- Measures of predictive performance. Spatiotemporal process independent in time Gaussian field

Long-term exposure

	MAPE	RMSE	Correlation	Coverage
PM_{10}	7.429%	2.640	0.838	0.917
NO_2	4.345%	2.930	0.937	0.973
O_3	6.795%	5.287	0.916	0.945
$PM_{2.5}$	16.037%	3.068	0.696	0.750

Short-term exposure – Training set Weeks 1-37

	MAPE	RMSE	Correlation	Coverage
PM_{10}	16.373%	6,582	0.396	0.445
NO_2	17.274%	8.020	0.796	0.849
O_3	20.604%	9.711	0.726	0.837
$PM_{2.5}$	49.897%	6.366	0.398	0.081

Short-term exposure – Training set Weeks 1-10 and 26-36 (excluding lockdown)

	MAPE	RMSE	Correlation	Coverage
PM_{10}	14.412%	6.091	0.462	0.493
NO_2	7.521%	5.811	0.894	0.905
O_3	9.233%	7.192	0.853	0.890
$PM_{2.5}$	39.936%	5.589	0.398	0.081

MAPE: Mean absolute percentage error

RMSE: Root mean square error

Correlation: Correlation coefficient

Coverage: Actual coverage of the 95% prediction credibility intervals

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 3.- Sensitivity analyses. Measures of predictive performance

NO_2. Long-term exposure 2011-2019.

Total of active monitoring stations in the period: 77

Spatiotemporal process independent in time Gaussian field

	MAPE	RMSE	Correlation	Coverage
Training set 58 monitoring stations (75%)	4.973%	3.205	0.930	0.963
Training set 55 monitoring stations (70%)	5.902%	6.464	0.851	0.925
Training set 41 monitoring stations (50%)	6.072%	7.201	0.786	0.870
Training set 35 monitoring stations (45%)	8.020%	8.088	0.776	0.633
Training set 18 monitoring stations (20%)	8.806%	10.159	0.609	0.590

MAPE: Mean absolute percentage error

RMSE: Root mean square error

Correlation: Correlation coefficient

Coverage: Actual coverage of the 95% prediction credibility intervals

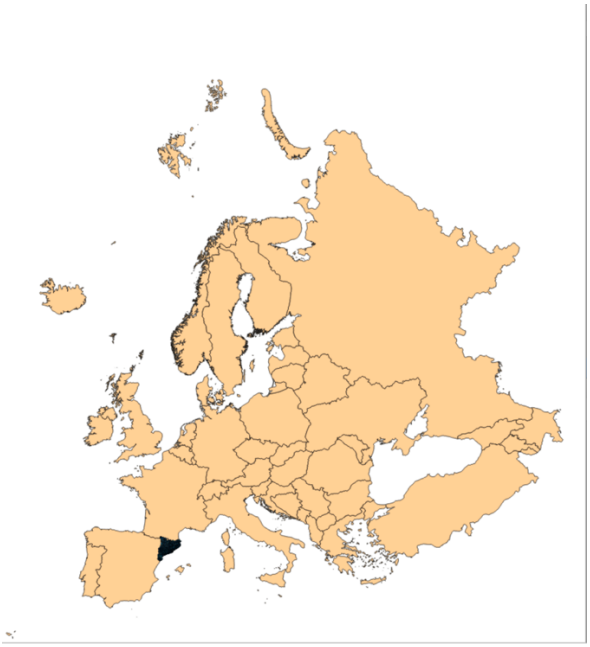
Spatiotemporal Gaussian field that changed in time according to an autoregressive of order one (AR(1))

Long-term exposure

	MAPE	RMSE	Correlation	Coverage
NO_2	8.843%	3.388	0.957	0.958

medRxiv preprint doi: https://doi.org/10.1101/2021.06.06.21258419; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 1.- Location of Catalonia (Spain).



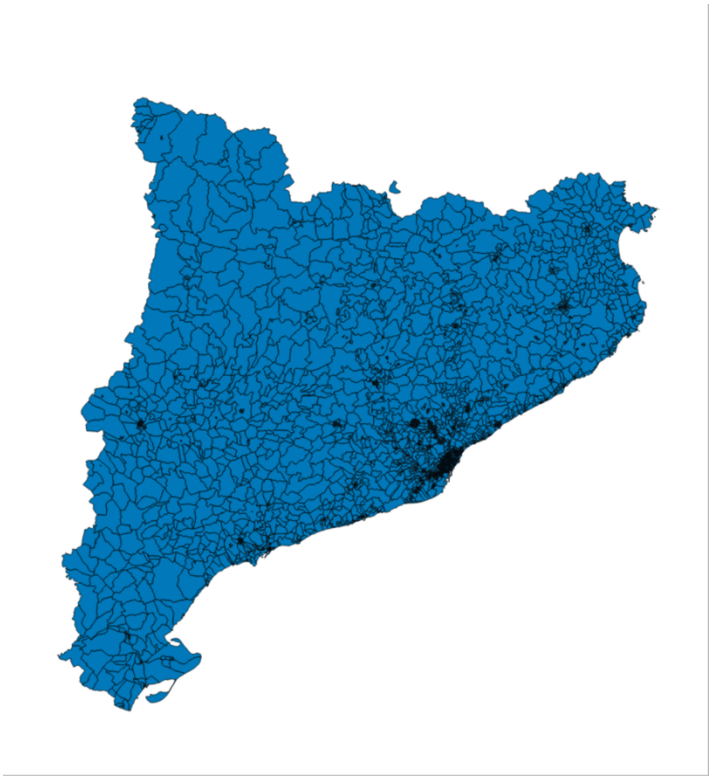


Figure 2.- Distribution in the territory of Catalonia (Spain), of the air pollution monitoring stations, according to where air pollutants (PM_{10}, NO_2, O_3 and $PM_{2.5}$) are measured.

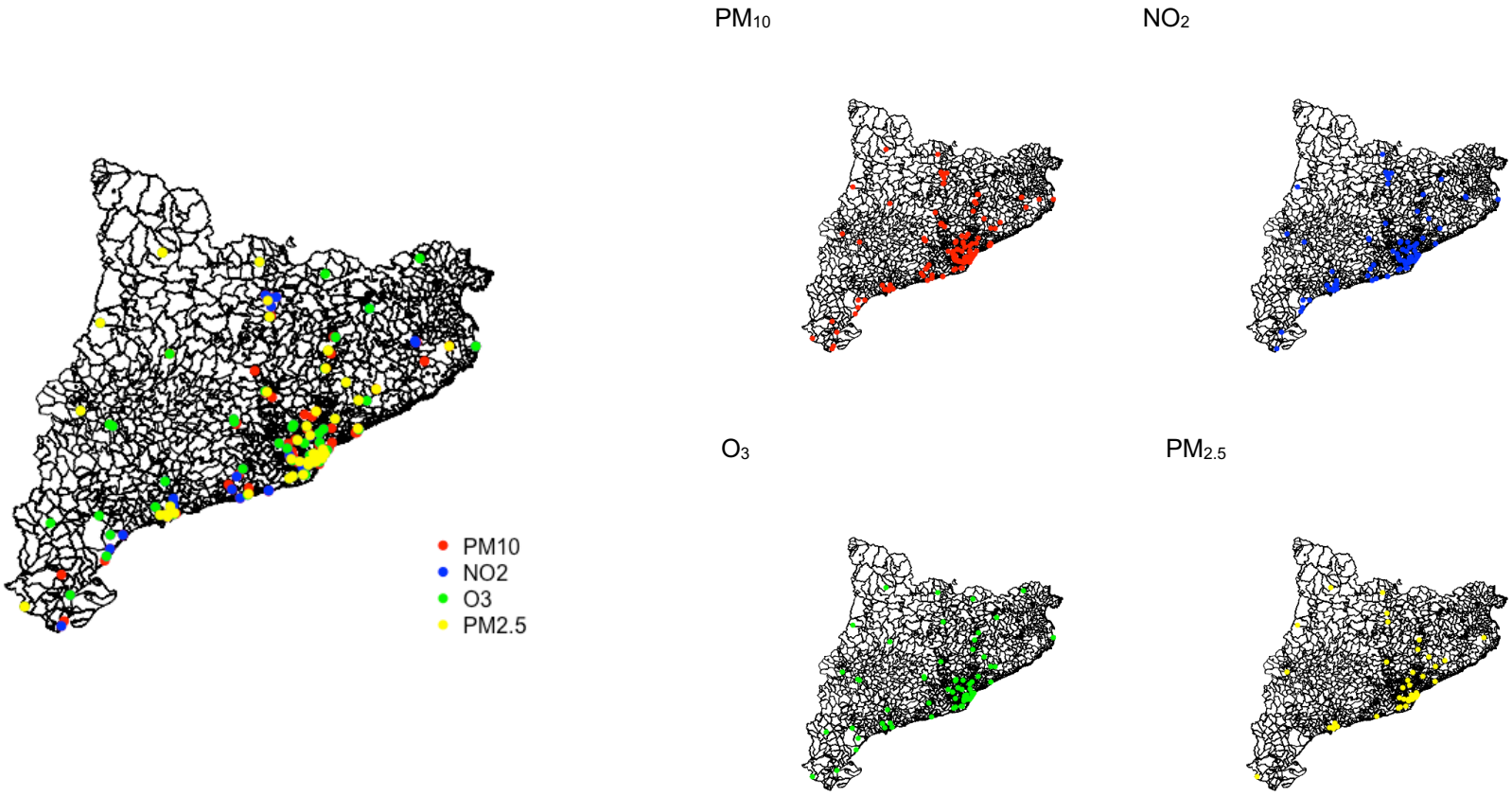


Figure 3a.- Posterior mean and posterior standard deviation of PM_{10} for 2019. Spatiotemporal process independent in time Gaussian field

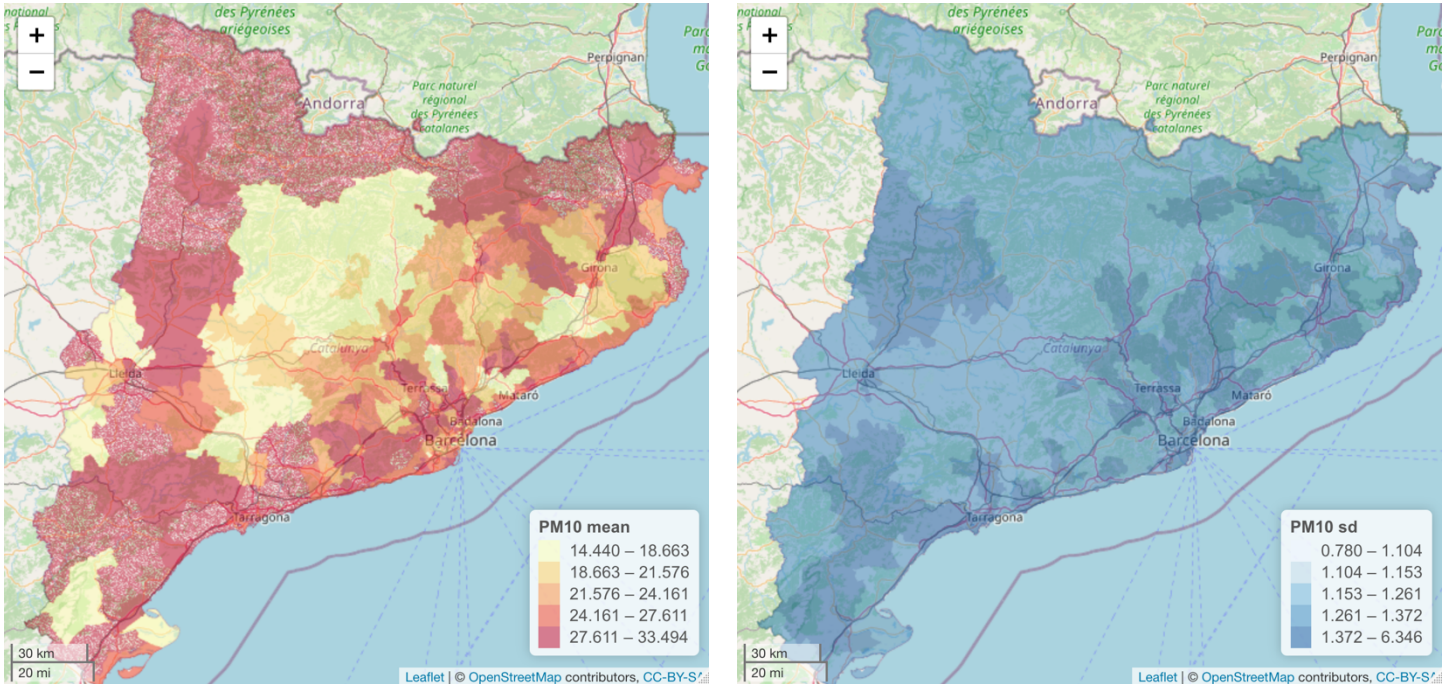


Figure 3b.- Posterior mean and posterior standard deviation of NO_2 for 2019. Spatiotemporal process independent in time Gaussian field

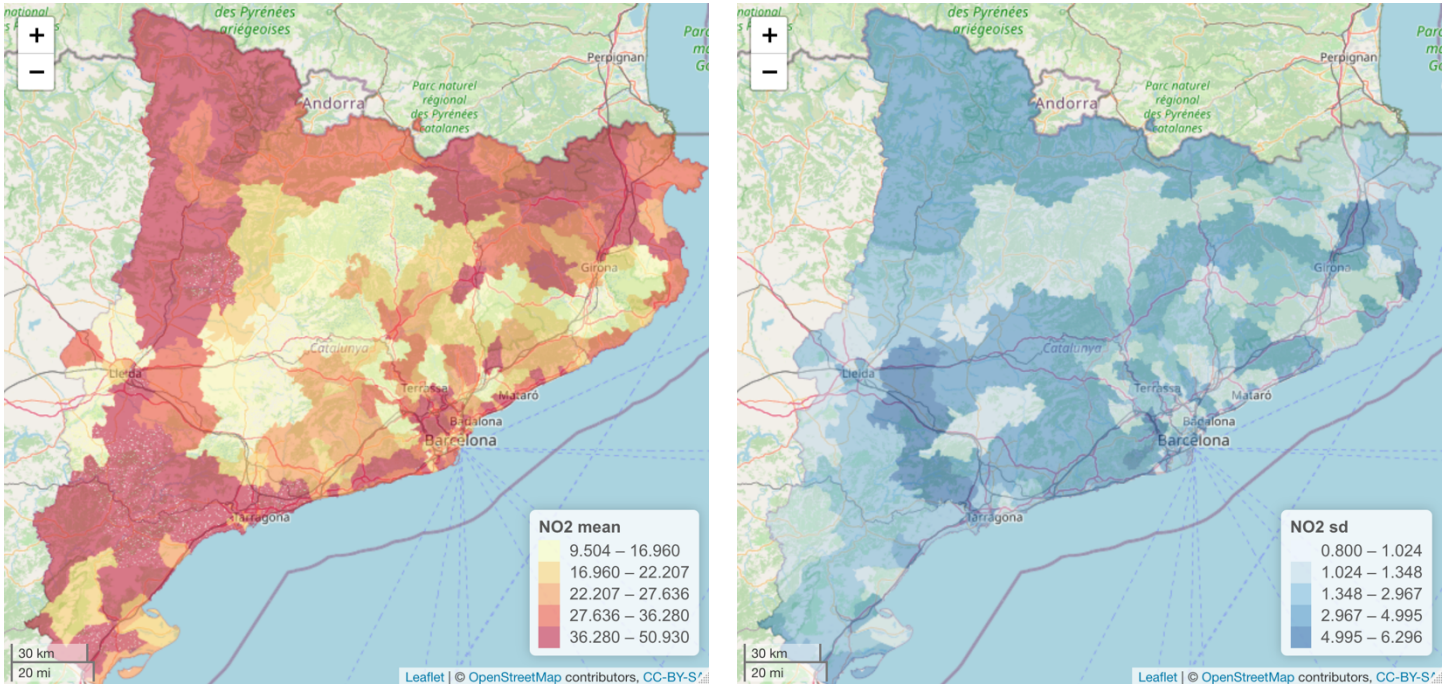
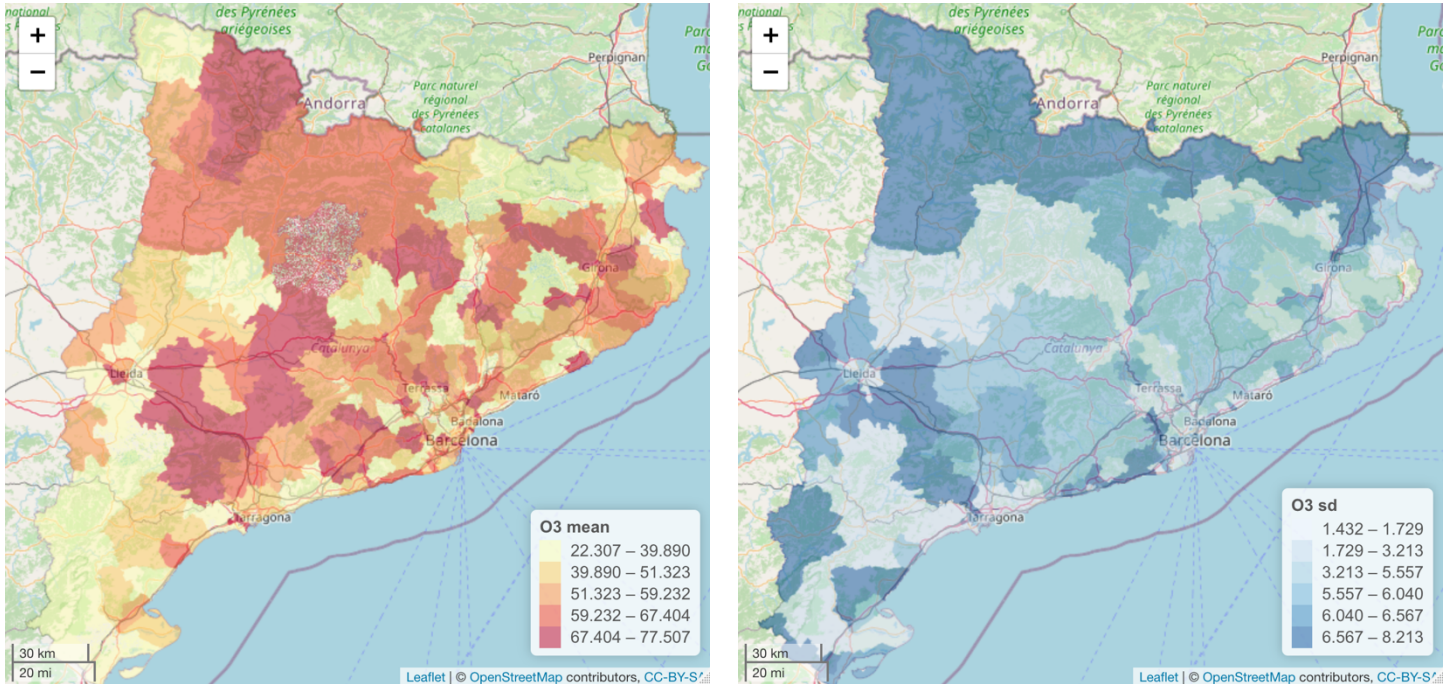



Figure 3c.- Posterior mean and posterior standard deviation of O_3 for 2019. Spatiotemporal process independent in time Gaussian field