# Modeling the outbreak and spread of infectious diseases using a Bayesian machine learning approach

*Dissertation submitted in partial fulfillment of the requirements for the Degree of Master of Science in Geospatial Technologies*

**February 24, 2021**

---


**Poshan Niraula**
*pniraula@uni-muenster.de*

**Supervised by:**
Edzer Pebesma
Institute of Geoinformatics
University of Münster

**Co-supervised by:**
Jorge Mateu
Department of Mathematics
Universitat Jaume I

**Co-supervised by:**
Roberto Henriques
NOVA Information Management School
Universidade Nova de Lisboa


---

UNIVERSITAT JAUME I

NOVA IMS Information Management School

ifgi Institute for Geoinformatics University of Münster

# Declaration of Academic Integrity

I hereby confirm that this thesis on *Modeling the outbreak and spread of infectious diseases using a Bayesian machine learning approach* is solely my own work and that I have used no sources or aids other than the ones stated. All passages in my thesis for which other sources, including electronic media, have been used, be it direct quotes or content references, have been acknowledged as such and the sources cited.

February 24, 2021

---

I agree to have my thesis checked in order to rule out potential similarities with other works and to have my thesis stored in a database for this purpose.

February 24, 2021

---

# Acknowledgements

First of all, I am eternally grateful to **Prof. Jorge Mateu** who has not only guided and motivated me in this thesis but also inspired me to pursue spatial statistics. I find myself privileged to have been given a chance to work so closely with Prof. Mateu. Equally, I would like to thank my supervisor **Prof. Edzer Pebesma** for his guidance throughout this thesis. In addition, I am equally thankful to **Prof. Roberto Henriques** for his valuable remarks. I am truly grateful to have this advisory team.

I would like to thank **Somnath Choudhuri**, for his continuous support and feedback in this work. I would like to thank my friends **David Payares** and **Mateen** especially for their support and motivation throughout the thesis process. Apart from that, we had some great discussions regarding this work in the lab and those discussions really helped me implement and improve the concepts in this thesis.

Among my friends, I would like to mention specially the name of my best friend **David Alsina**, who constantly inspired me to continue my work even at difficult moments. I would like to mention my friends **Janak, Anu and Ganesh** for being wonderful flatmates, keeping me well fed and constantly checking on me.

I would like to pay my highest gratitude to my family back home in Nepal, for their ever-enduring support. It is all due to their wishes, especially my parents, that I have come this far in life. They are the real strength behind every success that came my way throughout my pursuits.

Finally, I pay my respect, love and gratitude to my wife, **Kranti**, who patiently took care of our son **Adwait** back home so that I could focus on this work. Adwait's gifted smile have inspired me in his own way to successfully complete this work. Thank you **Kranti** for your sacrifices, patience and tolerance, and for keeping me sane over the past few months. But most of all, thank you for being a constant support system. I owe you everything.

*This work is dedicated to*
*everyone who lost their lives to COVID-19*
*and to those fighting against it.*

# Abstract

The modeling of infectious diseases and their predictions on space and time is very important as it helps in devising the policies for preventive measures. These predictions should be generated from a probabilistic model to provide the uncertainties and thus the confidence. The phenomenon of spread of infectious diseases is so complex that there are lots of uncertainties in the data and in the process itself. Machine learning methods like neural networks are useful in modeling this complex problem, however, these approaches lack handling of uncertainties. Similarly, it is seen in literature that a combined approach of neural networks and Bayesian inferences have not been explored much. Thus to fill these gaps this thesis aims to develop a combined model containing neural network method and Bayesian inference for modeling and predicting the number of cases of infectious diseases in areal units such as municipalities or health-zones.

To introduce the impact of human movement on the spread of infectious disease, the movement data has been used combined with the daily infection data to form a spatial factor and used as a covariate in this study. In addition to this, the spatial correlation due to spatial neighborhood as well as the mobility is taken into account in the model along with the temporal dependencies.

The model was evaluated on the COVID-19 dataset for 245 health-zones of the autonomous community of Castilla-Leon, Spain. The results show that the model is generally able to predict the number of cases of infectious diseases with good accuracy. Similarly, the mobility factor was also found to have an influence on the model. However, the flexibility of the model still needs to be evaluated by applying the model to different scenarios.

*Keywords:* Bayesian Inference, Human movement, Infectious diseases, Neural networks

# Contents

# List of Tables

# List of Figures

# List of Acronyms

| | |
|---|---|
| ANN | Artificial Neural Network |
| BYM | Besag York Mollie |
| COVID-19 | Corona Virus Disease 2019 |
| CPO | Conditional Predictive Ordinate |
| DIC | Deviance Information Criterion |
| GLM | Generalized Linear Model |
| GMRF | Gaussian Markov Random Fields |
| INLA | Integrated Nested Laplace Approximation |
| LSTM | Long Short Term Memory |
| MCMC | Markov Chain Monte Carlo |
| RMSE | Root Mean Square Error |
| RNN | Recurrent Neural Network |
| WAIC | Watanabe-Akaike Information Criterion |
| WHO | World Health Organization |

# Chapter 1

# Introduction

## 1.1 Context and motivation

Infectious diseases are the main cause of health hazards in the world (WHO, 2019). Various outbreaks of these diseases have occurred throughout human history. Dengue and Malaria caused by the bite of mosquitoes infect around 4 million and 228 million people per year respectively and is widely spread in underdeveloped countries in African and Asian countries (Ak et al., 2018; Organization et al., 2019). The highly infectious and fatal Ebola outbreak occurred during 2014-2016 in Western Africa, which infected 28500 and killed around 11000 people. Severe Acute Respiratory Syndrome (SARS) outbroke in China in 2003 affecting 26 countries and in 2012, Middle East Respiratory Syndrome affected 27 countries infecting overall 2494 people (WHO, 2019). From December 2019, there has been an outbreak of the novel Coronavirus disease (COVID-19), from China, Wuhan, and has infected more than 90 million people and has taken the lives of more than 2 million people (Worldometer, n.d.; Wu et al., 2020) as of January 2021. To contain the spread of this disease, governments around the world are making various efforts which include social distancing, travel restrictions, and city-level or, even nation-wide lockdown measures. These precautions, although effective in controlling the spread of the disease, have impacted the daily lives of people, the social behaviors and have a considerable impact on the global supply chain (Jones et al., 2008). Infectious diseases exhibit certain patterns and can be predicted based on socio-economic, environmental, and ecological factors. Prediction of these infections is important for the government and health workers to plan for controlling the rate of infection (Remuzzi & Remuzzi, 2020). More importantly, spatio-temporal analysis and prediction of the dynamics of the disease is very important for prioritizing the actions (Ak et al., 2018; Yang et al., 2020). Similarly, for the infectious diseases that could turn into pandemics, the control mechanism is very important along with a very good spatial and temporal prediction (Zhou et al., 2020).

The introduction of any infectious disease into a new area is generally stimulated by human movements. There are various examples where a region-specific disease in the world has been imported to a new region due to international travels (Nunes et al., 2014; Stoddard et al., 2009). Apart from this, the spread of the diseases in an area locally is also very relevant to the human movement patterns within the area (Stoddard et al., 2013). In the case of COVID-19, according to

the study done by (Gross et al., 2020), the infection of COVID-19 were found to be highly correlated with the propagation of the disease. With the development of advanced technologies for the precise location and the concepts of sharing the location information (anonymously), the introduction of the human mobility dimension into the epidemiological studies has been easier. Various works have been done in the modeling of various kinds of infectious diseases considering the human movement (M. U. G. Kraemer et al., 2020; Wesolowski et al., 2015) at different spatial scales and also on the resource-poor regions with no available mobility data set (M. Kraemer 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. It is also reported that during pandemics of infectious diseases, the behavior of human mobility changes (Pan et al., 2020) compared to that of normal times which makes the problem more complex and difficult to analyze. Deep learning methods have proved to be suitable methods for the modeling of these complex problems. In the studies (Akhtar et al., 2019; Kapoor et al., 2020; Wieczorek et al., 2020), neural network methods have been employed along with human mobility to model the spread of infectious diseases. Although these methods have performed well, they are unable to provide the uncertainties in the predictions which are important to consider in this case. Predictions with uncertainties give confidence to the users of the results from the models (Beale & Lennon, 2012), so it is important to have uncertainties in the predictions. To incorporate the uncertainties in the neural network-based methods, Bayesian Neural Network has been developed (Kononenko, 1989). These methods have been applied in various spatio-temporal contexts as well (McDermott & Wikle, 2019). But limited research works have been conducted in the field of modeling and understanding the dynamics of infectious diseases using neural network with Bayesian inference. These methods rely on the hidden stage of the neural networks to learn from the data and are unable to explicitly account for the spatial and spatio-temporal randomness.

These limitations are the main motivation of the current thesis. A particular focus of this work lies in the use of the combination of deep learning methods together with Bayesian inference to model and predict the spread and outbreak of infectious diseases with uncertainties. In the present, study the human mobility data along with socio-demographic variables will be incorporated in the combined model to predict the dynamics of COVID-19 pandemic. Likewise, the thesis attempts to analyze the importance of human mobility in modeling the dynamics of infectious diseases.

## 1.2 Related Works

The modeling of infectious diseases has generally focused on compartmental models, where the population is divided into various compartments. Susceptible Infected Removed (SIR) models developed by (Kermack & McKendrick, 1927) have been widely used and have been modified to other forms (Brauer, 2008). These forms of compartmental models are used in the modeling of different types of diseases including SARS, Ebola, HIV, and COVID-19. Similarly, these models are also applied to various vector-borne diseases like Malaria and Dengue. Table

1.1 provides a summarized view of the use of compartmental models in modeling different diseases.

Table 1.1: Summary of related works done in compartmental modeling

| References | Disease | Model type |
|---|---|---|
| (Chowell et al., 2003) | SARS | Susceptible Exposed Infectious Diagnosed Recovered |
| (Rivers et al., 2014) | Ebola | Susceptible Exposed Infectious Hospitalized Funeral Recovered/Removed |
| (Chitnis et al., 2008) | Malaria | Susceptible Exposed Infectious Recovered |
| (Pandey et al., 2020) | COVID-19 | Susceptible Exposed Infectious Recovered |
| (Giordano et al., 2020) | COVID-19 | Susceptible Infected Diagnosed Ailing Recognized Threatened Healed Extinct |

As spatio-temporal predictions help in understanding the spread of the disease better to identify the regions of high risks, various works can be found on the spatio-temporal modeling of the diseases. Among them, generalized linear models with the addition of spatial effect of nearby places and/or temporal effects from past events are found to be used often and have proved to be useful in predicting as well (Cabrera & Taylor, 2019; Giuliani et al., 2020; Guo et al., 2017). For example, (Giuliani et al., 2020) have used the generalized linear models to predict the COVID-19 infections in the regions of Italy and found the spatial interactions of nearby places to have a high influence on the modeling, which shows the importance of accounting for the spatial effects explicitly. Along with these, the Bayesian modeling methods have also been applied in the case of infectious disease modeling in various works (Aswi et al., 2019; Song et al., 2019) and they have been useful in the prediction and more importantly able to predict with associated uncertainties (Gelman et al., 2013).

Various machine learning methods have been applied in the forecast and modeling of the diseases (Ak et al., 2018; Anno et al., 2019; Titus Muurlink et al., 2018). In particular neural network and deep learning methods are explored as they can model the diseases' dynamics in space and time with good accuracy (Kapoor et al., 2020; Wieczorek et al., 2020). Bayesian Neural Network (Kononenko, 1989) applied in the modeling field have been able to perform better than the Neural networks (Dhamodharavadhani et al., 2020). Table 1.2 shows some of the works that have been able to perform spatio-temporal modeling of the diseases with the neural network methods. (Cabras, 2020) presented a method of combining the neural network method with the Bayesian Inference to model the COVID-19 infections on the autonomous regions of Spain. In this study, the author has not considered other spatial variables and spatial dependencies.

On the other hand, human mobility has been proved to be an important factor in the transmission of diseases. Thus, various studies have incorporated the human movement factors into the modeling of the spread of the diseases (M. Kraemer et al., 2019; Massaro et al., 2019; Mukhtar et al., 2020). 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 as one of the reasons for the transmission. In the case of COVID-19, it is also seen that the measures related to human movements like travel restrictions and social distancing have been effective in containing the diseases (M. U. G. Kraemer et al., 2020) (Fang et al., 2020). Availability of technologies like cell phone tower positioning records or global navigation satellite systems or Wifi positioning systems have made it easier to study the mobility (Gonzalez et al., 2008) (Toch et al., 2019)).

Table 1.2: Some of related works done with the use of Deep Learning methods

| References | Disease | Spatial Resolution | Method |
|---|---|---|---|
| (Akhtar et al., 2019) | Zika | Country wise | Dynamic Neural Network |
| (Wieczorek et al., 2020) | COVID-19 | Country and region wise | Neural Networks |
| (Kapoor et al., 2020) | COVID-19 | County wise | Graph Neural Network |
| (Dhamodharavadhani et al., 2020) | COVID-19 | Country wide | Probabilistic Neural Network |
| (Cabras, 2020) | COVID-19 | Region wise | Neural Network and Bayesian Inference |

## 1.3 Aim and Objectives

The principal aim of this research is to analyze and model the spatio-temporal dynamics of infectious diseases considering the influence of mobility. This study seeks to propose an approach to infectious diseases modeling with the use of a neural network method complemented by Bayesian inference.
The research questions that guide this study are:

1. How can the spread of infectious diseases in space, and their outbreaks in time be modeled and predicted using neural network methods?
2. How can uncertainties be introduced in the prediction of infectious diseases?
3. How can mobility be introduced to quantify its influence in the modeling of infectious disease?

## 1.4 Thesis Outline

The thesis is organised as follows: Chapter 2 provides a theoretical background of the methodological concepts used in this thesis. In Chapter 3, the methodology of the proposed model is described into detail. Chapter 4 presents the data, experiment design and implementation of the model with COVID-19 data. In Chapter 5, the results of the experiments are interpreted and discussed. Finally, Chapter 6 ends with the conclusion of this thesis.

# Chapter 2

# Theoretical Background

This chapter is meant to serve as the theoretical foundation of the concepts used in the thesis. The first section presents a brief explanation about the recurrent neural networks, and the background concepts of Long Short Term Memory (LSTM) method in detail. The second section starts with some discussions on Bayesian inferences, hierarchical models, and Integrated Nested Laplace Approximation (INLA) in detail.

## 2.1 Artificial Neural Networks

Artificial Neural Networks are types of machine learning techniques inspired by the functioning of human brains and work on the principle of parallel processing. They consist of interconnected processors called neurons which learn from the input data and optimize the output. Deep learning refers to the deeper networks with multiple layers of the neurons, thus providing better learning and prediction capabilities (Pascanu et al., 2014). Recurrent neural networks are the special kind of neural networks that have been designed to learn from sequential or time-series data. The major division of the deep learning methods includes deep neural networks, convolution neural network, and recurrent neural networks. These methods have a wide range of application areas which includes computer vision, natural language, time series prediction, etc (Medsker & Jain, 1999).

Since RNN is the algorithm chosen for this study, the following subsections deal with the detailed architecture of RNN.

### 2.1.1 Recurrent Neural Networks

RNNs are a subset of supervised machine learning models made up of one or more feedback loops of artificial neurons which are recurrent over time or sequence (Fausett, 1994; Haykin & Network, 2004). RNN has a stack of non-linear units that can learn even long-term dependencies of time series data (Bengio et al., 1994). In RNN, 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 it to learn from the past data and thus long term dependencies are learnt(Mikolov et al., 2014). This makes RNN an excellent choice for learning and predicting time-dependent data. In the next section, the basic architecture of RNN is explained.

**Architecture of Recurrent Neural Network**

A basic RNN consists of three layers: input layers, recurrent hidden layers, and output layers. A simple architecture of Recurrent Neural Network is shown in figure 2.1 reprinted from (Salehinejad et al., 2018) here input layer has N input units.

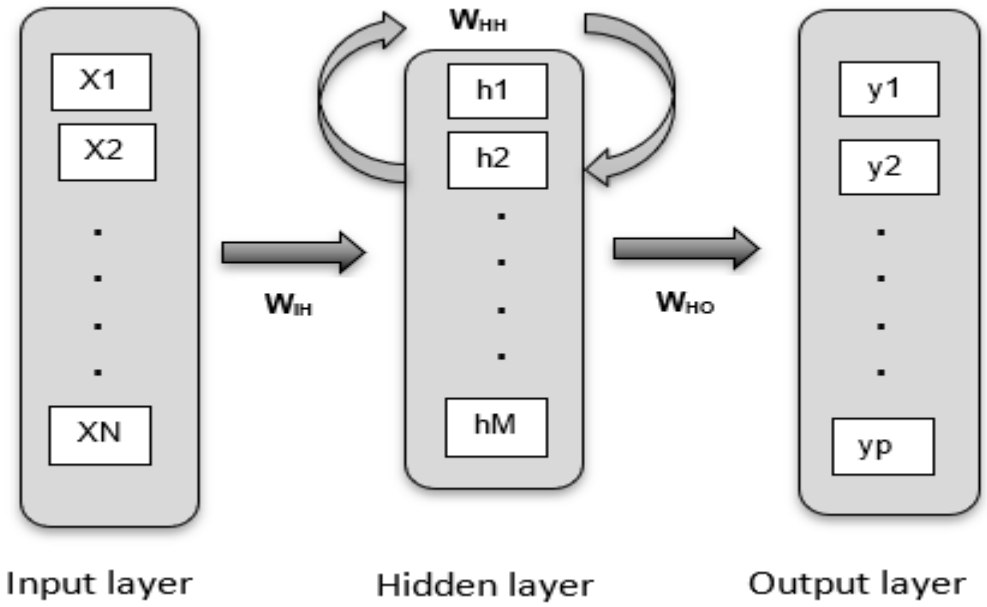



Figure 2.1: Architecture of Recurrent Neural Network
(Salehinejad et al., 2018)

The input layer is a sequence of vectors through time $\{x_1 \ldots . x_{t-1}, x_t, x_{t+1} \ldots . x_T\}$ where every $x_t = (x_1, x_2, \ldots \ldots, x_N)$. These input features are connected to hidden units of the hidden layers, these connections are dependent on a weight $W_{IH}$. The hidden layer in the example architecture contains M hidden units $h_t = (h_1, h_2, \ldots \ldots, h_M)$ which are interconnected through time. For each time of study t, a layer of M hidden units learns from the data and this layer is connected to the hidden layer with M hidden units in the next time t+1 and so on. The state of a hidden layer can be defined as:

$$\begin{gathered} \mathrm{h_t} = \mathrm{f_H}\left(\mathrm{O_t}\right) \\ where, \mathrm{O_t} = \mathrm{W_{IH}X_t} + \mathrm{W_{HH}h_t} - 1 + \mathrm{b_h} \end{gathered} \tag{2.1}$$

$f_h(.)$ is activation function for the hidden layer, and $b_h$ is bias vector of the hidden units.

The output layer is connected to the hidden units and are determined by the weights $W_{HO}$. The output layer contains P units. Here each output unit is given by:

$$\mathrm{y_t} = \mathrm{f_o}\left(\mathrm{W_{HO}h_t} + \mathrm{b_o}\right) \tag{2.2}$$

where, $f_o$ is activation function and $b_o$ bias vector of output layer.

### Activation Functions

In the learning process of the neural networks, all operations are linear except for the activation functions, thus it provides the non-linearity in the learning of a neural network and helps in solving complex problems (Sutskever et al., 2011). Different predefined activation functions are available and are selected based on the requirements. Activation functions used for classification problems are Sigmoid or SoftMax whereas Tanh (Hyperbolic Tan), ReLU (Rectified Linear Unit) are useful in regression (Sharma et al., 2020). Sigmoid is used in the gating of the LSTM since it outputs the values in the range of 0-1, whereas tanh is the activation function used in the conditions when gradient is less likely to vanish but also converges faster than the sigmoid. (Duch & Jankowski, 1999)

### Loss Function

Loss functions are the functions that help in evaluating the performance of a neural network. The comparison of the output from the network and the actual value is performed using the selection of loss functions as needed. In recurrent Neural Networks, if the output value for a timestamp t, $y_t$ and the actual value is $z_t$, the loss is the sum of loss for all the timestamps (Sutskever et al., 2011) as

$$\mathrm{L(y, z)} = \sum_{t=1}^{T} L_t \left(y_t, z_t\right) \tag{2.3}$$

### Gradient Descent

The training of recurrent neural networks involves the minimization of the loss, which is achieved by the optimization process. In the learning method of a recurrent neural network, gradient descent is one of the most popular and simple methods of optimization. These methods are based on the differential equation, where the derivatives of the error function are computed with respect to the weight. To minimize the loss in these methods, the weights assigned to each layer are adjusted proportionally to the derivatives. (Bengio et al., 1994)

In RNN, gradient descent through all the timestamps are applied which is called backpropagation through time, that unfolds the network in time and propagates error signals backward through time (Werbos, 1990). But with this, the problem of vanishing gradient may arise which is a problem when the gradient magnitudes are exponentially shrinking, and the RNN cannot learn from the long-range temporal dependencies (Bengio et al., 1994).

### Long Short Term Memory

Learning with the recurrent neural networks over an extended period via backpropagation usually falls into the problems like gradient vanishing while learning the long-range dependencies in temporal data (Mikolov et al., 2014). To overcome these problems, the method of LSTM has been proposed by (Hochreiter & Schmidhuber, 1997). In this approach, the structure of the hidden units is changed to memory cells, whose input and outputs are controlled by gates. The

gates control the flow of the information to hidden units also preserving the extracted features from previous time (Hochreiter & Schmidhuber, 1997; Le et al., 2015).

The gates are logistic units with their own learned weights on the connections with the input units and memory cells from the previous timestamp. There are three types of gates: the forget gate which learns weights that control the rate at which the value stored in the memory cell decays, the input gate, and the output gate. Figure 2.2 shows a simple LSTM unit, here the LSTM cells receive the activation signals from the previous memory cells state c(t-1), the previous activation h(t-1) and the input data x(t). Here the input gate is defined as

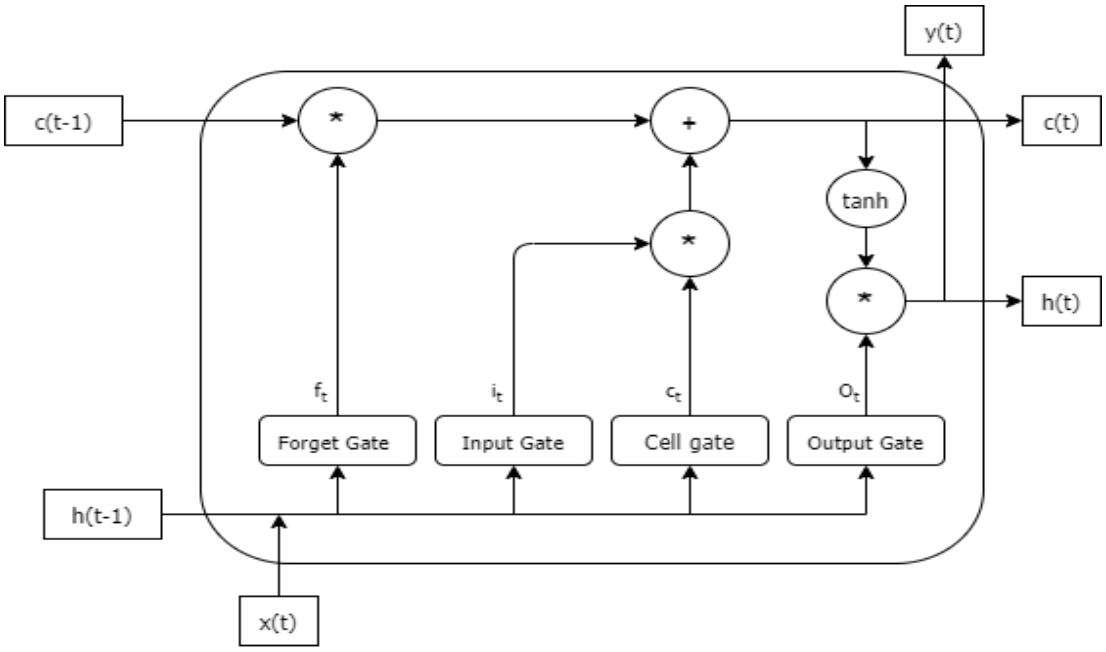



Figure 2.2: Architecture of an LSTM unit

$$i_t = \sigma\left(W_{ii}x_t + W_{Hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \tag{2.4}$$

where, $W_{ii}$ is the weight matrix from input layer to the input gate, $W_{Hi}$ is the matrix from hidden state to the input gate, $W_{Ci}$ is the matrix from the cell activation to the input gate and $b_i$ is the bias of the input gate. The forget gate is defined as

$$f_t = \sigma\left(W_{if}x_t + W_{Hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \tag{2.5}$$

where, $W_{if}$ is the weight matrix from input layer to the forget gate, $W_{Hi}$ is the matrix from hidden state to the forget gate, $W_{Ci}$ is the matrix from the cell activation to the input gate and $b_i$ is the bias of the forget gate. The cell gate is defined as

$$c_t = i_t \tanh\left(W_{ic}x_t + W_{Hc}h_{t-1} + b_c\right) + f_t c_{t-1} \tag{2.6}$$

where, $W_{ic}$ is the weight matrix from input layer to the cell gate, $W_{Hc}$ is the matrix from hidden state to the cell gate and $b_c$ is the bias of the cell gate. Similarly, the output gate is computed as

$$o_t = \sigma\left(W_{io}x_t + W_{Ho}h_{t-1} + W_{co}c_{t-1} + b_o\right) \tag{2.7}$$

where, $W_{io}$ is the weight matrix from ouput layer to the output gate, $W_{Ho}$ is the matrix from hidden state to the ouput gate, $W_{Cf}$ is the matrix from the cell activation to the output gate and $b_f$ is the bias of the output gate. The hidden state is defined as

$$h_t = O_t \tanh(c_t) \tag{2.8}$$

With the addition of these gates the LSTM is able to survive the vanishing gradient problem along with being able to learn the long term dependencies (Salehinejad et al., 2018).

### Optimization

To enhance the gradient descent, optimization are added to the neural networks training. Several optimization algorithms are available that makes the training process faster. Some of the notable ones are Adaptive Moment Estimation, Adaptive Gradient, Nesterov Accelerated Gradient, Root Mean Square Propagation and Stochastic Gradient Descent with momentum (Bengio et al., 2013). In this case Adaptive Moment Estimation optimization is used thus, the next section provides a brief description of the Adaptive Moment Estimation optimization.

### Adaptive Moment Estimation

ADAM is an efficient stochastic optimization method that computes individual adaptive learning rates for different parameters from the first and second moment gradient descent calculations. This method combines the positives from two other methods Adaptive Gradient and RMSProp This method uses exponential moving averages of gradient and squared gradients and the hyperparameters $\beta_1$, $\beta_2$ control the decay rate of these moving averages (Kingma & Lei Ba, 2017). ADAM optimizers converge a neural network very quickly thus is very useful in the multilayered neural networks.

### Parameters and Hyperparameters

The weights and biases and other variables that are derived through the training of a neural network are parameters, whereas the variables which remain are predefined for the training are hyperparameters. The hyperparameters need tuning a lot of experimentations based on the type of data and output. Some of the hyperparameters are explained in next sections

### No of Epochs

The total number of cycles of forward and back propagation the data goes through in the neural network. In the case of the neural network training in each epoch the weights and the bias are updated, thus increasing the accuracy of the model. Running a model for many epoches may sometimes cause the model to overfit thus needs the early stopping.

### Batch Size

The data size that the neural network model takes into consideration at once step is batch size. Batch size are set based on the size of the neural networks.

#### Learning Rate

The learning rate refers to the no of steps a neural network takes to converge.Learning rates are to be chosen carefully as higher learning rate could result in divergence while lower learning rate could cause model to take much time to converge.

## 2.2 Bayesian Inference

Bayesian inference is the method of statistical inference that considers the Bayes theorem to update the probability of a hypothesis. Thus they are used to deduce the probability distribution of data using the Bayes theorem. The bayesian analysis allows the incorporation of the subjective information from outside the available dataset and they can provide the conclusion regarding the parameters in terms of probability statements. (Gelman et al., 2013)

In the first sub-section, the concepts of probability distributions and Bayesian inference is explained in brief and in the second section, the spatial analysis using Bayesian methods is presented.

### 2.2.1 Probability Distribution

Probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes in an experiment.The distribution is usually described in the form of probability mass function or probability density function.

#### Poisson Distribution

Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space. The probability mass function for Poisson distribution is given by

$$f(k;\lambda) = \mathrm{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{2.9}$$

Here k is number of occurance and $\lambda$ is a positive real number which is equal to the expected value of X and the variance.

$$\lambda = E(X) = Var(X) \tag{2.10}$$

#### Negative Binomial Distribution

The negative binomial distribution is a discrete probability distribution that models the number of successes in a sequence of independent and identically distributed Bernoulli trials. The probability mass function for the negative binomial distribution is given by

$$f(k;r,p) \equiv \Pr(X = k) = \begin{pmatrix} k+r-1 \\ r-1 \end{pmatrix} (1-p)^k p^r \tag{2.11}$$

here r is number of successes, k is the number of failures and p is probability of success. The mean of negative binomial distribution is given by $\frac{pr}{1-p}$ and the variance $\mu(1+\frac{\mu}{r})$. Thus it tends to become poisson distribution when r $\to \infty$.

Negative binomial distribution can be used as an alternative to the poisson distribution as it allows different mean and variance, thus in modeling the disease transmission, where the variance is high, negative binomial distribution is used (Lloyd-Smith et al., 2005).

### 2.2.2 Bayes' Theorem

Bayes´theorem is a tool that allows the use of prior knowledge or belief regarding any event to compute the probability of that event. The Bayes theorem can be stated as

$$P(\Theta \mid data) = \frac{P(data \mid \Theta) \times P(\Theta)}{P(data)} \tag{2.12}$$

where $\Theta$ is the parameter of the distribution, $P(\Theta \mid data)$ is the posterior that defines a probability distribution of parameters that fits the data, $P(data \mid \Theta)$ is the likelihood which is the probability that the data could be generated by the model with parameters $\Theta$ and $P(\Theta)$ is the prior information we have regarding the parameters $\Theta$.

$P(data)$ is the overall probability of data also referred as the marginal likelihood which is sometimes difficult or even impossible to compute because of the complexity. This complexity of the computations is usually addressed by restricting the models to conjugate priors, or finding the numerical solutions or generation of large number of combinations of representative parameters from the posterior distribution, these methods are known as Markov Chain Monte Carlo (MCMC). The development of these computing methods have boosted Bayesian inferences towards practical use cases.

### 2.2.3 Priors and Conjugate prior families

Priors have very important role in the Bayesian analysis as they help in defining the subjective information regarding the data before actually looking at the data. Priors are provided as the probability distribution, which are usually specified based on information accumulated from the past studies or from the opinions of the subject-area experts. While choosing the prior distribution from certain families the computation of the posterior distribution is easier for some distributions than others. Generally for the ease of computation the selection of the priors is done so that it is conjugate with the likelihood which would generate the posterior distribution in the same distribution as the prior.

For example for a counting variable $X$ the likelihood distribution is generally considered Poisson i.e.

$$\mathrm{P}(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}, x \in \{0, 1, 2, ...\}, \theta > 0 \tag{2.13}$$

if the prior is considered as the Gaussian distribution with parameters $\alpha$ and

$\beta$ i.e.

$$\mathrm{P}(\theta) = \frac{\theta^{\alpha-1}e^{-\frac{\theta}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}, \theta > 0, \alpha > 0, \beta > 0 \tag{2.14}$$

the posterior distribution is proportional to a gamma distribution with parameters $\alpha' = x + \alpha$ and $\beta' = (1 + \frac{1}{\beta})^{-1}$ as from Bayes theorem

$$\begin{aligned} &\mathrm{P}(\theta|x) \propto \mathrm{P}(x|\theta)\,\mathrm{P}(\theta) \\ &\propto (e^{-\theta}\theta^{x})(\theta^{\alpha-1}e^{-\frac{\theta}{\beta}}) \\ &= \theta^{x+\alpha-1}e^{-\theta(1+\frac{1}{\beta})} \end{aligned} \tag{2.15}$$

Apart from this conjugate combination of Poisson and gamma there are several other conjugate combinations like Bernoulli- beta, normal-inverse gamma etc. (Carlin & Louis, 2008)

### 2.2.4 Markov Chain Monte Carlo and Integrated Nested Laplace Approximation

MCMC is a simulation based method for the approximation of the marginal likelihood which combines the two methods Markov Chain and Monte Carlo, allowing random sampling of high dimensional probability distribution. There are various MCMC based algorithms used in Bayesian inference some of them are Gibbs sampling algorithm and Metropolis-Hastings algorithm.

INLA is method specially designed for Latent Gaussian variables(Rue et al., 2009).INLA is an anlytical approach using laplace approximations. An R-package for performing Bayesian inference using INLA package is available with the name R-INLA (Martino & Rue, 2010). The functions in this package provide a simple and easy way of performing bayesian inferences for spatial dataset allowing addition of covariates as well as the spatial and temporal interactions. Spatial analysis consists of the models that considers the concept the observations that are closer are likely to show similar values (Tobler, 1970). This is often referred as Spatial auto-correlation which is a systematic variance of a variable in a space (Haining, 2001). Spatial models account for the spatial autocorrelation to separate the general trend from the covariates with the spatial variation. Spatial modeling is divided into three major areas of study on the basis of the data and type of problem namely areal data, geostatistics and point patterns (Cressie, 2015). In the following section the concepts related for areal data are described.

#### Areal data

Areal data are the type of data which are observed or aggregated within a given boundary, these data are also known as lattice data. These boundaries are often the administrative boundaries or arbitrary grids. Some example of lattice data are the number of Covid-19 cases in a given district or county, the total number of trees within a defined grid.

#### Adjacency Matrix and Spatial Neighborhood

While performing spatial analysis on areal data the adjacency matrices play a vital role as they define the dependence of a region to the other nearby regions based on the shared boundary. This is an important step because it ensures that the residuals do not contain any spatial pattern (Bivand et al., 2008). The adjacency matrices are represented by matrix with non-zero entries at the intersection of rows and columns of neighboring areas (Gómez-Rubio, 2020).

### 2.2.5 Bayesian Disease Mapping

modeling the geographical distribution of infectious diseases have been very important task in the history of epidemiology (Wakefield, 2007). For a case of incidence of disease in an area i if $E_i$ be the expected number of people in risk and $y_i$ is the number of cases in the region i, then the no of cases are generally Poisson distributed with $E_i$ mean. i.e.

$$y_i|\theta_i \sim Poisson(E_i.\theta_i) \tag{2.16}$$

where $\theta_i$ represents the true area specific relative risk (Bernardinelli et al., 1995). Various Bayesian hierarchical models for estimating these $\theta_i$ over space have been proposed where the underlying random effects depend on the neighborhood structures (Bernadinelli et al., 1997).

A general model formulation by assuming the log risk $\eta_i$ which is given as

$$\eta_i = log(\theta_i) = \mu + z_i^T\beta + b_i \tag{2.17}$$

where $\mu$ denotes overall risk level, $z_i^T$ are set of covariates with the corresponding regression parameters $\beta$ and $b_i$ the random effects. (Riebler et al., 2016)

#### Besag and Besag York Mollie models

Besag model uses the approach of modeling the spatial correlation as an intrinsic Gaussian Markov Random Field (GMRF). The conditional distribution for $b_i$ is given by

$$b_i \mid \mathbf{b}_{-i}, \tau_b \sim \mathcal{N}\left(\frac{1}{n_{\delta i}}\sum_{j\in\delta i} b_j, \frac{1}{n_{\delta i}\tau_b}\right) \tag{2.18}$$

where $\tau_b$ is precision parameter and $b_{-i} = (b_1, ....b_{i-1}, b_{i+1}, ...b_n)$, $\delta i$ are the neighbours of region i and $n_{\delta i}$ the number of neighbours. (Besag et al., 1991)

The Besag model only assumes the spatially structured component through the neighborhood, along with including the random error or pure overdispersion in the area i as spatial correlation, which may lead to error in parameter estimation (Breslow et al., 1998). Thus in the BYM model this issue is addressed by decomposing the spatial effect $b$ into the unstructured and structured components. $b = u + v$, where $v \sim \mathcal{N}(0, \tau_v^{-1}I)$ accounts for the pure overdispersion and $u$ is the structured component from the besag model.

### 2.2.6 Model Evaluation

The evaluation of the performance of a regression model is generally done with the help of Root Mean Squared Error (RMSE) value and are standard statistical metric to measure the performance of models in the field of geo-sciences (Chai & Draxler, 2014). The Root Mean Squared Error value can be computed as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{2.19}$$

where, n is the number of the observations and $e_i$ are the error on each observations i = 1,2,..n

For comparison of the Bayesian models Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), which is a Bayesian model comparison criterion, are used. The DIC values are represented as

$$DIC = goodness\ of\ \ fit + complexity = D(\overline{\theta}) + 2p_D \tag{2.20}$$

where $D(\overline{\theta})$ is the deviance evaluated at the posterior mean of the parameters and $p_D$ denotes the effective number of parameters and it measures the complexity of the model (Spiegelhalter et al., 2002). When the model is true, $D(\overline{\theta})$ should be approximately equal to the effective degrees of freedom, $n - p_D$. One drawback of DIC is that it may underpenalize complex models with many random effects.

An alternative is the Watanabe Akaike information criterion (WAIC) which follows a more strict Bayesian approach to construct a criterion (Watanabe & Opper, 2010). Like DIC, WAIC estimates the effective number of parameters to adjust over-fitting. $pWAIC$ is similar to $p_D$ in the original DIC. (Gelman et al., 2014) scales the WAIC of (Watanabe & Opper, 2010) by a factor of 2 so that it is comparable to DIC.

Similarly, the conditional predictive ordinate (CPO) (Pettit, 1990), which expresses the posterior probability of observing the value (or set of values) of $y_i$ when the model is fitted to all data except $y_i$.

$$CPO_i = \pi\left(y_i^{obs} \mid y_{-i}\right) \tag{2.21}$$

Here, $y_{-i}$ denotes the observations $y$ with the i-th component removed. This facilitates computation of the cross-validated log-score (Gneiting & Raftery, 2007) for model choice $(-\left(mean\left(log(cpo)\right)\right))$.

# Chapter 3

# A Bayesian LSTM Model

This chapter presents the proposed model and describes the methods used in this study. These methods are based on the theories explained in Chapter 2. The structure of this chapter is as follows, the first section gives a brief overview of the model, the second section covers the input data to the model and the third and fourth section describes the LSTM model and the Bayesian Inferences.

## 3.1 Overview

The LSTM Bayesian Model aims to model the number of infections of infectious disease on an areal unit such as a municipality, province, health-zone, etc. based on the spatial covariates, the temporal trends, and the mobility matrices comprising all the mobility from and within each areal units in the study area. With this model, it is possible to predict the number of infections in the future in an areal unit given the spatial covariates and the mobility data. The model assumes that the temporal scale of data is uniform and the spatial extents are irregular lattices.

Figure 3.1 shows the overview of the model used. The input to the LSTM model are the cases of infectious diseases, which gives a prediction. The combination of the mobility data and the spatial variables is done to create spatial weights. These weights and the predictions from the LSTM model are the inputs to the Bayesian model whereas to model the spatial correlation, the neighborhood structures are defined based on the spatial characteristics and the mobility matrices. This model can be summarized as:

If $Y_{ti} \in 0, 1, 2, 3...$ be random variable representing the number of cases of infectious disease in an area i = 1,2... m at a time t = 1,2....T, this work is focused on the computation of

$$P(Y_{ti} = y|F_{ti}, D) \tag{3.1}$$

The value $F_{ti}$ is the *evolution* of the data until time t (Cabras, 2020) which is computed by the LSTM and finally predicted no of infections are computed with the help of Bayesian inference which is conditioned on the predictions of the LSTM and other covariate information such as spatial weights, D.

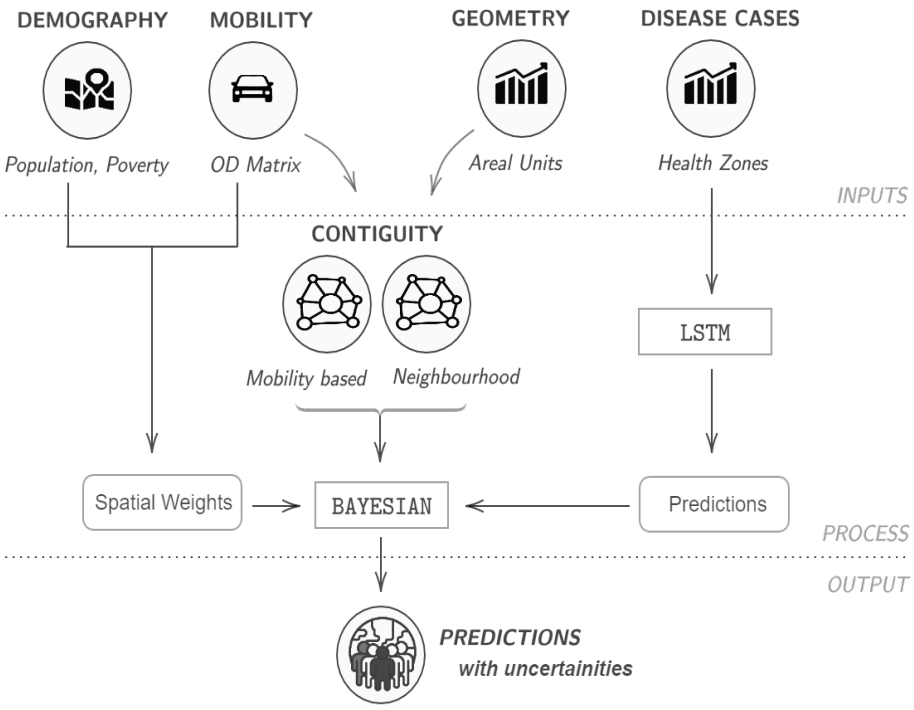



Figure 3.1: Overview of the Bayesian LSTM Model

## 3.2 Model Input

The inputs to the model are the daily number of infection cases in areal units, the spatial variables including the socio-demographical data, daily mobility matrices, and the neighborhood structure of the study area. The details of the input data required by the model are defined in the next sub-sections.

### 3.2.1 Sequential to Supervised Conversion

In performing the time series based analysis, the time series must be converted to a supervised problem i.e. the sequences should be converted to input output pair. Shifting of the sequential data is done to achieve this step (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.

### 3.2.2 Spatial Weights

Considering mobility from all other region $j = 1, 2, ..m$ into a region $i$ as the factor for the importing the infections of a disease into the region $i$, spatial weights are computed. This weight can be interpreted as the possibility of a moving person to import the infection of the disease into the region $i$ from all the other regions.

This spatial weight for a region $i$ for a day $t$, $W_{i,t}$, can be computed as

$$\mathbf{W}_{i,t} = \sum_{j=1}^{n} \sum_{t=t-1}^{\Delta t} (m_{j,i,t} * \frac{I_{j,t}}{P_j}) \tag{3.2}$$

where, $m_{j,i,t}$ is the mobility from all regions j to i on day t,
$I_{j,t}$ is the no of cases of infection on region $j$ at time $t$,
$P_j$ is the total population of the region $j$.

A time lag $\Delta t$ 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 from past days in all other regions of study area.

### 3.2.3 Neighborhood Structures

As discussed in section 2.2.5, in spatial analysis the neighborhood structures are key to accounting for spatial correlation. In this model, these neighborhood structures are spatial neighborhood as well as the neighborhood due to mobility. As the neighboring regions tend to have similar number of cases of infections, it is reasonable to consider the spatial neighborhood structure. Spatial neighborhood is a matrix containing the binary information, i.e. 1 if the regions are sharing common border and 0 if the regions dont share the border.

Along with this connectivity, the regions are also connected by the means of movement, i.e. even though some regions may not share the borders, there may be movement between them which could create a connection. Thus with this connection, spatial correlation may exist and to account for this correlation mobility based neighborhood structure is required. The mobility based matrix considered in this model is a median mobility matrix depicting the information of the median through out the study period. Median is chosen in this case to reduce the effects of outliers but this matrix can be any representative matrix from the study time period.

## 3.3 LSTM Model

The recurrent neural networks are effective models to model temporal events as they are able to predict the temporal events based on long term dependencies. In particular LSTM model is able to cope with the gradient vanishing problem. These LSTM models require data to be in supervised format thus the input data is expected in the format explained in the section 3.2.1.

The LSTM model in the LSTM Bayesian model accounts for the temporal trend of the disease infections within a particular area. It is assumed that the LSTM model learns the temporal patterns more than the spatial dependence and correlation, although some spatial covariate information are also part of input of this model. The aim of the LSTM model is to learn from the past events in an area with the LSTM and also incorporate spatial dependencies from some spatial covariate information. Thus, LSTM model is able to learn $F_{ti}$ in the equation 3.1 In the following section the model architecture is described.

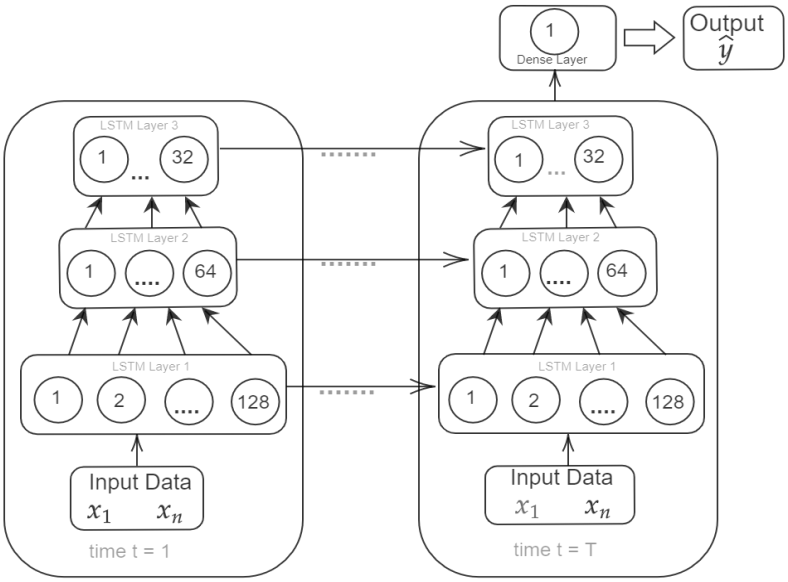



Figure 3.2: Architecture of LSTM model

**Architecture**

The LSTM model has 131,489 parameters consisting of three stacked LSTM layers which are recurrently used for the time period T. The first, second and the third LSTM layer has 128, 64 and 32 hidden units respectively. 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 model is shown in figure 3.2.

## 3.4 Bayesian Inference

The aim of performing Bayesian inference as a second stage is to model uncertainty in the prediction of number of infections in terms of a probabilistic spatio-temporal stochastic model. The count variable $Y_{ti}$ i.e. the number of infections on a area i at time t has a Poisson distribution expressed as

$$Y|\theta_{it} \sim Poisson(\theta_{it}) \tag{3.3}$$

The general log linear model is adopted (Wakefield, 2007)

$$\eta_{it} = log(\theta_{it}) = \mu + z_i^T \beta + b_{it} \tag{3.4}$$

where $\mu$ is the intercept, the covariates and their coefficients come in the term $z_i^T \beta$, and the $b_i$ are the random effects. These random effects are modelled as

$$b_{it} = \delta_t + \xi_i + \zeta_i \tag{3.5}$$

where the random effects are decomposed as a temporal trend $\delta_t$, and $\xi_i$ and $\zeta_i$ that account for the spatial correlation due to the spatial neighborhood relations and the mobility respectively. This model has been adopted and modified as used

by (Jalilian & Mateu, 2020). The spatial neighborhood effect $\xi_i$ on the model is the same but the $\zeta_i$ is modified following a mobility based neighborhood structure.

The temporal trend $\delta_t$ has been modelled using a Random Walk structure, which accounts for the short and long temporal trend (Fahrmeir & Kneib, 2008). The spatial correlation due to the neighborhood structure is modelled through a Besag York Mollie (BYM) model (Besag et al., 1991).The additional spatial correlation due to the mobility is modelled by assuming the random effect $\zeta_i$ following a Gaussian Markov Random Field (GMRF).

The predictions from the LSTM model are plugin into the Bayesian mechanism as expected values to further fit the Bayesian approach. Usually in Bayesian analysis, the values to be predicted are left empty and the model computes the mean prediction (Zuur et al., 2017). In order to avoid overfitting by the model, these LSTM predictions cannot be used as covariate information.

# Chapter 4

# Experiment Design and Implementation

This chapter presents the study area, data sources and the preprocessing applied to the data and then explains the experiment design, model implementation and evaluation.

## 4.1 Study area

The daily COVID-19 cases in the autonomous community of Castilla-Leon in Spain were analyzed in this study. Castilla-Leon is the largest autonomous community in Spain by area located in the northwest part of Spain. The autonomous community has a population of around 2.5 million and is ranked third among the autonomous communities in offering social services to the citizens. Figure 4.1 is the location map of Castilla-Leon showing the location of the community in Spain and the 245 health-zones in the community.

## 4.2 Data Sources

The daily covid -19 cases data were retrieved from the open data portal of Castilla-Leon [1]. The datasets are aggregated to the health-zones level, and although there are 247 health-zones in Castilla-Leon, after the initial preprocessing 245 health-zones data are used for the study [2]. The data set from March 1, 2020, until November 13, 2020, were used for the study.

The daily mobility data for the study area was acquired from Barcelona Supercomputing Center flowmap dashboard [3]. Similarly, the socio-demographic dataset and the health-zone boundary in the form of shapefile were downloaded using the open data platform [4].

[1] https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html

[2] In this study, the health-zones SORIA NORTE, SORIA SUR and SORIA RURAL are aggregated to form a single unit

[3] https://flowmaps.life.bsc.es/flowboard/

[4] https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html

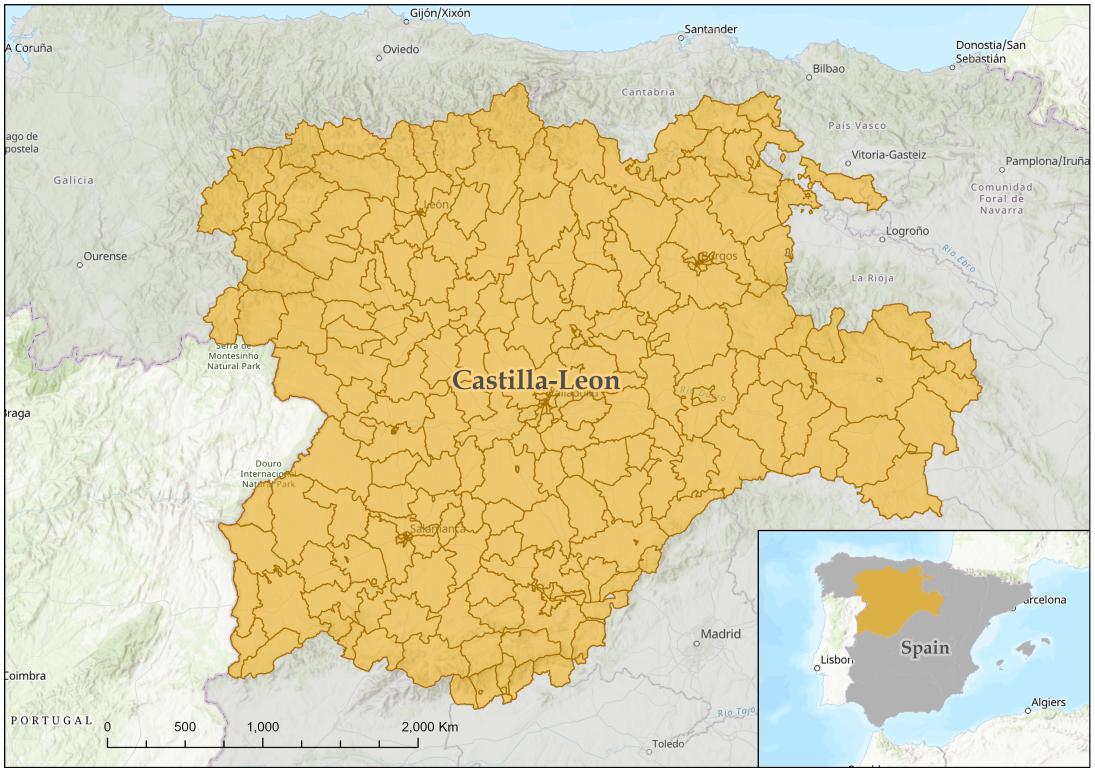



Figure 4.1: Study Area: Autonomous Community of Castilla-Leon, Spain

Table 4.1: 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 247 health-zones |

### 4.2.1 COVID-19 data

The health-zones level daily new infected cases of covid was acquired from the open data portal of Castilla-Leon. The dataset from the March 1, 2020 till November 13, 2020 were used for the study.

### 4.2.2 Mobility Data

The mobility data acquired from the data portal of Barcelona Supercomputing Center was prepared by the Ministry of Transport, Mobility, and Urban Agenda. In preparation of the data, the main data source was anonymized records from mobile phones. These recorded events contain both active events also known as Call Detail Records (CDR) and passive events with the 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 as shown in figure 4.2, districts and provinces level. Along with these records from the cell phones, landuse data, and 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).

```
fecha|origen|destino|periodo|distancia|viajes|viajes_km
20200221|01001_AM|01001_AM|00|0005-002|8.700|16.545
20200221|01001_AM|01001_AM|00|002-005|29.627|87.800
20200221|01001_AM|01001_AM|00|005-010|5.382|32.671
20200221|01001_AM|01001_AM|00|010-050|16.206|196.664
20200221|01001_AM|01001_AM|01|005-010|22.328|166.005
20200221|01001_AM|01001_AM|02|005-010|18.393|130.593
20200221|01001_AM|01001_AM|02|010-050|28.099|518.697
20200221|01001_AM|01001_AM|03|002-005|4.350|17.387
20200221|01001_AM|01001_AM|03|005-010|4.350|27.336
20200221|01001_AM|01001_AM|03|010-050|15.398|228.776
20200221|01001_AM|01001_AM|04|005-010|51.427|363.970
20200221|01001_AM|01001_AM|04|010-050|32.219|416.177
20200221|01001_AM|01001_AM|05|002-005|4.350|10.290
20200221|01001_AM|01001_AM|05|005-010|77.915|539.565
20200221|01001_AM|01001_AM|05|010-050|11.599|204.742
20200221|01001_AM|01001_AM|06|005-010|36.090|221.325
20200221|01001_AM|01001_AM|06|010-050|20.358|321.452
20200221|01001_AM|01001_AM|07|005-010|61.770|483.171
20200221|01001_AM|01001_AM|07|010-050|33.066|551.780
20200221|01001_AM|01001_AM|08|002-005|34.532|142.422
20200221|01001_AM|01001_AM|08|005-010|45.320|295.310
20200221|01001_AM|01001_AM|08|010-050|16.986|262.407
20200221|01001_AM|01001_AM|09|002-005|41.210|141.426
```

Figure 4.2: Mobility data Sample

### 4.2.3 Socio-demographic data

The socio-demographic dataset for the health-zones was acquired from the open data portal of Castilla-Leon. The following table 4.2 shows the socio demographic variables used in the study

Table 4.2: 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 |
| `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 |
| `hzone_type` | Type of health-zone (urban/rural) |

## 4.3 Data Preprocessing

As the data are from different sources, it requires merging and combining. Join operations were carried out to join these data. This also involved the preparation of unique keys for each health-zones (referred as hzcode). To ensure the consistency of the data, redundant information was removed. The repeating date was removed. Data with negative values and empty values in the number of cases were set to 0.

The preparation of the data involved the conversion of various data into the required format and the creation of new fields as shown in 4.3. In the computation of the travel restrictions, the dates considered reflecting the decisions by the Spanish Government. So, the ranges of dates considered are pre-lockdown, lockdown, post-lockdown, and restricted travels period. Each of these periods has different significance on the movement of people and also other human behaviors such as social distancing, awareness, etc.

Table 4.3: Summary of variable transformations.

| Variable Name | Description |
|---|---|
| Day of the week | Computed from the date |
| Travel Restrictions | Factor considering the travel restrictions time period<br>Factor consists following values:<br>0 - 2020-03-01 - 2020-03-13<br>1 - 2020-03-13 - 2020-07-16<br>2 - 2020-07-16 - 2020-10-01<br>3 - 2020-10-01 - 2020-11-13 |
| health-zone type | Factor based on health-zones' type<br>0- Rural<br>1- Urban |
| Average No of cases in neighboring health-zones | Average of no of cases in health-zones directly in contact |
| Shifted Cases | The shifted cases of COVID by 7 days |
| Mobility Matrices | Conversion from municipality to health-zone wise |

**Mobility data conversion**

The available daily mobility data was at the municipality level. These municipalities with population less than 1000 were combined to form aggregated zones. These aggregations were converted to the health-zone level by applying spatial overlay functions and dividing the movement data in proportion to the area of the overlay regions.

## 4.4 Experiment Design

### 4.4.1 Training Validation Test data division

During the study period, two waves of covid infections have been reported in the study area. With the start of covid infections worldwide, the first infection in the study area was seen in early March. The number of daily cases was rising very quickly. This stage of the infection is referred to as the first wave. During this time in most of the health-zones of the study area, the daily infections had reached the peak and the government had restricted the movement by applying the lockdown. By mid June, the daily infection rate had gone down and on 16 July, the lockdown was lifted and movement and other activities were allowed with some restrictions. Starting early August, the number of daily infections increased rapidly which is referred to as the second wave of infections which has continued until the end of the study period.

In this study, it is important to train the model with both of these waves of infections as they depict two different scenarios, the first wave shows the condition of lockdown and a trend of reducing the daily number of infections, while the second wave shows the condition of restriction in movements and other daily activities without lockdown. Thus the study period selected for the training includes both the waves data. The following figure 4.3 shows the training, validation and test period.

The training phase of the study is March 1, 2020, till October 22, 2020, were used while validation was done for the data from October 22, 2020, till November 6, 2020. Finally, the model was tested by predicting the daily infections for the last week of the study i.e. November 6, 2020, till November 13, 2020.

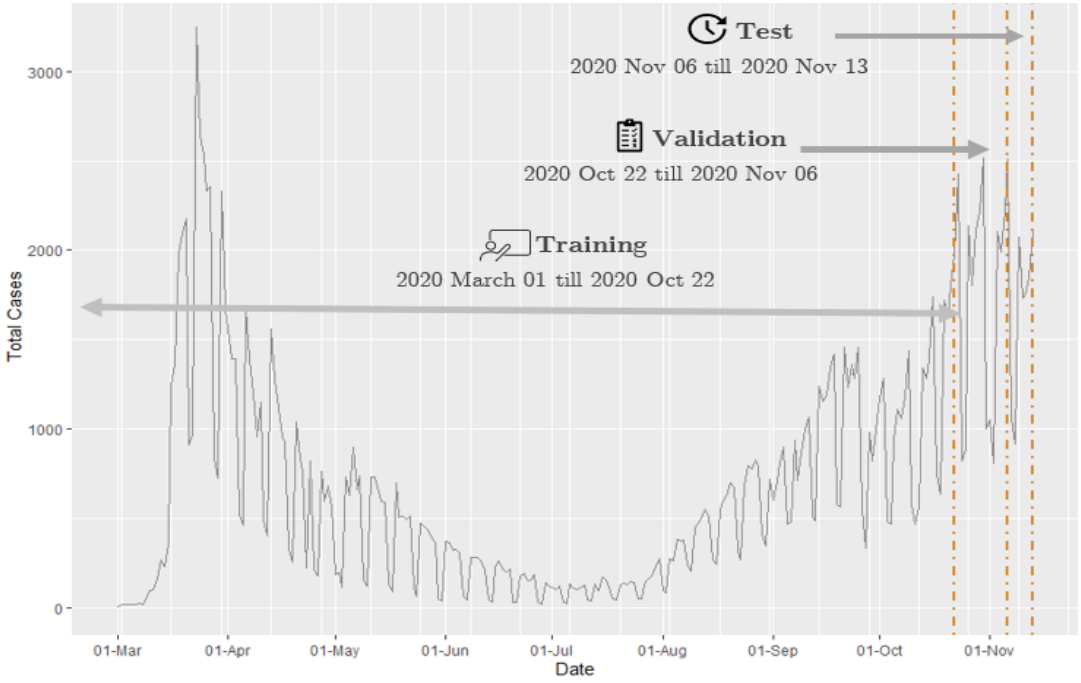



Figure 4.3: Division of training, validation and test dataset

## 4.5 Implementation

The model explained in Chapter 3 is the combination of two parts: Recurrent Neural Network with the LSTM and the Bayesian inference. The Bayesian inference is performed with the INLA (Rue et al., 2009). The model is referred to as LSTM-INLA model. The following sections describe the implementation details of these two sections along with the computation of the spatial weights of the model.

### 4.5.1 Computation of Spatial Weights

As described in section 3.2.2, the spatial weights are computed to incorporate the daily movement matrices into the model. Along with the mobility matrices, the daily infection within a time lag $\Delta t$ has been introduced. In the case of COVID-19, this lag period can be assumed equal to the incubation period as proposed in clinical studies (Guan et al., 2020). Thus, we used a 4 days lag time period to compute these spatial weights.

### 4.5.2 LSTM

Python programming language and the library Keras [5] is used for development of the model. The selection of the model was done by tuning the parameters and hyper parameters. The training and validation loss for the combination were analysed along with the Root Mean Squared Error (RMSE) value for all the regions. The following table 4.4 depicts the selection of the parameters and hyperparameters.

[5] https://keras.io/

Table 4.4: Summary of Parameters and Hyperparameters in LSTM model.

| **Parameter** | **Value** |
|---|---|
| Number of LSTM layers | 3 |
| Hidden Units in LSTM layers | Layer 1: 128<br>Layer 2: 64<br>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<br>Learning Rate: 0.001<br>$\beta_1$: 0.9<br>$\beta_2$: 0.999 |
| Batch Size | 10 |

### 4.5.3 INLA

In performing the Bayesian inference, R package R-INLA [6] is used. The R-INLA package provides all the required possibilities of covariates additions, prior distribution definition, and the definition for the spatial and temporal effects used in the models. The functions in the packages are used to define the regression, and run the model and perform the predictions. The model provides mean prediction values along with the posterior distributions of the parameters. The evaluation of the INLA model is done with the Deviance Information criterion (DIC), Watanabe-Akaike Information Criterion (WAIC) values, and Conditional Predictive Ordinate (CPO).

The following table 4.5 shows the spatial and temporal random effect components in the model and the corresponding functions from R-INLA used.

Table 4.5: Model Components and their implementation functions in R-INLA.

| Component | Description | R-INLA Function |
|---|---|---|
| $\delta_t$ | The temporal random effects | rw |
| $\xi_i$ | Spatial Random effect due to neighborhood structure | bym |
| $\zeta_i$ | Spatial Random effect due to mobility | generic1 |

## 4.6 Model Evaluation

The evaluation of the model was performed with two baseline models as shown in the table 4.6. As discussed in the section 2.2.6, the evaluation metrics chosen was RMSE. A mean value from all the RMSEs computed for all the health-zones

[6] www.r-inla.org

of the study was computed to compare the models. Similarly, for the comparison of the INLA models the WAIC, DIC and CPO values are compared.

Table 4.6: Baseline models for Evaluation.

| Model Name | Description |
|---|---|
| LSTM | A Recurrent Neural Network with same configuration that LSTM-INLA model has in the LSTM part |
| INLA | Regression model with the same configuration that LSTM-INLA model has in the INLA part |

# Chapter 5

# Results and Discussions

This chapter presents and discusses the results achieved from the experiment designed as described in the Chapter 4. The first section shows in brief the exploratory analysis performed on the available variables. The second section presents the results from the selected LSTM-INLA model and the comparison of the model with the baseline models i.e. LSTM and INLA. The third section describes the impact of the spatial weights factor on the results of the model. The fourth section presents the interpolation and predictions from the model. Finally, the last section describes the limitation of the models and possible further improvements are presented.

## 5.1 Data Exploration

For exploratory analysis on the data, the temporal variations and the spatial distributions of the available covariate information and the number of cases were analyzed. The spatial distribution of the number of COVID-19 infections per 10000 population in health-zones of the study area is shown in figure 5.1 and the distribution of the number of cases per 10000 population in health-zones in figure 5.2. This shows the number of cases in all the health-zones vary. The highest number of cases is 1770 per 10000 population and the lowest number of cases is 200 per 10000 population. There are many health-zones with very few cases and very few health-zones with the high number of cases.

The temporal trend of the number of COVID cases per 10000 population

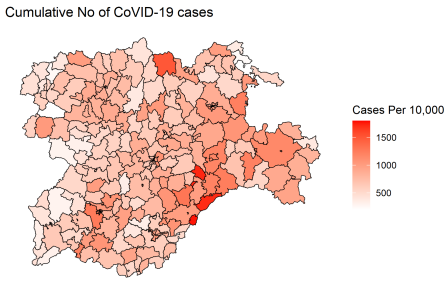



Figure 5.1: Spatial distribution of COVID-19 cases in the study area

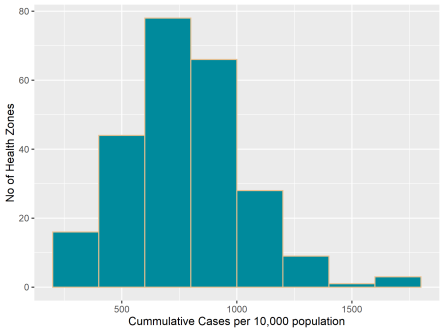



Figure 5.2: Distribution plot of the cumulative cases

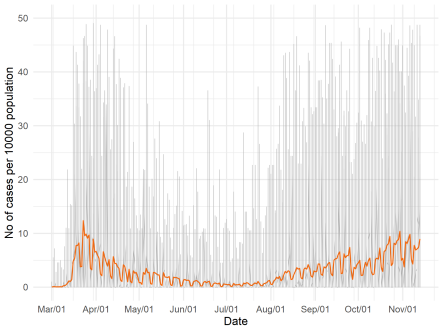



Figure 5.3: Temporal plot of COVID-19 cases

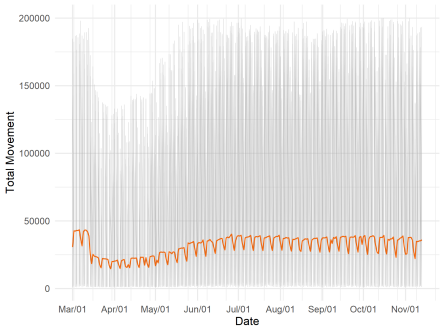



Figure 5.4: Temporal plot of the total mobility

and the total mobility in the health-zones are shown in the figures 5.3 and 5.4 respectively. The orange lines represent the mean for each day. In the initial days of the study period, it can be seen that there are some similarities between the mobility and the number of cases. But in the later period of the study, there is not a very clear pattern, although there is a slight reduction in mobility as the cases were high. Apart from this, the weekly trends in both data can be seen, as there are sudden drops each weekend.

## 5.2 Model Evaluation

As explained the section 4.6, the evaluation of the model was done with two baseline models LSTM and INLA. These models were trained or fitted with the same configuration and same covariates. The statistics for the comparison is the RMSE values and for the INLA based model, the WAIC and DIC values are also compared. Table 5.1 shows the RMSE for the predictions for the last week of the study i.e. (from 2020-11-06 to 2020-11-13) from the model and the baselines. The plots and the maps from the predictions are presented in the prediction section.

Table 5.1: Models comparision

| Model | RMSE | WAIC | DIC |
|---|---|---|---|
| LSTM-INLA | 9.11 | 202324.70 | 202145.79 |
| INLA | 58.11 | 193364.41 | 193530.17 |
| LSTM | 12.95 | - | - |

The RMSE value for the proposed model is 9.11 as compared to the value of 58.11 for the INLA model and 12.95 for the LSTM model. The RMSE value for the model LSTM-INLA is lower than that of the INLA model while the WAIC values and DIC values are higher. And, although the RMSE values in the LSTM model are similar and there is only a slight increase in the LSTM-INLA model, the ability to predict the number of cases with the credible interval gives advantages to the LSTM-INLA model. Thus, it can be said the proposed LSTM-INLA model is able to perform better than the baseline models as reported in figure 5.1.

The spatial covariates as well as the mobility are transferred into a spatial weight factor as described in section 3.2.2. These spatial weight factors were

introduced to the model in the form of covariate information. The impact of these spatial weights on the model predictions are described in this section. The model LSTM-INLA was fitted with and without the spatial Weights to find out the influence of the spatial weight on the model. The models were evaluated based on the RMSE, WAIC, and DIC values.

Table 5.2: Evaluation of Spatial Weights

| Model | RMSE | WAIC | DIC |
|---|---|---|---|
| LSTM-INLA with Spatial Weights | 9.11 | 202324.70 | 202145.79 |
| LSTM-INLA without Spatial Weights | 13.35 | 207052.62 | 207278.62 |

The table 5.2 shows that the model WAIC values and the DIC values are better in the case when the spatial weights are computed. Similarly, the RMSE values have improved in the case of the proposed model with the inclusion of the spatial weights.

The posterior mean of the significant fixed effect parameters of the model and their respective 95% credible interval are shown in the table 5.3. It is observed that the weekdays are equally significant but the weekends are less significant. This is a reasonable finding because the number of tests and reporting on the weekends are lower. Similarly, the spatial weights computed are also significant with a mean value of 0.026.

Table 5.3: Posterior Mean and credible interval of the significant parameters

| Parameters | Mean | Credible interval |
|---|---|---|
| Monday | -9.072 | -9.749, -8.396 |
| Tuesday | -9.076 | -9.748, -8.403 |
| Wednesday | -9.051 | -9.721, -8.382 |
| Thursday | -9.081 | -9.757, -8.404 |
| Friday | -9.052 | -9.739, -8.365 |
| Saturday | -9.86 | -10.544, -9.176 |
| Sunday | -10.005 | -10.685, -9.325 |
| Spatial Weights | 0.026 | 0.025, 0.027 |

In the figures 5.5 and 5.6, the maps showing the spatial random effects due to the neighborhood structure and the mobility respectively are shown. The values suggest both random effects have an influence on the model. The spatial random effect due to the neighborhood structure $\xi_i$ have clusters around the major cities like Leon, Valladolid, and Burgos whereas the random effect due to the mobility $\zeta_i$ are distributed evenly with some exceptional peaks. The figure 5.7 shows the mean and the 95 % credible interval for the temporal trend $\delta_t$. This temporal trend suggests similar findings to that of section 5.1 that there exist two peaks or waves of infections in the study period: one in early April 2020 and another one starting after August 2020. Similarly, it is also seen that the first wave reduced down quickly whereas the second wave does not have a quick downward trend but has been consistent afterwards. In the plot of the temporal effect, the trend

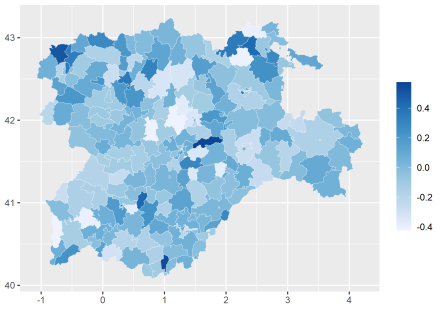



Figure 5.5: Spatial Random effect due to neighborhood structure

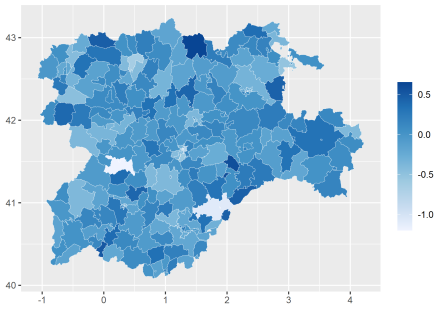



Figure 5.6: Spatial Random effect due to the mobility

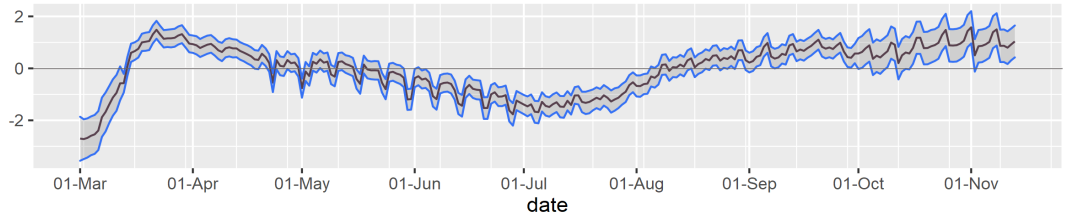



Figure 5.7: Trend of temporal effect

on both sides of the zero line with fluctuating values supports the inclusion of the temporal effect in the model design.

## 5.3 Interpolation and Predictions

The fitted model was used for health-zone wise one week ahead prediction. The prediction was done for the last week of the study which is from 2020-11-06 till 2020-11-13. The LSTM model was used to initially predict for the same time range, and these results were used in the prediction for the INLA part of the model. In general, to predict the values in R-INLA package, the values we want to predict are set as null values (Zuur et al., 2017). But in this case, since the predictions from the LSTM model are considered the expected values, instead of setting the values as null values the predictions from LSTM were used, to provide the model with a reference of the values to predict.

Figure 5.8 shows the predicted values from the LSTM-INLA model and the 95% credible interval for a few selected health-zones of the study area for the prediction period. The actual number of cases on that day are shown in red-colored dotted lines whereas the number of cases predicted by the LSTM model are also shown for comparison (green in color). Generally, the LSTM-INLA is predicting results better than that of the LSTM model. It can be seen that the LSTM-INLA model's mean prediction and the 95% credible interval is close to the observed values. Similarly, LSTM model has not been able to follow the pattern of the observed cases and the predictions also lack the credible interval. Furthermore, the predictions from the INLA model for the same health-zones have very large credible intervals and are not able to follow the pattern of the observed actual cases, which are shown in the figure B.1 and B.2. The results

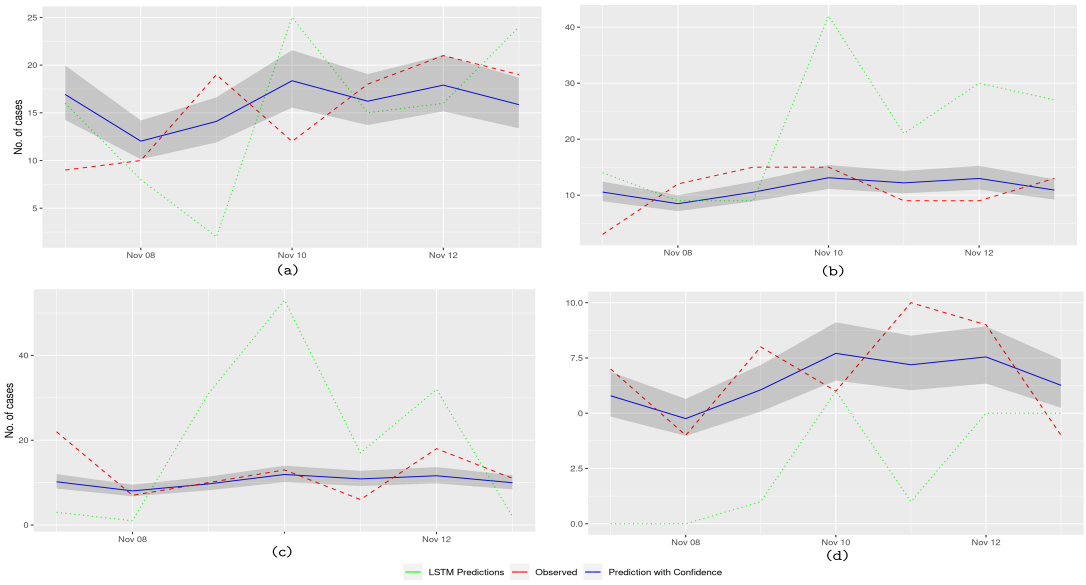



Figure 5.8: Prediction of daily COVID-19 cases for dates 2020-11-07 till 2020-11-13 from the LSTM-INLA model for selected health-zones (a) San Agustin, (b) Portillo, (c) Tortolo and (d) Canterac

from the interpolation i.e. the fitting of the model for the whole time scale is shown in the figure A.1 and A.2 for these 4 health-zones.

The prediction map for the day 2020-11-12 is shown in figure 5.9. In the figure 5.9, (a) shows the predictions for each health-zone and (b) shows the observed values on that day. It can be seen that LSTM-INLA model is able to predict the spatial distribution in a good way. The clusters are similar for the major cities of the study area i.e. Burgos, Salamanca, Leon and Valladolid and the places with lower daily number of cases. The prediction is visualized in shiny app [1].

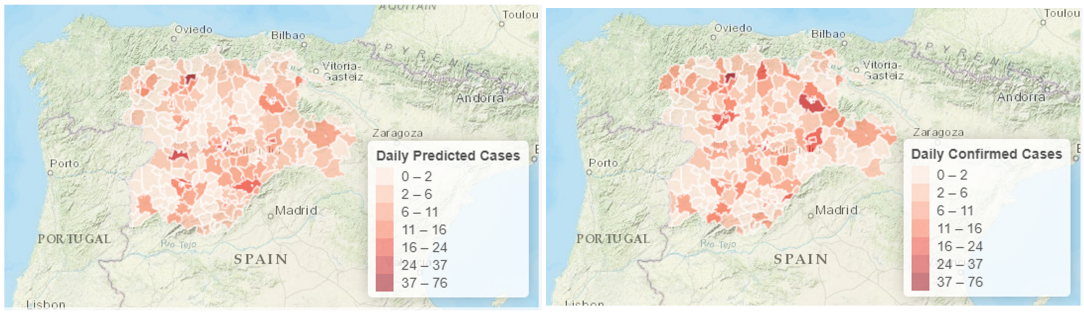



Figure 5.9: Map showing the predictions and observed values for the day 2020-11-12 (a)Predictions from the LSTM-INLA model and (b) Observed Number of cases

[1] https://poshan-niraula.shinyapps.io/inla-results/

## 5.4 Limitations and Future Directions

The limitations and possible future improvements in this work are presented in this section.

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, the environmental conditions, and most importantly, the 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 major factors considered 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. Generally, the prediction results are good with low RMSE values but in some cases, the sudden rise in the number of cases on a day and sudden fall on the next day were not predicted properly by the model. Similarly, one assumption common to most disease modeling, in this case, is that the number of reported cases is assumed to be equal or at least representative of the actual number of infections.

The focus of this work is on the combination of neural networks and Bayesian inference. The predictions from neural networks were used as expected values for the Bayesian inferences which can be improved by transferring the predictions to a prior distribution and use them as the prior information in the Bayesian inference. Similarly, this task was performed by working on them separately which has the associated complexities in the development of the model. A combined solution such as spatio-temporal recurrent neural networks able to predict results with uncertainties can be a possible alternative.

In this study, the daily mobility matrices were converted into covariates and the median mobility as a neighborhood structure. Instead of using one median mobility, an approach to generate the daily neighborhood matrices and use them to account for the spatio-temporal correlation could be an enhancement of this work. Similarly, more detailed mobility data should be used to have a better evaluation of the impact. graph-based neural networks (Dhamodharavadhani et al., 2020) .

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.

# Chapter 6

# Conclusions

For modeling the spread and outbreak of infectious diseases, a model comprising the combination of neural network and Bayesian inference has been presented. This model is able to model the number of cases of infectious diseases in areal units such as municipalities or health-zones. The predictions from the model have uncertainties associated with them. The model accounts for the spatial correlation due to the spatial neighborhood relation and also due to a median mobility matrix. In addition to this, the daily matrices of movements were used in the computation of the spatial weight which is added in the model as one of the covariate information.

The model was evaluated with the case study of COVID-19 data from the autonomous community of Castilla-Leon in Spain consisting of 245 health-zones. The dataset used were daily COVID-19 cases from March 1, 2020, till November 13. 2020. The model was able to predict the number of daily infections in each health-zones, and these predictions and the credible interval were compared with the observed data. The results from the evaluation showed that the model performed well generally. The model outperformed the model with only neural networks and only bayesian regression. The mobility transformed as a spatial weight as well as the spatial correlation introduced as a result of the mobility was found influential. However, the results also highlighted some challenges and limitations in terms of the addition of covariate information, and the inability to predict sudden peaks and lows.

In future works, 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. Furthermore, detailed mobility information may be introduced as a spatio-temporal effect with the use of graph concepts.

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 the Bayesian inference to associate the uncertainties in the predictions also considering the spatial dependencies due to the mobility. In conclusion, this thesis is able to present a model that can provide accurate predictions of infectious diseases and help in a way to mitigate the impacts.

# Bibliography


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). 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*.

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., Boulanger-Lewandowski, N. & Pascanu, R. (2013). Advances in optimizing recurrent networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8624–8628.

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

Bernadinelli, L., Pascutto, C., Best, N. & Gilks, W. (1997). Disease mapping with errors in covariates. *Statistics in Medicine*, *16*(7), 741–752.

Bernardinelli, L., Clayton, D. & Montomoli, C. (1995). Bayesian estimates of disease maps: How important are priors? *Statistics in medicine*, *14*(21-22), 2411–2431.

Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1–20.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V. & Pebesma, E. J. (2008). *Applied spatial data analysis with r* (Vol. 747248717). Springer.

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

Brauer, F. (2008). Compartmental models in epidemiology. *Mathematical epidemiology* (pp. 19–79). Springer.

Breslow, N., Leroux, B. & Platt, R. (1998). Approximate hierarchical modelling of discrete data in epidemiology. *Statistical Methods in Medical Research*, *7*(1), 49–62.

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.

Carlin, B. P. & Louis, T. A. (2008). *Bayesian methods for data analysis*. CRC Press.

Chai, T. & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, *7*(3), 1247–1250.

Chitnis, N., Hyman, J. M. & Cushing, J. M. (2008). Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model. *Bulletin of mathematical biology*, *70*(5), 1272.

Chowell, G., Fenimore, P. W., Castillo-Garsow, M. A. & Castillo-Chavez, C. (2003). Sars outbreaks in ontario, hong kong and singapore: The role of diagnosis and isolation as a control mechanism. *Journal of theoretical biology*, *224*(1), 1–8.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

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*.

Duch, W. & Jankowski, N. (1999). Survey of neural transfer functions. *Neural Computing Surveys*, *2*, 163–213.

Fahrmeir, L. & Kneib, T. (2008). On the identification of trend and correlation in temporal and spatial regression. *Recent advances in linear models and related areas* (pp. 1–27). Springer.

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

Fausett, L. (1994). *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. Prentice-Hall, Inc.

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

Gelman, A., Hwang, J. & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, *24*(6), 997–1016.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A. & Colaneri, M. (2020). Modelling the covid-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine*, 1–6.

Giuliani, D., Dickson, M. M., Espa, G. & Santi, F. (2020). Modelling and predicting the spatio-temporal spread of coronavirus disease 2019 (covid-19) in italy. *Available at SSRN 3559569*.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Gómez-Rubio, V. (2020). *Bayesian inference with inla*. CRC Press.

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*.

Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D. S. et al. (2020). Clinical characteristics of coronavirus disease 2019 in china. *New England journal of medicine*, *382*(18), 1708–1720.

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.

Haining, R. (2001). Spatial sampling. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social behavioral sciences* (pp. 14822–14827). Pergamon. https://doi.org/https://doi.org/10.1016/B0-08-043076-7/02510-9

Haykin, S. & Network, N. (2004). A comprehensive foundation. *Neural networks*, *2*(2004), 41.

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

Jalilian, A. & Mateu, J. (2020). A hierarchical spatio-temporal model to analyze relative risk variations of covid-19: A focus on spain, italy and germany. *arXiv preprint arXiv:2009.13577*.

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*.

Kermack, W. O. & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, *115*(772), 700–721.

Kingma, D. P. & Lei Ba, J. (2017). *ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION* (tech. rep.).

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

Kraemer, M. U. G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., Covid, O., Hanage, W. P., Brownstein, J. S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. G. & Scarpino, S. V. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China, 6.

Kraemer, M., 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.

Le, Q. V., Jaitly, N. & Hinton, G. E. (2015). A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, *438*(7066), 355–359.

Martino, S. & Rue, H. (2010). Case studies in bayesian computation using inla. *Complex data modeling and computationally intensive statistical methods* (pp. 99–114). Springer.

Massaro, E., Kondor, D. & Ratti, C. (2019). Assessing the interplay between human mobility and mosquito borne diseases in urban environments [Num-

ber: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *9*(1), 16911. https://doi.org/10.1038/s41598-019-53127-z

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.

Medsker, L. & Jain, L. (1999). Recurrent neural networks: design and applications.

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

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 Negl Trop Dis*, *8*(4), e2769.

Organization, W. H. et al. (2019). World malaria report 2019.

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.

Pandey, G., Chaudhary, P., Gupta, R. & Pal, S. (2020). Seir and regression model based covid-19 outbreak predictions in india. *arXiv preprint arXiv:2004.00958*.

Pascanu, R., Gulcehre, C., Cho, K. & Bengio, Y. (2014). *How to Construct Deep Recurrent Neural Networks* (tech. rep.).

Pettit, L. (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

Riebler, A., Sørbye, S. H., Simpson, D. & Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, *25*(4), 1145–1165.

Rivers, C. M., Lofgren, E. T., Marathe, M., Eubank, S. & Lewis, B. L. (2014). Modeling the impact of interventions on an epidemic of ebola in sierra leone and liberia. *PLoS currents*, *6*.

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.).

Sharma, S., Sharma, S. & Athaiya, A. (2020). *ACTIVATION FUNCTIONS IN NEURAL NETWORKS* (tech. rep.). http://www.ijeast.com

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. et al. (2013). House-to-house human movement drives dengue virus transmission. *Proceedings of the National Academy of Sciences*, *110*(3), 994–999.

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.

Sutskever, I., Martens, J. & Hinton, G. (2011). *Generating Text with Recurrent Neural Networks* (tech. rep.).

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

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, *46*(sup1), 234–240.

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

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

Watanabe, S. & Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, *11*(12).

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, *78*(10), 1550–1560.

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen, K. & Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in pakistan. *Proceedings of the National Academy of Sciences*, *112*(38), 11887–11892.

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. (n.d.). Coronavirus Update (Live): 4,654,991 Cases and 309,133 Deaths from COVID-19 Virus Pandemic - Worldometer [Library Catalog: www.worldometers.info]. Retrieved May 16, 2020, from 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 [Number: 7 Publisher: Multidisciplinary Digital Publishing Institute]. *International Journal of Environmental Research and Public Health*, *17*(7), 2563. https://doi.org/10.3390/ijerph17072563

Zhou, C., Su, F., Pei, T., Zhang, A., Du, Y., Luo, B., Cao, Z., Wang, J., Yuan, W., Zhu, Y. et al. (2020). Covid-19: Challenges to gis with big data. *Geography and Sustainability*.

Zuur, A., Ieno, E. & Saveliev, A. (2017). *Beginner's guide to spatial, temporal, and spatial-temporal ecological data analysis with r-inla: Using glm and glmm*. Highland Statistics Limited. https://books.google.es/books?id=QvODzQEACAAJ

# Appendices

# Appendix A

# Interpolation Results from LSTM-INLA model

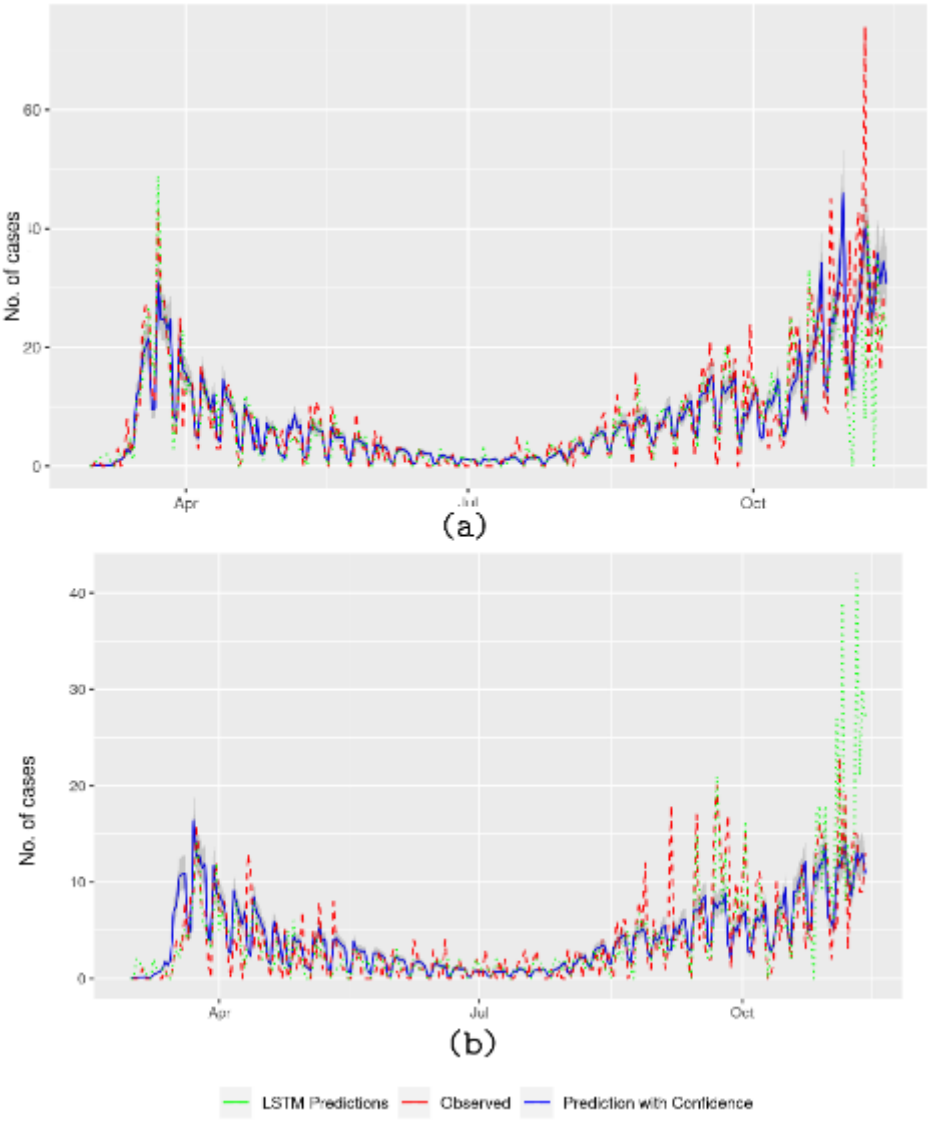



Figure A.1: Interpolation results from the LSTM-INLA model for health-zones a) San Agustin and (b) Portillo

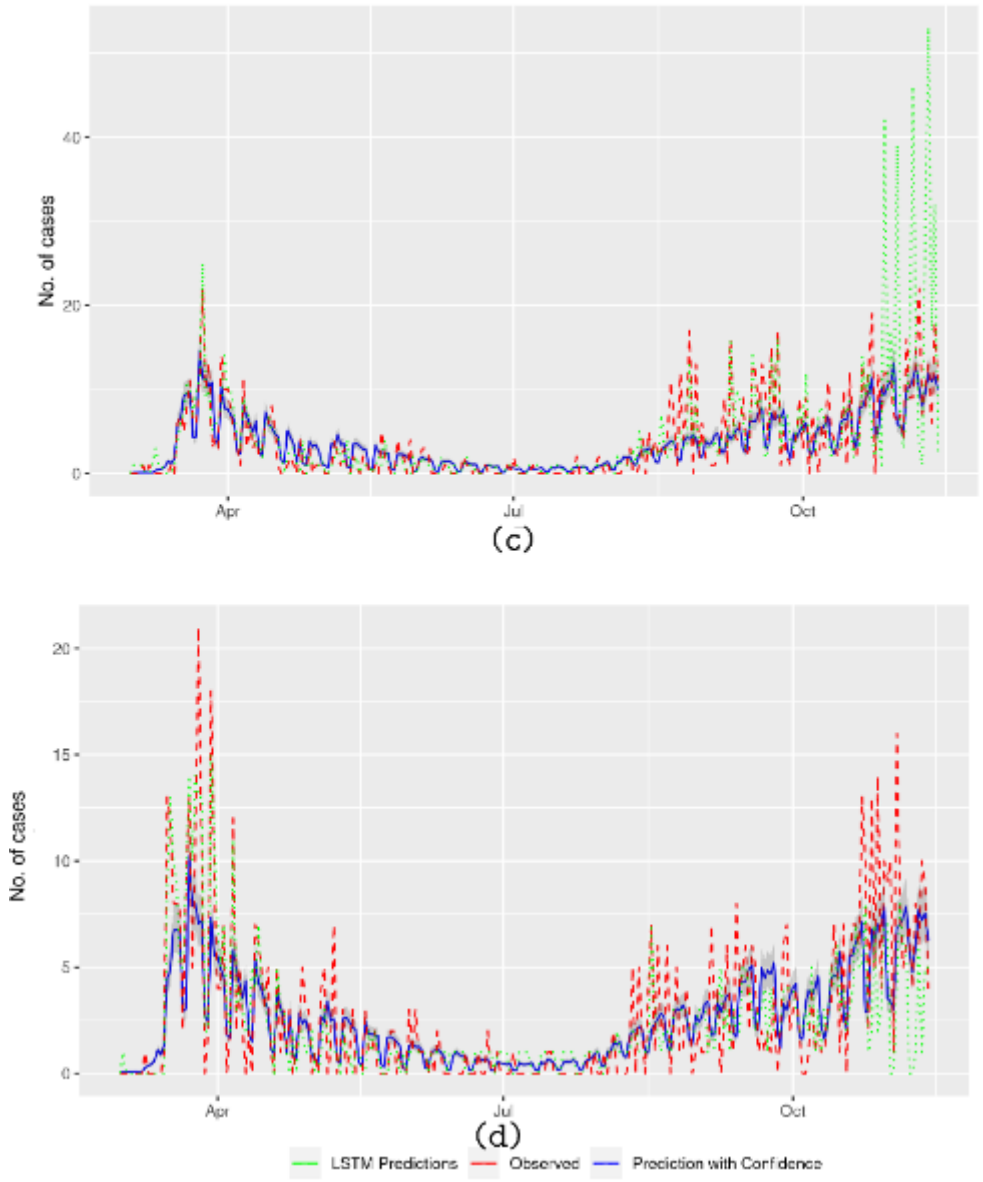



Figure A.2: Interpolation results from the LSTM-INLA model for health-zones (c) Tortolo and (d) Canterac

# Appendix B

# Prediction Results from INLA model

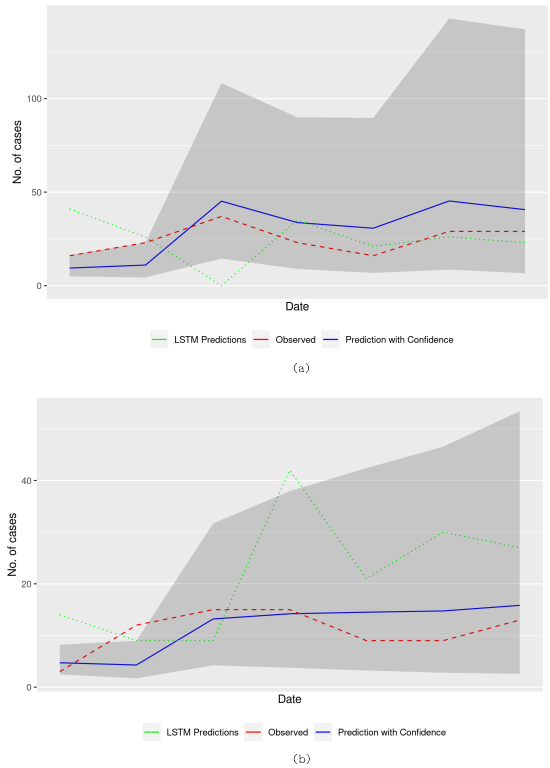



Figure B.1: Predictions results from the INLA model for health-zones a) San Agustin and (b) Portillo

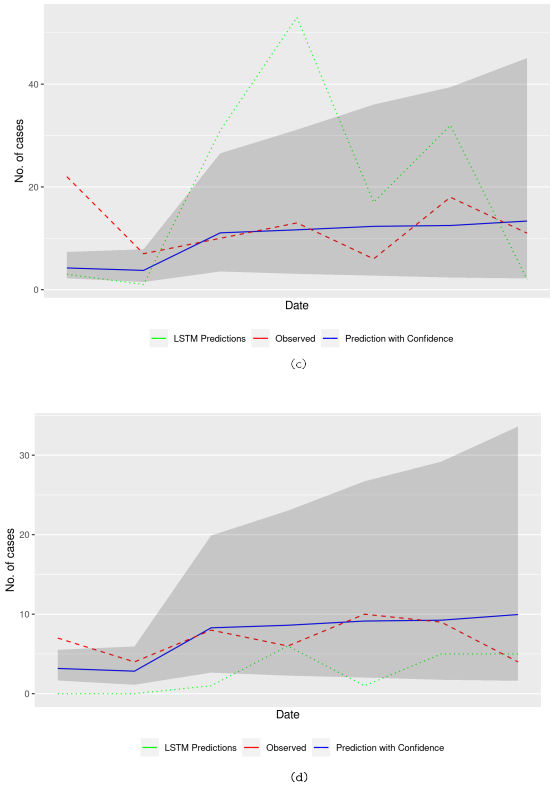
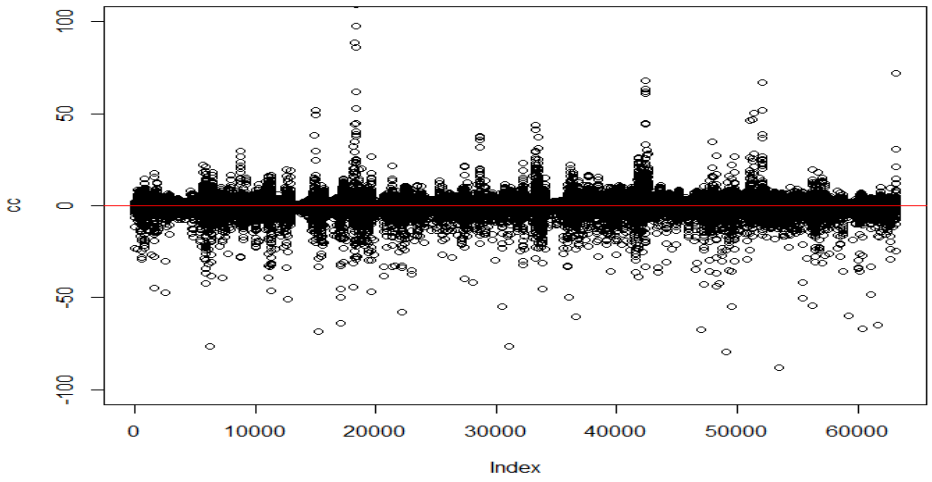


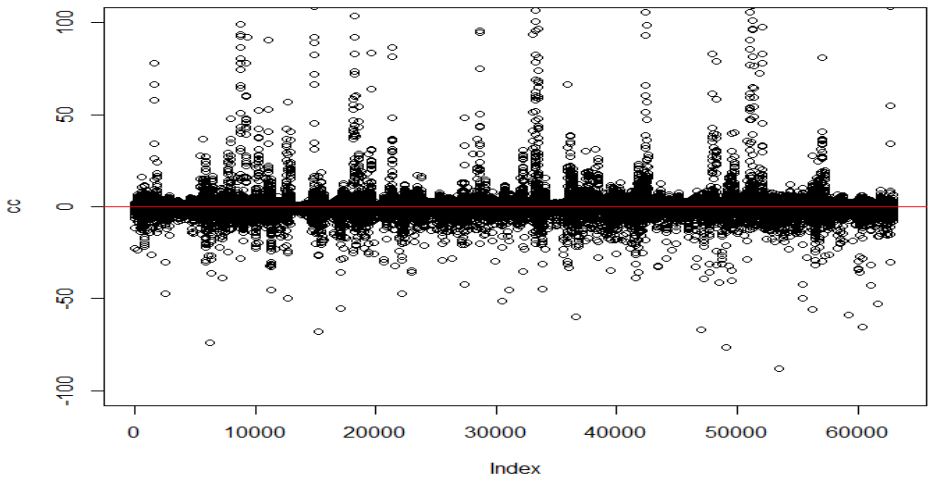

Figure B.2: Predictions results from the INLA model for health-zones (c) Tortolo and (d) Canterac

# Appendix C

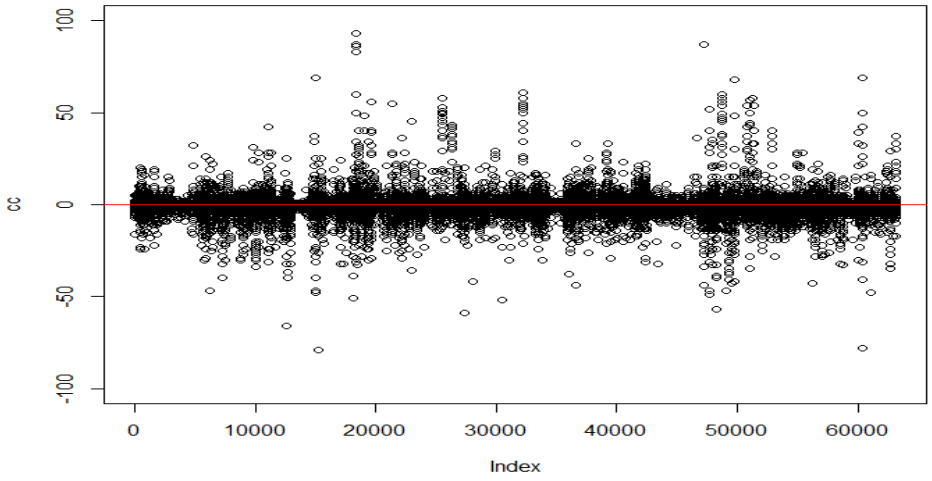
# Residual Plots




(a) LSTM-INLA model




(b) INLA model




(c) LSTM model

Figure C.1: Residual Plots for prediction from different models