A Bayesian Machine Learning Approach for Spatio-Temporal Prediction of COVID-19 Cases

Poshan Niraula (al394267@uji.es)
Universitat Jaume I https://orcid.org/0000-0001-6527-3593
Jorge Mateu
Universitat Jaume I
Somnath Chaudhuri
University of Girona: Universitat de Girona

Research Article

Keywords: Bayesian inference, COVID-19, Neural network, Poisson regression, Public mobility

Posted Date: July 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-636809/v1

SERRA manuscript No.
(will be inserted by the editor)

A Bayesian Machine Learning Approach for Spatio-Temporal Prediction of COVID-19 Cases

Poshan Niraula · Jorge Mateu · Somnath Chaudhuri

Received: date / Accepted: date

Abstract Modeling the behavior and spread of infectious diseases on space and time is key in devising public policies for preventive measures. This behavior is so complex that there are lots of uncertainties in both the data and in the process itself. We argue here that these uncertainties should be taken into account in the modeling strategy. Machine learning methods, and neural networks, in particular, are useful in modeling this sort of complex problems, although they generally lack of probabilistic interpretations. We thus present here a neural network method embedded in a Bayesian framework for modeling and predicting the number of cases of infectious diseases in areal units. A key feature is that our combined model considers the impact of human movement on the spread of the infectious disease, as an additional random factor to the also considered spatial neighborhood and temporal correlation components.

Our model is evaluated over a COVID-19 dataset for 245 health zones of Castilla-Leon (Spain). The results show that a Bayesian model informed by a neural network method is generally able to predict the number of cases of COVID-19 in both space and time, with the human mobility factor having a strong influence on the model.

Keywords Bayesian inference, COVID-19, Neural network, Poisson regression, Public mobility

P. Niraula
Department of Mathematics, University of Jaume I, Castellón, Spain
E-mail: al394267@uji.es

J. Mateu
Department of Mathematics, University of Jaume I, Castellón, Spain
E-mail: mateu@uji.es

S. Chaudhuri
Department of Mathematics, University of Jaume I, Castellón, Spain

Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain
E-mail: al383341@uji.es

1 Introduction

Infectious diseases are the main cause of health hazards in the world and are responsible for deaths of millions of people around the world (WHO, 2019). Various outbreaks of infectious diseases have occurred throughout human history, and indeed there is currently a global health pandemic caused by the novel Coronavirus disease (COVID-19). More than 90 million people have been infected and more than 2 million people have lost their lives (Worldometer, 2020; Wu et al., 2020) as of January 2021 due to COVID-19. To contain the spread of this virus, various regulations such as social distancing measures, travel restrictions, and city or nation-wide lockdowns have been put in place by policy makers around the world. These regulations, although effective in containing the spread of the disease, have also impacted the daily lives of people, social behavior and the global supply chain (Jones et al., 2008). The transmission of general infectious diseases (e.g. COVID-19) exhibits spatio-temporal patterns and can be predicted based on ecological, environmental and socio-economic factors (Anno et al., 2019; Yang et al., 2020). Prediction of these infections is important for government and health workers to plan for effective mitigation by prioritizing the actions of prevention and control measures (Remuzzi & Remuzzi, 2020).

Human movement typically stimulates the introduction of infectious diseases into a new region. There are various evidences that due to human movement, a region-specific disease is introduced to a new region (Stoddard et al., 2009; Nunes et al., 2014) and spreads locally (Stoddard et al., 2013; Gross et al., 2020). Indeed, a number of recent studies have incorporated human movement factors into the modeling strategy (M. U. G. Kraemer et al., 2019; Massaro et al., 2019; Mukhtar et al., 2020). For example, the increased human mobility in western Africa had a high impact in making the Ebola virus catastrophic (Farrar & Piot, 2014). Bogoch et al. (2015) studied the air transport data of flights going out of the Ebola virus affected countries, finding air transport also one of the reasons for the transmission. In the case of COVID-19, it is also seen that the measures related to human movements, such as travel restrictions and social distancing, have been effective in containing the diseases (Fang et al., 2020; M. Kraemer et al., 2020). It is a fact that the introduction of human mobility in epidemiological studies has been more accessible due to technological advancements in locational services and availability of movement data (Guinness, 2016; Sedlar et al., 2019). In this context, availability of technologies such as WiFi or cell phone tower positioning systems and global navigation satellite systems have made the analysis of mobility much easier (Gonzalez et al., 2008; Toch et al., 2019).

The spread of infectious diseases in space and their outbreak in time constitute a complex spatio-temporal problem, which is an effect of complex dynamics of human behavior, environment, and their interactions. Furthermore, as reported in Pan et al. (2020) , during pandemics the human mobility pattern changes compared to that of other times which makes the problem more complex and difficult to analyze. Deep learning methods have proven to be suitable for modeling such complex problems (Mosavi et al., 2020). Indeed, (Ak et al., 2018; Titus Muurlink et al., 2018; Akhtar et al., 2019; Anno et al., 2019; Kapoor et al., 2020; Wieczorek et al., 2020) have used neural networks (some with human mobility data)

to model the spread of infectious diseases. Neural network-based methods rely on a hidden stage to learn from the data and are unable to explicitly account for the spatial and spatio-temporal random effects. However, although these methods have performed well, they are unable to provide uncertainties in the predictions, which we believe are essential in statistical inference and probabilistic forecasting. We argue that predictions accompanied with uncertainties provide further confidence on the results (Beale & Lennon, 2012). To incorporate uncertainties in neural networks, Bayesian neural networks have been developed (Kononenko, 1989; Dhamodharavadhani et al., 2020) and applied over various spatio-temporal problems (McDermott & Wikle, 2019). However, in the field of modeling and understanding the dynamics of COVID-19, the use of neural networks in combination with Bayesian inference is limited. Cabras (2020) presented a method of combining neural networks with Bayesian inference having a focus on COVID-19 infections in Spain. However, mobility and its influences were not considered. As spatio-temporal predictions help in understanding the spread of the disease to further identify the regions of high risk, a large number of papers can be found in the field of spatio-temporal modeling of diseases. Among them, generalized linear models (GLM) with the addition of spatial effects of nearby places and/or temporal effects from past events are found to be often used and proven to be useful in prediction (Guo et al., 2017; Cabrera & Taylor, 2019; Giuliani et al., 2020). For example, Giuliani et al. (2020) have used GLM to predict COVID-19 infections in regions of Italy, and found the spatial interactions of nearby places to have a high influence on modeling; this shows the importance of accounting for the spatial effects explicitly. In a parallel vein, Bayesian modeling methods have also been used in this epidemiological context (Gelman et al., 2013; Aswi et al., 2019; Song et al., 2019; Torres-Signes et al., 2020).

In this context, our approach focuses on the use of deep learning methods (using a Long Short Term Memory-LSTM) informing a Poisson regression model in a Bayesian framework to model and predict the spread and outbreak of COVID-19 with uncertainties. In particular, human mobility data along with socio-demographic variables are incorporated in the combined model to predict the dynamics of COVID-19. In doing so, we highlight the importance of human mobility in modeling the dynamics of infectious diseases.

The plan of the paper is as follows. Section 2 presents the data along with all covariates considered in the model to motivate the proposed statistical model. We also consider some spatial weights built from the movement data. Section 3 presents the statistical model, and the results come in Section 4. The paper ends with some conclusions and a discussion in Section 5.

2 Study area and data

Daily COVID-19 infections aggregated per 245[1] health zones in the community of Castilla-Leon (Spain) were used in this paper. The temporal range goes from March 1, 2020 to February 5, 2021. Castilla-Leon is the largest community in Spain

[1] Here, the health zones SORIA NORTE, SORIA SUR and SORIA RURAL are aggregated to a single unit.

by area located in the northwest part of Spain. This region has a population of around 2.5 million and is ranked third among the communities in offering social services to the citizens. Figure 1 shows the location map of the study area and also cumulative COVID-19 cases per 10000 inhabitants per health zone until February 5, 2021. We note that COVID-19 has spread throughout the study area with clusters around major urban areas.

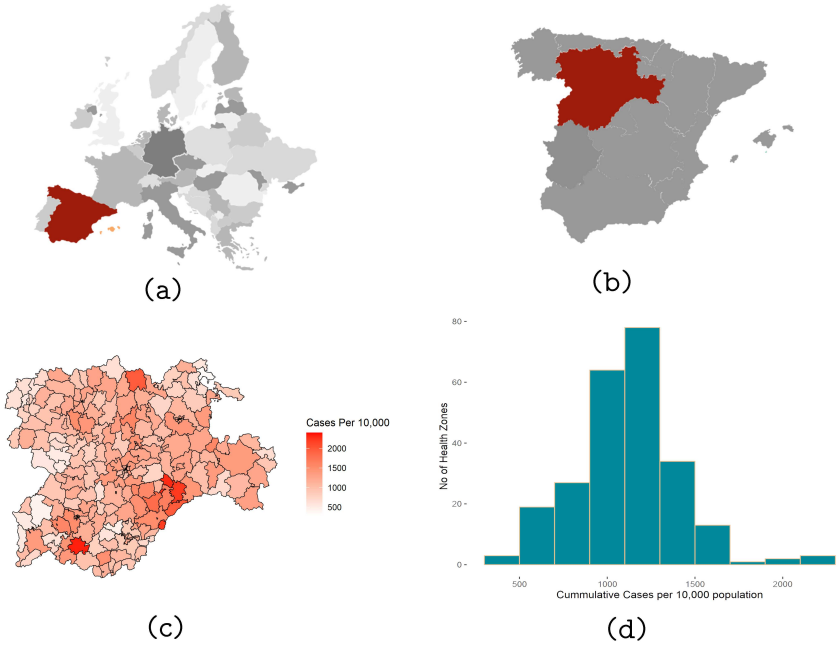

Fig. 1: (a) Location of Spain; (b) Location of Castilla-Leon in Spain; (c) Cumulative numbers of COVID-19 cases per 10000 inhabitants and health zones; (d) Histogram of cumulative cases per health zones of Castilla-Leon

COVID-19 cases data were retrieved from the open data portal of Castilla-Leon [2]. Similarly, the socio-demograhic datasets and the health zone boundary, in shapefile form, were downloaded from the open data platform of Instituto Nacional de Estadística [3]. The human mobility data for the study area was acquired from Barcelona Supercomputing Center flowmap dashboard [4]. A brief description and source of the datasets used in the current paper are reported in Table 1.

Figure 2 shows the daily number of COVID-19 cases per 10000 inhabitants. The highlighted red line represents the daily mean number of cases per 10000 inhabitants. The cases increased in March and April 2020 (defining the first wave), and then started to decrease until August 2020 due to the imposed lockdown

[2] https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html
[3] https://www.ine.es/en/index.htm
[4] https://flowmaps.life.bsc.es/flowboard/

Table 1: Summary of data used and their sources

Data	Data Sources	Description of data
COVID-19	Open data portal of Castilla-Leon	Daily infected cases at health zone level
Mobility data	Barcelona Supercomputing Center	Daily human mobility matrices at municipality level
Socio-demographic	Open data portal of Castilla-Leon	Individual health zone total population, unemployment level and number of urban offices
Geometry	Open data portal of Castilla-Leon	Boundary shapefiles of 245 health zones

measures. However, due to a certain relaxation towards the summer period, the cases started to increase late August to end up with a second wave in October and November 2020. A third wave of infection is noted in January and February 2021, and started to decrease again due to some partial restrictions and the onset of the vaccination program. Similarly, weekly trends in the number of cases is visible with a drop of cases on weekends, due to the reduced number of tests done over the weekends.

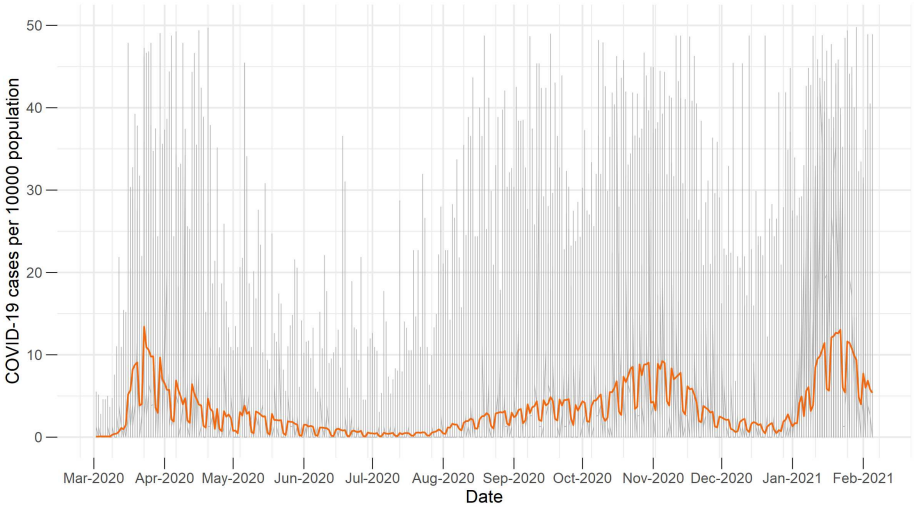

Fig. 2: Temporal trend of COVID-19 cases in the study area. The red line represents the daily mean number of cases per 10000 inhabitants.

The mobility data acquired from the data portal of Barcelona Supercomputing Center was prepared by the Ministry of Transport, Mobility, and Urban Agenda. The data was preprocessed to guarantee anonymized records from mobile phones.

These recorded events contain both active events also known as Call Detail Records (CDR) and passive events with a periodic update of device position, change of coverage area, etc. The location information is at the level of the coverage area of each antenna, which is merged to create origin-destination matrices at municipality, districts and provinces level. Along with these records from the cell phones, landuse data, population data, transport network data such as train lines, and location of airports have been used to create the merged matrices (Ministry of Transport & Agenda, 2020). The available daily mobility data was at the municipality level; those municipalities with population less than 1000 were combined to form aggregated zones. As all other available data were at the health zones level, these aggregations were converted to the health zone level by applying spatial overlay functions and dividing the movement data in proportion to the area. The socio-demographic covariates considered in this paper were the following: total population per health zone, number of people demanding for employment, number of unemployed people, number of commercial units, office units, and industrial units in the urban areas of each health zone (see a description in Table 2). Additionally, we also considered some built-in variables (see Table 3). In particular, we computed the average number of cases and average number of deaths in the direct neighborhood. The cumulative cases of COVID-19 for the last 14 days were also computed to consider the aggregated impact for a short time frame.

Table 2: Summary of socio-demographic variables

Variable Name	Description
`total_pop`	total population of the health zone
`demanding_total_employment`	Number of people demanding for employment
`registered_unemployed_total`	Number of people registered as unemployed
`number_of_urban_commercial_units`	Number of commercial offices in the urban areas
`number_of_urban_industrial_units`	Number of industrial units in the urban areas
`number_of_urban_office_units`	Number of offices units in the urban areas

Table 3: Summary of built-in variables

Variable Name	Description
Day of the week	Computed from the date
Cumulative cases	Cumulative number of cases for last 14 days
Average number of cases in neighboring health zones	Average of number of cases in health zones that share a common border
Average number of deaths in neighboring health zones	Average number of deaths in health zones that share a common border

Last, but not least, we introduce new spatial weights based on the movement data that represent the associated movement-based risk. These weights are com-

puted per health zone and day. We add a temporal lag to handle past-term movement data and the daily data are weighted depending on the temporal distance.

These spatial weights take into account the mobility from all other regions j into region i, and the weights are interpreted as the chance of a moving person to import the infection of the disease into region i from all the other regions. This spatial weight for a region i and day t, $W_{i,t}$, can be computed as

$$W_{i,t} = \sum_{j=1}^{n} \left[\sum_{t'=t-1}^{t-\Delta t} m_{ji,t'} * w'_{t'} \right] * \frac{I_{j,t}}{P_j} \tag{1}$$

where n is total number of regions, $m_{ji,t}$ is the mobility from all regions j to i on day t, $I_{j,t}$ is the number of infected cases at region j at time t, P_j is the total population of the region j and w'_t is the weight given to the mobility data on day t.

A time lag Δt is added to the computation of the spatial weights as the spread of a disease on the region is dependent on the mobility and infections on past days in all other regions of the study area. We used a 7-day lag as infection is assumed to act a week before first symptoms. We assigned the following weights: given t, we give $t-1$ and $t-2$ only a weight of 5%, this weight increases up to 10% for $t-3$ and $t-4$, then goes up to 20% for $t-5$ and $t-6$, and finally the weight is 30% for $t-7$.

Figure 3 shows the temporal series of the spatial weights for 4 selected health zones along with the daily number of COVID-19 cases for the study period. It is evident that increasing weights correspond to increased COVID-19 cases. Similarly, Figure 4 shows the flowmap of the median mobility for the week 2021-01-30 till 2021-02-05, prepared with the flowmapblue R package [5], and the spatial distribution of the spatial weights for the same period.

Summarizing, our model is feeded by COVID-19 covariates, socio-demographic covariates and human movement-related covariates. COVID-19 covariates include cumulative cases, average number of cases in neighboring health zones, deaths and average number of deaths in neighboring health zones, and spatial weights computed from the daily mobility matrices and infection. A temporal covariate, day of the week, was computed as a factor from the date.

[5] https://github.com/FlowmapBlue/flowmapblue.R

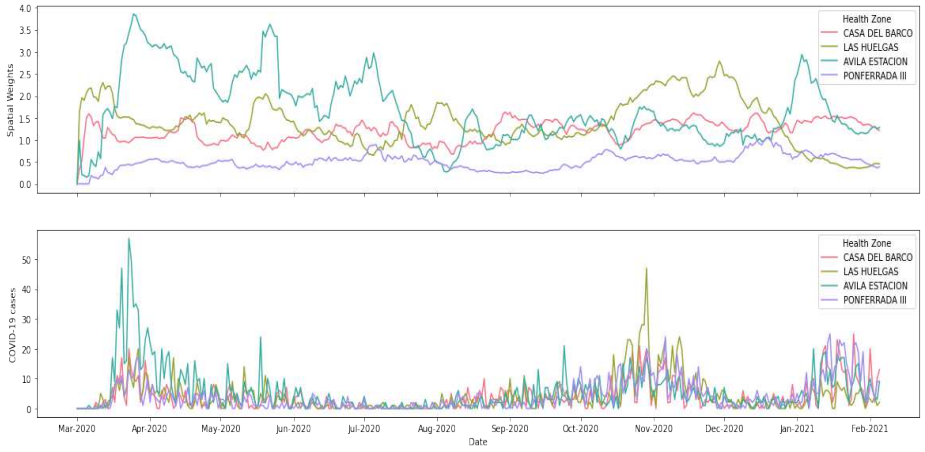

Fig. 3: Spatial weights and COVID-19 cases for the selected health zones

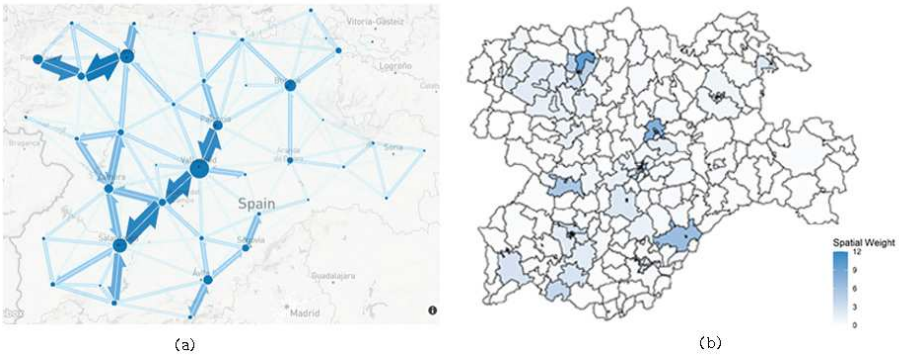

Fig. 4: For the last week of study period 2021-01-30 till 2021-02-05: (a) Flowmap of the study area with the median mobility; (b) Spatial distribution of median values of spatial weights

3 A Bayesian LSTM method

We use here the term Bayesian LSTM method, to indicate that we use a statistical model within a Bayesian framework informed by the output of a Long Short Term Memory (LSTM) neural network method. We aim to model the number of infections on an areal unit, in our case health zones, based on spatial covariates, temporal trends, and mobility matrices. Thus our combined model considers temporal and spatial dependence structures, and provides predictions in space and time of the number of infections.

Figure 5 shows a graphical overview of the proposed model which contains two major components: (a) a deep learning method (LSTM), and (b) a Bayesian spatial Poisson regression model. The input to the LSTM method are the temporal

series of the cases of infections. The LSTM method learns from these temporal series and predicts the number of cases in the future. Predictions from the LSTM method are embedded into the Poisson regression as an expected value. The spatial correlation structure is modeled using a stochastic partial differential equation (SPDE) method through the INLA approach.

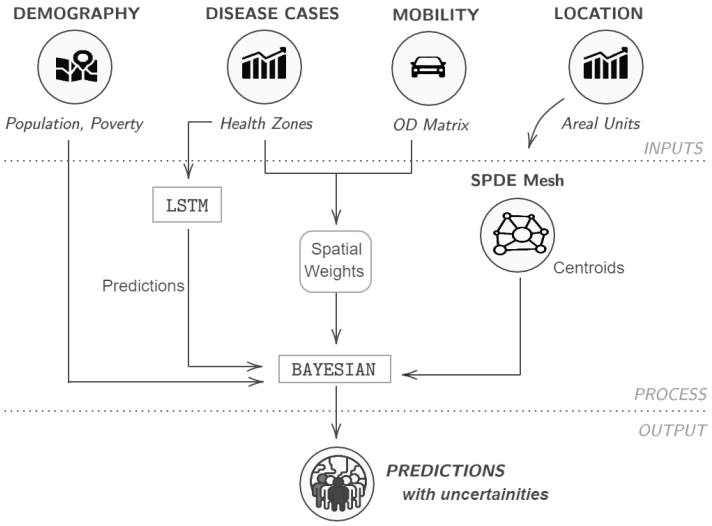

Fig. 5: Graphical overview of the Bayesian LSTM method

3.1 LSTM method

Artificial neural networks are a class of machine learning methods inspired by the functioning of human brain and work on the principle of parallel processing. They consist of layers of interconnected processors known as neurons, which have a vector of weights associated with them. Artificial neural networks models consist of input data also known as input layer, layers of interconnected neurons also known as hidden states, and the output layer which is the output of the model. Fitting an artificial neural network involves estimating the optimal value of these weights which are able to accurately reproduce and mimic some training data. The optimization of these weights is done through the gradient descent method, and the weights assigned to each layer are adjusted proportionally to the derivatives (Bengio et al., 1994).

Among many types of artificial neural networks, recurrent neural networks are arguably the most useful ones for sequential data (as time series) as they have a stack of non-linear units that can learn even long-term dependencies of time series

data (Bengio et al., 1994). In recurrent neural networks, the configuration of hidden states acts as the network memory and the hidden layer state at a time is dependent on its previous state which enables to learn from past data, thus handling long-term dependencies (Mikolov et al., 2014). This makes recurrent neural network an excellent choice for learning and predicting time-dependent data. However, despite having these advantages, as the recurrent neural networks perform the gradient descent method with each timestamp of the data, they are likely to fall into the gradient vanishing problem. Due to this problem, as the recurrent neural network loops through the networks recurrent connections, the effect of a given input on hidden layers, and consequently on the output, either decays or explodes exponentially (Hochreiter, 1991). One alternative approach to tackle this problem comes from using a LSTM method (Hochreiter & Schmidhuber, 1997), that solves the gradient vanishing problem by introducing LSTM memory cells instead of the hidden units. These LSTM cells consist of input, output and forget gates; the input and output gates are used for the control of the flow of memory cell input and output into the rest of the model, whereas the forget gates are responsible for learning the weights that control the rate at which the value stored in the memory cell decays. With the addition of these gates, the LSTM is able to bypass the vanishing gradient problem while also learning from the long term dependencies in the data (Salehinejad et al., 2018).

In our case, the LSTM method accounts for the temporal trend of the COVID-19 spread, learning from the temporal trend of the infected cases on individual health zones separately, rather than considering the spatial cross-correlation amongst the regions. Since the LSTM methods are mostly applicable for temporal series data, it can be assumed that the LSTM method learns more from the temporal trend in the infection on individual health zones separately than from the spatial relationship between the health zones, which will be further accounted for in the Bayesian regression model.

3.1.1 Architecture

We used a four layered LSTM, for which the first layer is the input layer given by the daily time series of COVID-19. In order to create a supervised learning problem, the temporal series of infected cases were converted to an input-output pair which is performed by shifting the data (Brownlee, 2017). Thus, for every time step t of the time series, one day ahead shifting is done in the data to create a shifted prediction at $t + 1$. The second layer of the model consists of the 128 LSTM memory cells; similarly, the third and fourth layers consist of 64 and 32 memory cells, respectively. This number of memory cells in each layer comes from experimentation and also motivated by previous works (Shahid et al., 2020). With this configuration, the model has 131489 parameters consisting of three stacked LSTM layers which are recurrently used for the time period T (equal to the total number of days under study). Finally, a dense layer connects all the recurrent layers and connects them to the output layer. The dense layer has the linear activation function. The architecture of the LSTM method is shown in Figure 6. Additional parameters and hyper-parameters that define the LSTM method are shown in more detail in Appendix B (Table 6).

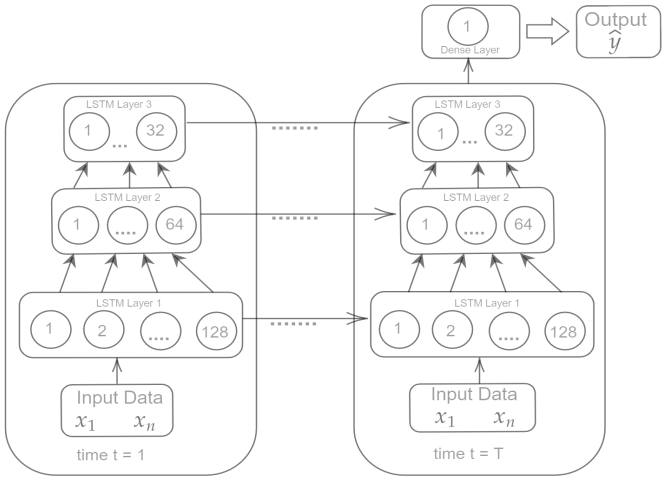

Fig. 6: Architecture of the LSTM method

3.2 Spatio-temporal Poisson regression and Bayesian inference

To deal with uncertainty, we consider in a second stage a spatio-temporal stochastic model for the counts of COVID-19 infected cases, which is informed by the output of LSTM run at a first stage.

Let Y_{it} and E_{it} be the number of observed and expected cases in the i-th area (health zone) and the t-th period (day), $t = 1, \ldots, T$. We assume that conditional on the relative risk, ρ_{it}, the number of observed cases follows a Poisson distribution

$$Y_{it}|\rho_{it} \sim Po(\lambda_{it} = E_{it}\rho_{it})$$

where, the log-risk is modeled as

$$log(\rho_{it}) = \beta_0 + Z_{it}^T\beta_{it} + S(x_i) \quad (2)$$

with $S(.)$ a spatially structured random effect, and the Z_{it} stand for the covariates (as mentioned in Section 2). We assigned a vague prior to the vector of coefficients $\beta = (\beta_0, \ldots, \beta_p)$ which is a zero mean Gaussian distribution with precision 0.001. Finally, all parameters associated to log-precisions are assigned inverse Gamma distributions with parameters equal to 1 and 0.00005.

To compute the joint posterior distribution of model parameters, Bayesian inference has traditionally relied upon Markov Chain Monte Carlo (MCMC) (Gilks, 1996; Brooks, 2011). This distribution is often in a high dimensional space and thus it is computationally very expensive. As an alternative computationally faster solution, Rue et al. (2009) developed a new approximation to the posterior marginal distributions of model parameters based on a Laplace approximation, and named it as integrated nested Laplace approximation (INLA). INLA focuses on models that can be expressed as latent Gaussian Markov random fields (GMRF). In

particular, we use a stochastic partial differential equation (SPDE) method, as introduced by (Lindgren et al., 2011). SPDE consists in representing a continuous spatial process like a Gaussian field (GF) using a discretely indexed spatial random process such as a Gaussian Markov random field (GMRF). In particular, the spatial random process $S(.)$ follows a zero-mean Gaussian process with Matérn covariance function represented as

$$Cov(S(x_i), S(x_j)) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa||x_i - x_j||)^{\nu} K_{\nu}(\kappa||x_i - x_j||) \tag{3}$$

where $K_{\nu}(.)$ is the modified Bessel function of second order, and $\nu > 0$ and $\kappa > 0$ are the smoothness and scaling parameters, respectively. INLA approach constructs a Matérn SPDE model, with spatial range r and standard deviation parameter σ. The model parameterization is expressed as

$$(\kappa^2 - \Delta)^{(\alpha/2)}(\tau S(x)) = W(x)$$

where $\kappa = \sqrt{8\nu}/r$ is the scale parameter, $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ is the Laplacian operator, $\alpha = (\nu + d/2)$ is the smoothness parameter, τ is inversely proportional to σ and $W(x)$ is a spatial white noise (Blangiardo & Cameletti, 2015). Note that we have $d = 2$ for a two-dimensional process, and we fix $\nu = 1$, so that $\alpha = 2$ in our case.

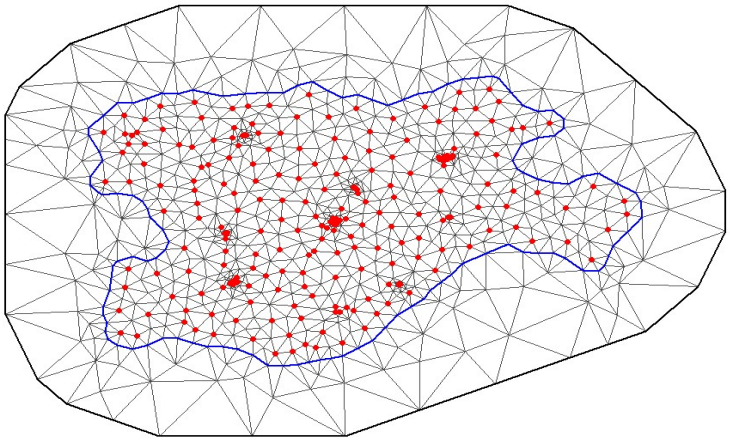

Fig. 7: SPDE triangulation for the study area of Castilla-Leon

We use the centroids of each health zone as the target locations over which we build the mesh. The mesh is formed by smaller triangles within the region of interest, and by larger ones outside the region. The constrained refined Delaunay triangulation is illustrated in Figure 7. The blue line highlights the outline boundary of the study area, with the red dots indicating the centroids of the individual health zones. Note that some few regions show sort of clusters due to the close proximity of health zones. We generate the projection matrix to project the spatially continuous Gaussian random field from the observations to the mesh nodes.

Centroids of individual health zones and the triangulations in the mesh are used to generate the projection matrix. We fixed $r = 0.1$ and $\sigma^2 = 1$. Parameters τ and κ are renamed as $\theta_1 = log(\tau)$ and $\theta_2 = log(\kappa)$, and we assign them zero mean vague Gaussian independent priors with precisions equal to 0.1.

4 Results

We fitted our Bayesian neural network approach (named as LSTM-INLA throughout this section) and compared it with two other baseline models, one which is only using a LSTM method (named as LSTM) and the other one that only fits a spatial Poisson regression with INLA and no LSTM (named as INLA). We fitted the models for all the temporal range except for the last week, and used these last 7 days for prediction. The models were evaluated using the averaged Root Mean Squared Error (RMSE) from all health zones. Additionally, we also considered the Bayesian metrics Watanabe Akaike information criterion (WAIC) (Watanabe, 2010), deviance information criterion (DIC) (Spiegelhalter et al., 2002) and conditional predictive ordinate (CPO) (Pettit, 1990).

Table 4 shows the corresponding metrics, with RMSE evaluated over the training period (RMSE Training) and over only the prediction period (from 2021-01-30 to 2021-02-05, RMSE Prediction).

Table 4: Metrics for model evaluations

Model	RMSE Training	RMSE Prediction	DIC	WAIC	CPO
INLA	5.33	14.24	373184.93	375164.6	-2.29
LSTM	4.44	6.07	-	-	-
LSTM-INLA	*4.14*	*5.51*	*354601.13*	*355510.1*	*-2.17*

*The best model is marked in italics.

The RMSE for the LSTM-INLA model is lower than the INLA and LSTM methods for both the training and prediction periods. We note that although the RMSE for the training set is quite as good as for the other two methods, the RMSE for the prediction set for INLA and LSTM is far larger. This suggests that inclusion of LSTM as an expected value for the spatial Poisson regression plays an important role. Similarly, the comparison of INLA and LSTM-INLA models with DIC, WAIC and CPO metrics, shows that the LSTM-INLA combination provides the best fit. Table 5 shows the correlation between the observed values and the predicted ones for the prediction period (recall this is the last week of the overall temporal range). The correlation is largest when using the combined LSTM-INLA model (0.80) reinforcing the goodness-of-fit of our proposal.

Figure 8 depicts the observed cumulative cases of COVID-19 at three selected weeks within the overall temporal range and chosen at different phases of the pandemic. We also show the corresponding predictions from the LSTM method and the combined LSTM-INLA model. In particular, first row of Figure 8 represents the cumulative number of cases on the initial week of COVID-19 spread in Spain,

Table 5: Correlation between observed and predicted cases (2021-01-30 to 2021-02-05)

Model	Correlation
INLA	0.77
LSTM	0.75
LSTM-INLA	*0.80*

2020-03-22 to 2020-03-28, second row is for the week 2020-10-18 to 2020-10-24, and third row stands for the 7-days prediction ahead period, from 2021-01-30 to 2021-02-05. A map depicting the prediction from the LSTM-INLA model and observed cases for the final week of the study period is published in an R-Shiny app, which can be accessed through the link[6]. A sample view of the shiny app is presented in Figure 13 in Appendix A.

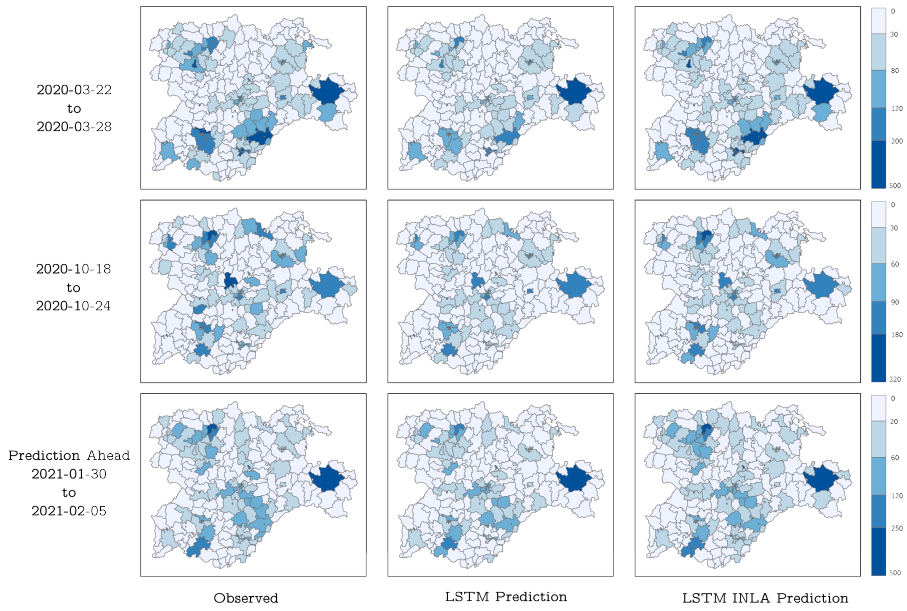

Fig. 8: Spatial distribution of the observed cases (left column) of COVID-19 for three selected weeks. Prediction from the LSTM method (central column) and from LSTM-INLA model (right column)

To visualize the temporal trends, Figure 9 shows the observed cases together with the predicted ones for four selected health zones (Avila Estacion, Las Huelgas, Casa del Barco and Ponferrada-II). In particular, we note that we can draw, together with the predictions under LSTM-INLA, the corresponding 95% credible

[6] https://poshan-niraula.shinyapps.io/CYLCovidPrediction/

interval, providing a measure of the uncertainty associated to the prediction, thing that we can not obtain under LSTM alone. Comparing the prediction from the LSTM method (green lines), the LSTM-INLA prediction with 95% credible interval (blue lines) with the observed cases (red lines), we note the better prediction results when using LSTM-INLA. Figure 14 in Appendix D shows the corresponding residual plots. They suggest the better behavior of the LSTM-INLA model as they are lower in magnitude and symmetrically distributed around the zero line. This is also true to the prediction ahead case.

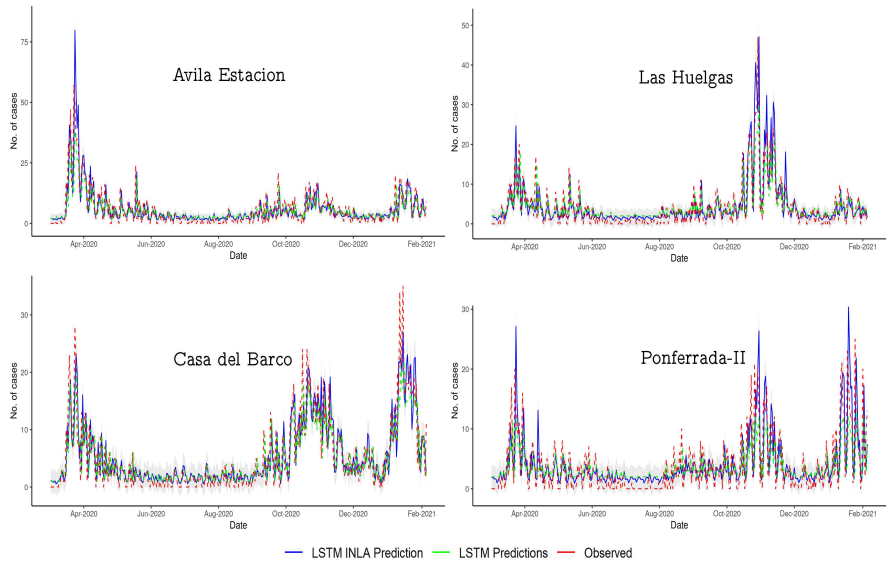

Fig. 9: Temporal trend plots of the observed and predicted cases with LSTM and LSTM-INLA models for four selected health zones. The grey band stands for the 95% credible interval under the LSTM-INLA model

Having in mind the model described in equation 2, we now put in place some information related to the posterior distribution of fixed and random effects. In particular, Figure 10 depicts the marginal posterior mean and 95% credible intervals of spatial random effect $S(.)$. ID in the X-axis of Figure 10 represents 799 triangulation nodes of the SPDE mesh used in the model. A stronger and significative spatial effect is observed basically on the nodes of smaller triangles within the region of interest (as shown in Figure 7). The nodes outside the region show no spatial effect.

Additionally, Figure 11 and Table 7 in Appendix C depict the marginal posterior distributions of all fixed effects including the intercept (β_0) and the other covariates. We note that four covariates, namely number of people demanding for employment, number of commercial offices, number of industrial units and number of office units in the urban areas, have no influence in our model. The positive

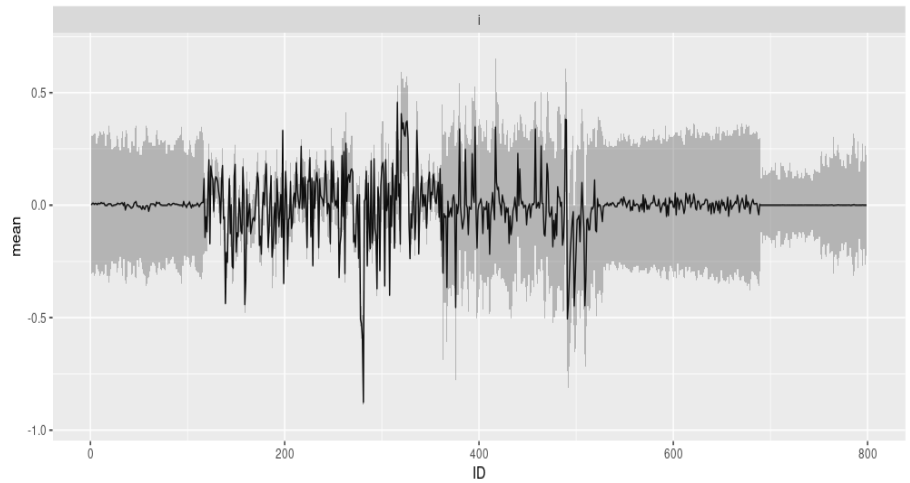

Fig. 10: Marginal posterior mean of the spatial random effect $S(\cdot)$

mean values for covariates such as average cases in neighbouring health zones, cumulative cases, or deaths indicate positive influence in the model. The covariate associated to daily movement (spatial weight) has the highest positive mean value which indicates strong positive influence of human mobility on the model.

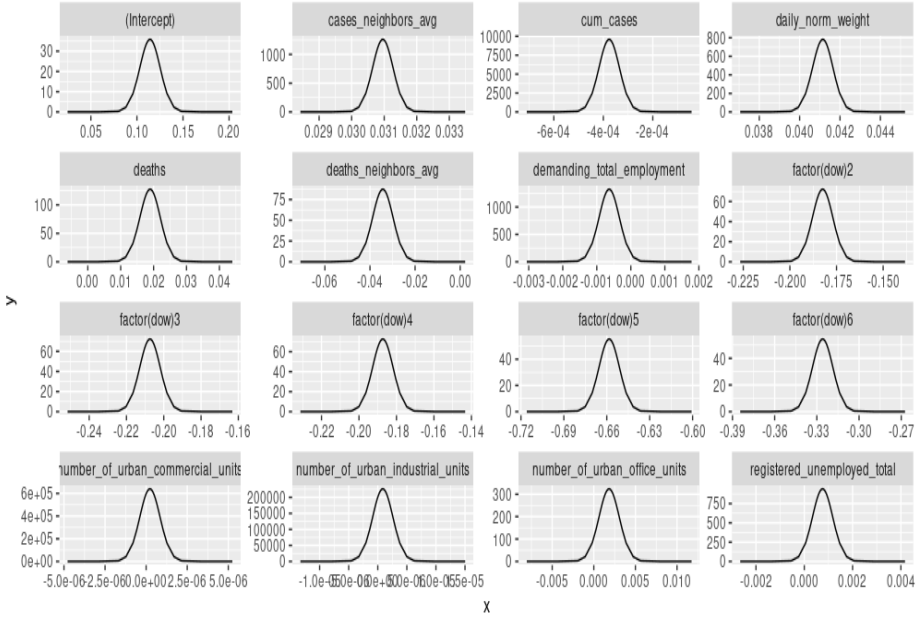

Fig. 11: Marginal posterior distributions of covariate coefficients

Finally, Figure 12 shows the marginal posterior Gaussian distributions of the two hyperparameters for the spatial random field θ_1, θ_2. Mean and variance for the two hyperparameters are $\theta_1 = (-3.10, 0.142)$, and $\theta_2 = (3.35, 0.099)$.

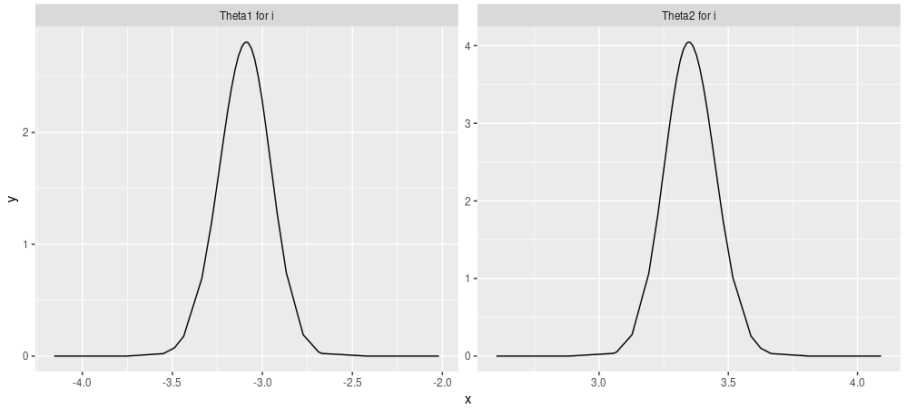

Fig. 12: Hyperparameters θ_1 and θ_2 for the spatial random field $S(\cdot)$

5 Conclusions and discussion

For modeling the spread and outbreak of infectious diseases, a model comprising the combination of neural network and Bayesian inference for a spatio-temporal Poisson regression has been proposed. This model is able to provide good predictions of further cases of COVID-19 while handling uncertainties. In particular, our model has two components, a LSTM neural network, which learns from the temporal patterns, and a spatial Poisson regression with expected values the predictions coming from the LSTM. The spatio-temporal Poisson regression considers various spatial and temporal covariates. It is noteworthy that we consider daily matrices of population movement that are transformed into spatial weights and act as additional covariates in the model.

The proposed model was evaluated with COVID-19 daily infected cases in Castilla-Leon (Spain), consisting of 245 health zones, and within a temporal range running from March 1, 2020 to February 5, 2021. The combined model was able to predict the number of daily infections in each health zone, outperforming two other cases, one with only a neural network method and the other with only a spatio-temporal Poisson regression. A key and novel aspect is the introduction as a spatial weight of the population movement, being highly influential in the overall fit. However, we note that sudden increasing peaks or abrupt decreasing magnitudes can not be finely fitted by our model. We believe this is due to typos, errors or under-reporting actions, and they clearly mean a challenge for modeling purposes of this sort of data.

Clearly, the accuracy of prediction may be improved by the addition of other variables relevant to the disease of study which may include the weather conditions and preventive measures. The phenomenon of infectious disease spread has a lot of complexities and is dependent on numerous factors. These factors include the organism causing the disease, the mode of transmission, human behaviors, environmental conditions, and most importantly, some potential preventive measures applied. All of these factors are not quantifiable but a maximum number of these factors are to be considered while modeling the diseases. In this study, one of the most relevant considered factors is human mobility. Some socio-demographic variables were considered but we believe more variables associated with the socio-demography and climatic conditions can be introduced. Similarly, the variables related to human behavior and preventive measures such as social distancing and personal hygiene should be incorporated in future works.

The focus of this work is on the combination of neural networks and Poisson regression within a Bayesian framework. The predictions from neural networks were used as expected values for the Poisson regression which can be improved by transferring the predictions to a prior distribution and use them as prior information in the Bayesian inference. Here we followed a two-stage procedure, but ideally it would be better a joint solution such as spatio-temporal recurrent neural networks able to predict results with uncertainties. Finally, the proposed method is applied only in one scenario of COVID-19 infection for a short period. Thus, data with a longer period and different spatial scales should be used to test the versatility of the model.

The model is believed to be useful for the governments in monitoring any infectious diseases. The results from the model can be used in formulating health-related policies such as the application of preventive measures or vaccination. The contribution of this work is that it is able to take advantage of the neural network methods in learning complex dependencies from the data, as well as from a Bayesian paradigm to associate the uncertainties in the predictions. In conclusion, this work is able to present a model that can provide accurate predictions of infectious diseases and help in a way to mitigate the impacts.

Acknowledgements J. Mateu has been partially funded by grants PID2019-107392RB-I00 from Ministry of Science and Innovation and UJI-B2018-04 from University Jaume I. P. Niraula has been funded through the Erasmus Mundus programme by the European Commission under the Framework Partnership Agreement, FPA-2016-2054.

References

Ak, Ç., Ergönül, Ö., Şencan, İ., Torunoğlu, M. A., & Gönen, M. (2018). Spatiotemporal prediction of infectious diseases using structured Gaussian processes with application to Crimean–Congo hemorrhagic fever (M. A. Rabaa, Ed.). *PLOS Neglected Tropical Diseases*, *12*(8), e0006737. https://doi.org/10.1371/journal.pntd.0006737

Akhtar, M., Kraemer, M. U., & Gardner, L. M. (2019). A dynamic neural network model for predicting risk of Zika in real time. *BMC Medicine*, *17*(1), 171.

Anno, S., Hara, T., Kai, H., Lee, M.-A., Chang, Y., Oyoshi, K., Mizukami, Y., & Tadono, T. (2019). Spatiotemporal dengue fever hotspots associated with climatic factors in Taiwan including outbreak predictions based on machine-learning. *Geospatial Health*, *14*(2), 183–194. https://doi.org/10.4081/gh.2019.771

Aswi, A., Cramb, S. M., Moraga, P., & Mengersen, K. (2019). Bayesian spatial and spatio-temporal approaches to modelling dengue fever: A systematic review. *Epidemiology and Infection*, *147*, e33. https://doi.org/10.1017/S0950268818002807

Beale, C. M., & Lennon, J. J. (2012). Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1586), 247–258.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, Ltd.

Bogoch, I. I., Creatore, M. I., Cetron, M. S., Brownstein, J. S., Pesik, N., Miniota, J., Tam, T., Hu, W., Nicolucci, A., Ahmed, S., Yoon, J. W., Berry, I., Hay, S. I., Anema, A., Tatem, A. J., MacFadden, D., German, M., & Khan, K. (2015). Assessment of the potential for international dissemination of Ebola virus via commercial air travel during the 2014 west African outbreak. *The Lancet*, *385*(9962), 29–35. https://doi.org/10.1016/S0140-6736(14)61828-6

Brooks, S. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press.

Brownlee, J. (2017). *Introduction to time series forecasting with Python: How to prepare data and develop models to predict the future*. Machine Learning Mastery.

Cabras, S. (2020). A Bayesian - Deep Learning model for estimating COVID-19 evolution in Spain. http://arxiv.org/abs/2005.10335

Cabrera, M., & Taylor, G. (2019). Modelling spatio-temporal data of dengue fever using generalized additive mixed models. *Spatial and Spatio-temporal Epidemiology*, *28*, 1–13.

Dhamodharavadhani, S., Rathipriya, R., & Chatterjee, J. M. (2020). COVID-19 mortality rate prediction for india using statistical neural network models. *Frontiers in Public Health*, *8*.

Fang, H., Wang, L., & Yang, Y. (2020). *Human Mobility Restrictions and the Spread of the Novel Coronavirus (2019-nCoV) in China* (tech. rep. w26906). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w26906

Farrar, J. J., & Piot, P. (2014). The Ebola Emergency — Immediate Action, Ongoing Strategy. *New England Journal of Medicine*, *371*(16), 1545–1546. https://doi.org/10.1056/NEJMe1411471

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

Gilks, W. R. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Giuliani, D., Dickson, M. M., Espa, G., & Santi, F. (2020). Modelling and predicting the spatio-temporal spread of COVID-19 in Italy. *BMC Infectious Diseases*, *20*(1), 700. https://doi.org/10.1186/s12879-020-05415-7

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns [arXiv: 0806.1256]. *Nature*, *453*(7196), 779–782. https://doi.org/10.1038/nature06958

Gross, B., Zheng, Z., Liu, S., Chen, X., Sela, A., Li, J., Li, D., & Havlin, S. (2020). Spatio-temporal propagation of COVID-19 pandemics. *medRxiv*.

Guinness, R. (2016). Advances in mobile location services and what it means for privacy. *European Journal of Navigation*, *14*, 19–24.

Guo, C., Du, Y., Shen, S., Lao, X., Qian, J., & Ou, C. (2017). Spatiotemporal analysis of tuberculosis incidence and its associated factors in mainland China. *Epidemiology & Infection*, *145*(12), 2510–2519.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, *91*(1).

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, *451*(7181), 990–993. https://doi.org/10.1038/nature06536

Kapoor, A., Ben, X., Liu, L., Perozzi, B., Barnes, M., Blais, M., & O'Banion, S. (2020). Examining COVID-19 forecasting using spatio-temporal graph neural networks. *arXiv preprint arXiv:2007.03113*.

Kononenko, I. (1989). Bayesian neural networks. *Biological Cybernetics*, *61*(5), 361–370.

Kraemer, M. U. G., Golding, N., Bisanzio, D., Bhatt, S., Pigott, D., Ray, S., Brady, O., Brownstein, J., Faria, N., Cummings, D., et al. (2019). Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. *Scientific Reports*, *9*(1), 1–11.

Kraemer, M., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., Brownstein, J. S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. G., & and, S. V. S. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, *368*(6490), 493–497. https://doi.org/10.1126/science.abb4218

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498.

Massaro, E., Kondor, D., & Ratti, C. (2019). Assessing the interplay between human mobility and mosquito borne diseases in urban environments. *Scientific Reports*, *9*(1), 1–13.

McDermott, P. L., & Wikle, C. K. (2019). Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. *Entropy*, *21*(2), 184.

Mikolov, T., Joulin, A., Chopra, S., Mathieu, M., & Ranzato, M. (2014). Learning Longer Memory in Recurrent Neural Networks. *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*. http://arxiv.org/abs/1412.7753

Ministry of Transport, M., & Agenda, U. (2020). *Analysis of mobility in Spain with Big Data technology during the state of alarm for the management of the COVID-19 crisis* (tech. rep.). Madrid. https://cdn.mitma.gob.es/portal-web-drupal/covid-19/estudio/MITMA-Estudio_Movilidad_COVID-19_Informe_Metodologico_v012.pdf

Mosavi, A., Ardabili, S., & Varkonyi-Koczy, A. (2020). List of deep learning models. https://doi.org/10.1007/978-3-030-36841-8_20

Mukhtar, A. Y. A., Munyakazi, J. B., & Ouifki, R. (2020). Assessing the role of human mobility on malaria transmission. *Mathematical Biosciences*, *320*, 108304. https://doi.org/10.1016/j.mbs.2019.108304

Nunes, M. R., Palacios, G., Faria, N. R., Sousa Jr, E. C., Pantoja, J. A., Rodrigues, S. G., Carvalho, V. L., Medeiros, D. B., Savji, N., Baele, G., et al. (2014). Air travel is associated with intracontinental spread of dengue virus serotypes 1–3 in Brazil. *PLOS Neglected Tropical Diseases*, *8*(4), e2769.

Pan, Y., Darzi, A., Kabiri, A., Zhao, G., Luo, W., Xiong, C., & Zhang, L. (2020). Quantifying human mobility behaviour changes during the COVID-19 outbreak in the United States. *Scientific Reports*, *10*(1), 1–9.

Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, *52*(1), 175–184.

Remuzzi, A., & Remuzzi, G. (2020). COVID-19 and Italy: What next? *The Lancet*, *395*(10231), 1225–1228. https://doi.org/10.1016/S0140-6736(20)30627-9

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319–392.

Salehinejad, H., Sankar, S., Barfett, J., Colak, E., & Valaee, S. (2018). *Recent Advances in Recurrent Neural Networks* (tech. rep.).

Sedlar, U., Winterbottom, J., Tavcar, B., Sterle, J., Cijan, J., & Volk, M. (2019). Next generation emergency services based on the Pan-European Mobile Emergency Application (PEMEA) Protocol: Leveraging mobile positioning and context information. *Wireless Communications and Mobile Computing*, *2019*.

Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons & Fractals*, *140*, 110212.

Song, C., Shi, X., Bo, Y., Wang, J., Wang, Y., & Huang, D. (2019). Exploring spatiotemporal nonstationary effects of climate factors on hand, foot, and mouth disease using Bayesian spatiotemporally varying coefficients (STVC) model in Sichuan, China. *Science of The Total Environment*, *648*, 550–560.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stoddard, S. T., Forshey, B. M., Morrison, A. C., Paz-Soldan, V. A., Vazquez-Prokopec, G. M., Astete, H., Reiner, R. C., Vilcarromero, S., Elder, J. P., Halsey, E. S., Kochel, T. J., Kitron, U., & Scott, T. W. (2013). House-to-house human movement drives dengue virus transmission. *Proceedings of the National Academy of Sciences*, *110*(3), 994–999. https://doi.org/10.1073/pnas.1213349110

Stoddard, S. T., Morrison, A. C., Vazquez-Prokopec, G. M., Soldan, V. P., Kochel, T. J., Kitron, U., Elder, J. P., & Scott, T. W. (2009). The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl Trop Dis*, *3*(7), e481.

Titus Muurlink, O., Stephenson, P., Islam, M. Z., & Taylor-Robinson, A. W. (2018). Long-term predictors of dengue outbreaks in Bangladesh: A data mining approach. *Infectious Disease Modelling*, *3*, 322–330. https://doi.org/10.1016/j.idm.2018.11.004

Toch, E., Lerner, B., Ben-Zion, E., & Ben-Gal, I. (2019). Analyzing large-scale human mobility data: A survey of machine learning methods and applications. *Knowledge and Information Systems*, *58*(3), 501–523. https://doi.org/10.1007/s10115-018-1186-x

Torres-Signes, A., Frıas, M. P., & Ruiz-Medina, M. D. (2020). Spatiotemporal prediction of COVID–19 mortality and risk assessment. *arXiv preprint arXiv:2008.06344*.

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

WHO. (2019). *World Health Statistics 2019: Monitoring health for the sdgs, sustainable development goals*. [OCLC: 1133205496]. World Health Organization.

Wieczorek, M., Siłka, J., & Woźniak, M. (2020). Neural network powered covid-19 spread forecasting model. *Chaos, Solitons & Fractals*, *140*, 110203.

Worldometer. (2020). Coronavirus Update (Live): 4,654,991 Cases and 309,133 Deaths from COVID-19 Virus Pandemic - Worldometer [Library Catalog: www.worldometers.info]. https://www.worldometers.info/coronavirus/

Wu, Y.-C., Chen, C.-S., & Chan, Y.-J. (2020). The outbreak of COVID-19: An overview. *Journal of the Chinese Medical Association*, *83*(3), 217.

Yang, W., Deng, M., Li, C., & Huang, J. (2020). Spatio-Temporal Patterns of the 2019-nCoV Epidemic at the County Level in Hubei Province, China. *International Journal of Environmental Research and Public Health*, *17*(7), 2563.

Conflict of interest

The authors declare that they have no conflict of interest.

Data availability statement

The `pre-processed data`, the `Python` and `R` code for implementing the proposed model can be made available upon request.

Appendices

A Shiny App

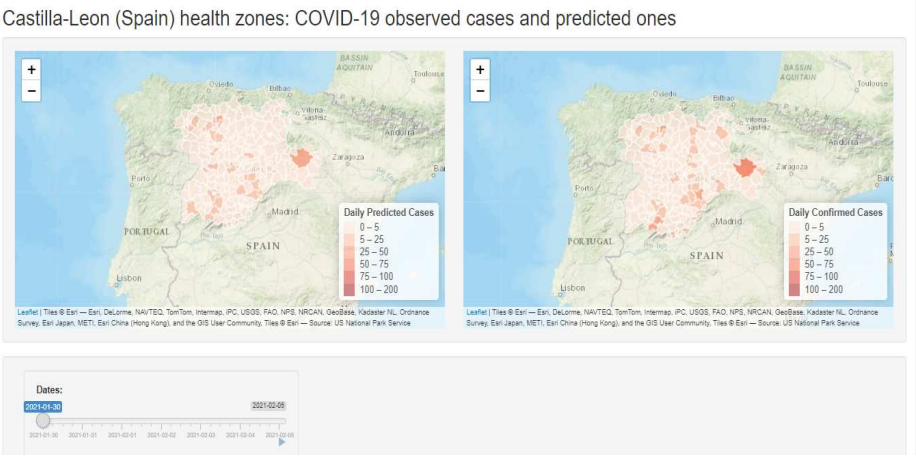

Fig. 13: View of R-Shiny App visualizing observed and predicted COVID-19 cases. See: https://poshan-niraula.shinyapps.io/CYLCovidPrediction/

B LSTM method Parameters

Table 6: Summary of parameters and hyperparameters in the LSTM model

Parameter	**Value**
Number of LSTM layers	3
Hidden Units in LSTM layers	Layer 1: 128 Layer 2: 64 Layer 3: 32
Number of dense layers	1
Activation function of dense layer	Linear
Number of epochs	100
Loss Function	Mean Squared error
Optimizer	ADAM Learning Rate: 0.001 β_1: 0.9 β_2: 0.999
Batch Size	10

C Marginal posterior distributions of covariate coefficients

Table 7: Marginal posterior mean and credible interval of fixed effects

Covariate	**Mean**	**Credible interval**
Monday	-0.182	-0.193, -0.172
Tuesday	-0.207	-0.218, -0.197
Wednesday	-0.187	-0.198, -0.177
Thursday	-0.658	-0.672, -0.644
Friday	-0.326	-0.340, -0.312
Saturday	-0.206	-0.212, -0.192
Sunday	-0.321	-0.336, -0.310
Average cases in neighboring health zones	0.031	0.030, 0.032
Cumulative cases	0.025	0.019, 0.031
Deaths	0.019	0.013, 0.025
Average deaths in neighboring health zones	-0.034	-0.043, -0.025
Daily normal weight (spatial weight)	0.041	0.040, 0.042
Number of people demanding employment	-0.001	-0.001, 0.000
Total registered unemployment	0.001	0.000, 0.002
Number of urban commercial units	0.000	0.000, 0.000
Number of urban industrial units	0.000	0.000, 0.000
Number of urban office units	0.002	-0.001, 0.004

D Residual plots of fitted models and their predictions

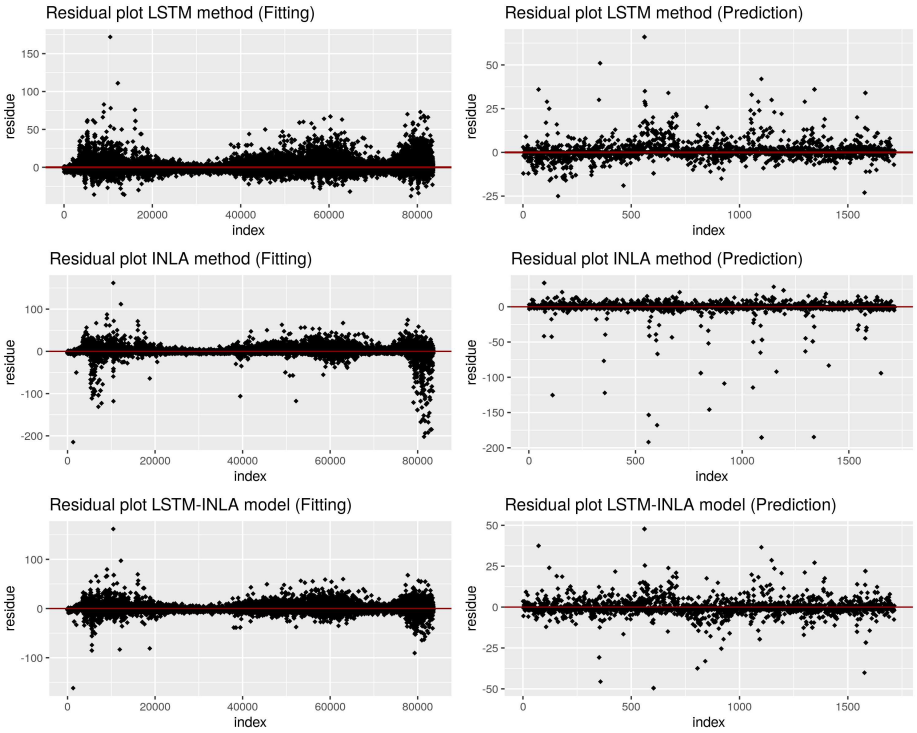

Fig. 14: Residual plot of the fitted models (left) and predictions (right)