Monte Carlo Deep Neural Network Model for Spread and Peak Prediction of COVID-19

G. N. Baltas, F. A. Prieto, M. Frantzi, C. Garcia, P. Rodriguez

March 2020

1 Introduction

Just a few days before the beginning of this year a new virus, widely known as the COVID-19[1], was detected in Wuhan, capital of the province Hubei, China. Since then, COVID-19 has spread all across the globe infecting more than half a million people[2] resulting to the passing of nearly 25000 patients. Beside the social pain that this new pandemic is causing, the measures put in force to halt the spreading of the virus are stressing the global economy indicating a domino effect that can last even longer than the probable eradication of COVID-19. Yet, these measures are necessary to prevent health system reach their capacity, an occasion where difficult decisions will need to be made such as prioritization of patients to be treated.

Estimating the evolution of COVID-19 is imperative for enhancing the efficiency of health systems and allocating resources. In this study, an Artificial Intelligence (AI) approach, based on Deep Neural Networks (DNN), is designed to predict the peak of the virus in Spain. The method consists of a data generation process based on Monte Carlo simulations of SIR epidemiology models and the development of the DNN prediction model. In Section 2 a brief summary of the virus evolution in Spain is presented. Section 3 is dedicating on describing the methodology that this work is based on while Section 4 presents the results. Finally Section 5 summarizes this paper findings.

2 Evolution of the COVID-19 in Spain

The first case in Spain was detected on January 31^{st} with an isolated case in Las Islas Canarias where the virus began to spread, with six cases reported by the 24^{th} of February. A second case was detected in Baleares on the 9^{th} of February.

[1] Coronavirus disease 2019

[2] At moment of writing this paper

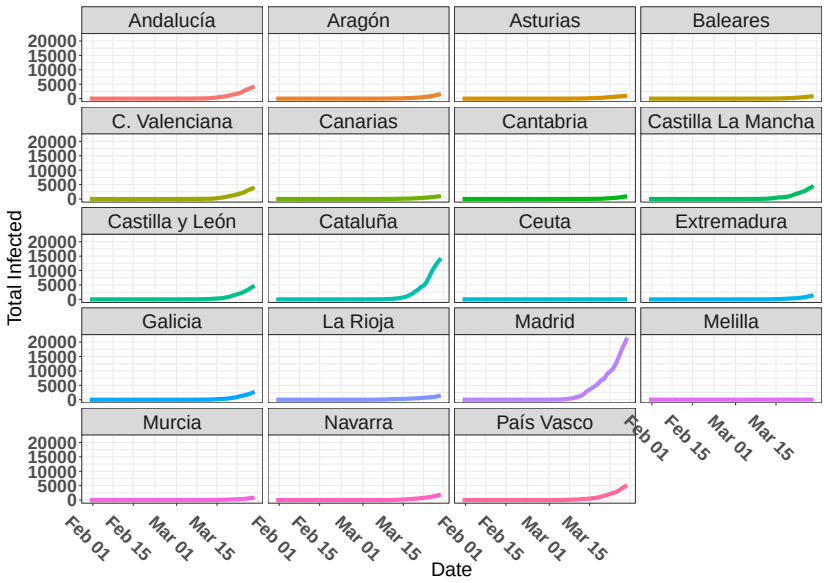

Figure 1: Total Infected Until 28^{th} of March per Spanish province

Until March 24^{th} there was no official registration of the virus outside the Spanish islands, however on the following day the virus was detected in Spain's largest cities, i.e. Madrid, Catalonia and Valencia.

Subsequently, from then on the number of cases within Spain started to grow exponentially. In Figure 1 the evolution of COVID-19 is depicted for each Spanish province separately [2]. According to this figure, Madrid and Catalonia are the two provinces that COVID-19 spread radically possible due to the population and density. The majority of the provinces that had only a few cases where protected by the strict measures put in force by the government. For instance, in Spain measures started on the 15^{th} of March, while in other COVID infected countries, similar measures began between 12^{th}-14^{th} of March [5].

3 Methodology

3.1 The SIR model

A widely used mathematical tool for analysing the spread of a virus is the SIR (Susceptible-Infected-Recovered) model. In its simplest form the model is based on a few strong assumptions [6, 7]. First, the number of population is constant for the duration of the analysis, meaning there are no births or deaths. Second, the incubation period of an infected is considered to be instant and third infectivity has a duration similar to the disease [7]. Yet, despite

these assumptions this model can provide good estimates on the evolution of an epidemic.

Briefly, the model consists of three coupled ODE[3], namely $S(t)$ representing the susceptible people, $I(t)$ for the infected and $R(t)$ for the recovered, all as function of time t. Using numerical integration techniques, these equations are solved for a pre-specified period of time, as in (1) - (3) where β and γ are the infection and recovery rate parameters, respectively. The ratio defined by (4) is known as the epidemiological threshold and is a key indicator about the evolution of a disease.

$$\dot{S} = -\beta SI \tag{1}$$

$$\dot{I} = \beta SI - \gamma I \tag{2}$$

$$\dot{R} = \gamma I \tag{3}$$

$$R_0 = \beta/\gamma. \tag{4}$$

3.2 Artificial Intelligence and Deep Neural Networks

Inspired by the human brain cells, DNN are the cornerstone of modern AI. A typical DNN consists of hundreds (or thousands) of neurons grouped in at least four layers: an input, an output and two hidden layers. At each neuron two operations occur: a summation of the weighted neuron inputs and a transformation of that sum through a mapping function. The type and complexity of the problem dictates the selection of the mapping function. Overall, DNN can differ in both size and structure however, all are typically known as universal function approximators due to their ability to solve any possible problem [4].

DNN can be implemented for Supervised, Unsupervised and Reinforced Learning related tasks. Reasonably, the structure will differ depending on the type of learning. Regardless, in the estimation of the COVID-19 peak the problem is formulated as a Supervised Learning task where partially observed SIR curves are labeled by their parameters of their full curves. This, in essence, will try to provide the parameters of an actual case, where the disease is in its first days of spreading and determination of the parameters of a realistic SIR model can be difficult.

To illustrate, in Spain as of March the 28^{th} the number of infected people followed a steep upwards trend, as depicted in Figure 2. According to these dates we extract the corresponding curves along with their parameters from the Monte Carlo SIR model database. The DNN is trained to detect from this partial information the true parameters β and γ. Once trained the data shown in Figure2 are fed into the DNN to deduce the β and γ parameters.

[3]Ordinary Differential Equations

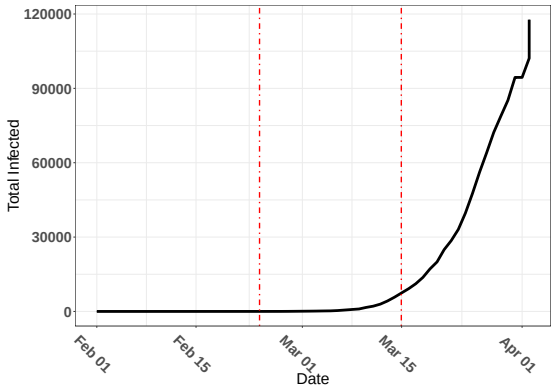

Figure 2: Total Infected Until 28^{th} of March in Spain.

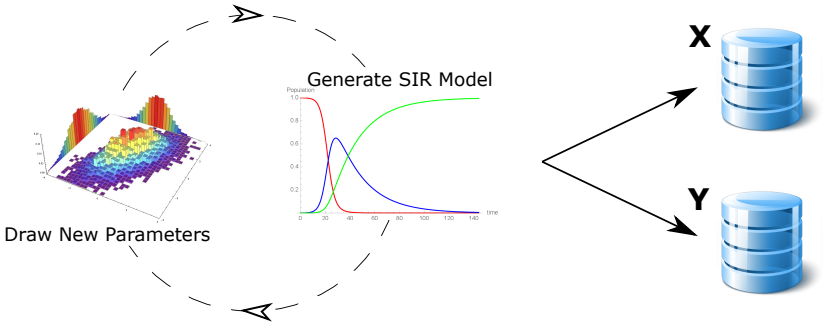

Figure 3: Database Generation

3.3 Monte Carlo Database Generation

Based on the hypothesis that a SIR model can explain the behavior of COVID-19, Monte Carlo simulations are implemented for generating a set of probable SIR models by treating β and γ parameters as random variables. The workflow presented in Figure3 shows the process that generates the database for training and testing the estimation model.

For iteration $m = [1, \cdots, M]$, where M is the total number of iterations, a random set of parameters is drawn, which is used to develop a unique SIR model. Ultimately, the two sets of data consist by the numerical integration of SIR for the infected agents (i.e. $\mathbf{X} \in \mathbb{R}^{M \times 365}$) and their corresponding parameters (i.e. $\mathbf{Y} \in \mathbb{R}^{M \times 3}$). To limit the simulations to generate only relevant data a constraint has been added to maintain beta and gamma within the the estimated range of WHO [1]. That is, the epidemiological threshold is constrained within the open set $1 < R_0 < 4$.

In this work, the number of simulated SIR models are $M = 4 \cdot 10^5$ where a fraction of the resulted infected curves are plotted in Figure 4. The random

selection of β and γ generates the different possible outcomes of the virus given a fixed population.

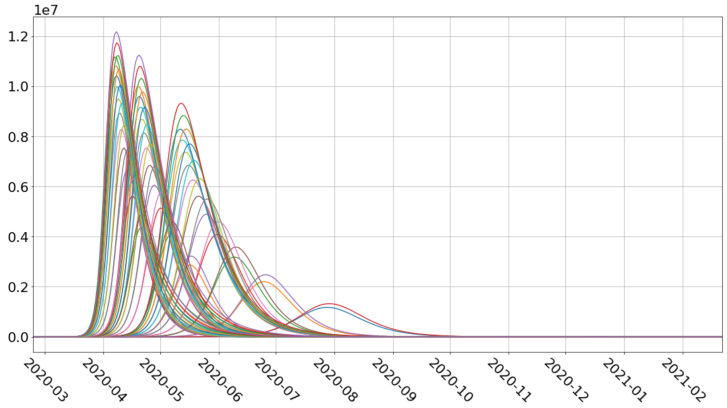

Figure 4: Monte Carlo Database Generation Profiles

4 Results

4.1 DNN

The developed DNN consists of one input layer with 17 neurons (one for each day shown in Figure 2) and an output layer with 3 neurons representing the β, γ and initial population as depicted in Figure 5. The hidden layers consist of 50 neurons and 1 bias neuron each, all with ReLu[4] activation function. The total number of trainable parameters can be found in Table1. The loss function of the DNN is the Mean Absolute Error (MAE), as in (5) where y_i and $\hat{y}_i$ are the true and predicted values, respectively, of the i-th sample. The loss function is minimized using the Adam[3] optimizer, whereas the Mean Absolute Percentage Error (MAPE), as in (6), is used for evaluating the performance of the model. This is because it is more convenient to compare performance between different DNN designs.

$$MAE = (\frac{1}{n}) \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{5}$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{6}$$

The aforementioned DNN model is trained in Python using Tensorflow and Keras backend for 30 epochs using an early stopping callback to prevent overfitting. The training and validation loss are depicted in Figure 6. As it can be

[4]Rectified Linear unit

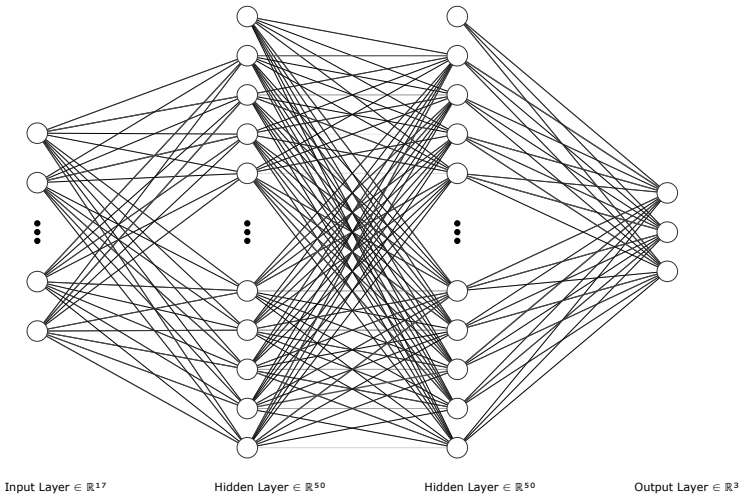

Figure 5: Structure of DNN

Table 1: DNN parameter List per layer

Layer	**Parameters**
Input	0
Hidden 1	900
Hidden 2	2550
Output	153
Total	3603

observed the callback terminated training at epoch 6 and rolled-back the model parameters to epoch 5 where the lowest MAPE is achieved.

4.2 Evaluation on Unseen Cases

A separate test set was kept aside from the development process of the DNN. The reason is that tuning a model according to the validation set can lead to overfitting the model on that set. This will hinder the ability of the model to generalize in actual cases.

Concretely, a random sample is chosen to be used for predicted and plotting the true and estimated SIR model. As shown in Figure 7 the DNN predicts accurately the parameters of the SIR model according to the partial curve of infected people. In fact, the true and estimated parameters are given in Table 2. For the case presented in Figure 7 the overall error is 1.4% whereas the total MAE and MAPE in the test set is 0.0144 and 13.293114%, respectively.

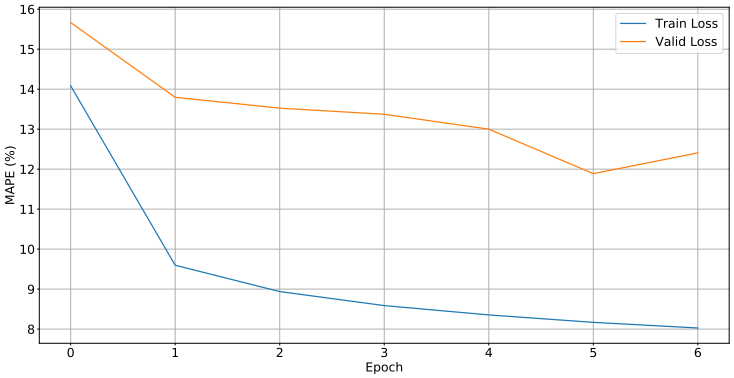

Figure 6: Performance of DNN

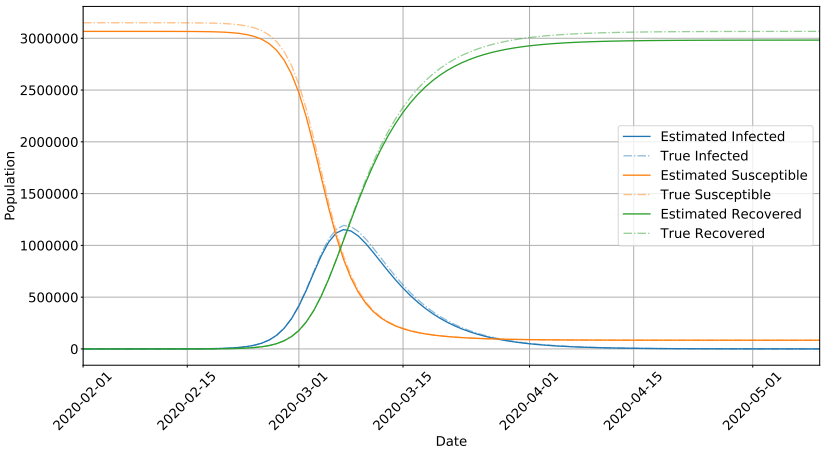

Figure 7: Estimation of SIR parameters on a random unseen case

Table 2: True and predicted parameters of random unseen case

SIR parameters	Predicted	True	Error (%)
β	0.6238	0.6219	0.3
γ	0.168	0.167	1.1
Population (%)	0.068	0.07	2.9

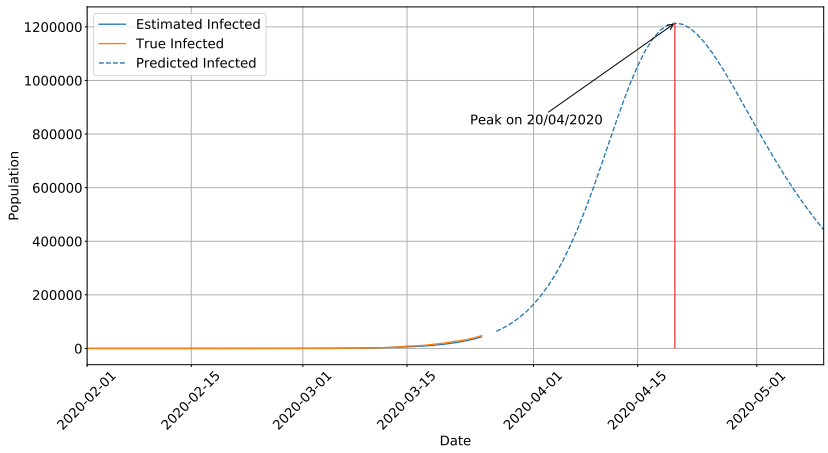

Figure 8: Prediction of DNN on Peak and Total Infected people in Spain with $4.5 \cdot 10^5$ initial population

4.3 Estimated SIR model for Spain

The purpose of the DNN is to predict what the parameters of a SIR model would be for the recorded total infected people over a time span of 17 days. The period that the DNN was developed on is from the 09/03/2020 up to 25/03/2020 because that is the period where the rapid growth of infected people is observed. The results presented in Figure 8 show the predicted evolution of the COVID-19 in Spain given the data up until 25/03/2020. The estimated curve fits quite well with the actual data while the peak of infected people is on the 20/04/2020, i.e. 79 days after the first case of COVID-19 recorded in Spain.

The SIR model is a parameterized model that requires a population to be defined from the beginning. The peak value and peak occurrence are sensitive to the aforementioned parameter. Therefore, using the predicted parameters β and γ and accepting them as true, SIR models for different levels of population reveal that day of the peak ranges from 68 to 91 days after 1^{st} of February i.e. 09/04/2020 − 02/05/2020 as illustrated in Figure 9.

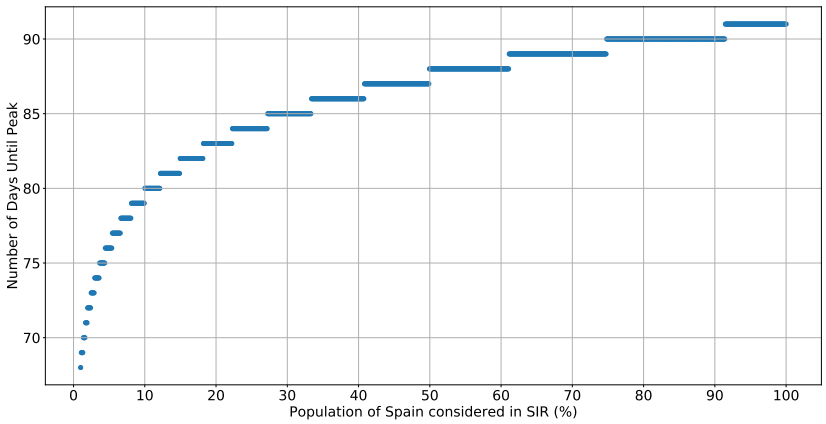

Figure 9: Days until peak as a function of population consider in SIR

5 Conclusions

In this paper, the application of a DNN has been studied for the identification of the parameters of the SIR model that modulates the COVID-19 virus. The DNN is an advanced technique that has made possible to know the parameters of the SIR model that better adapts to the data of Spain (being this the studied case) although the DNN can be trained for the rest of countries as well as of provinces.

The simplicity of the proposed approach with the DNN allows to identify the SIR parameters for different COVID-19 evolution curves what it could help the scientific community to identify curves from different population sizes in contact with the virus. Further studies on COVID-19 evolution curves are required, however in this case of study the model has been able to correctly obtain the SIR model parameters, thus generating a population-dependent model.

References

[1] Statement on the meeting of the international health regulations (2005) emergency committee regarding the outbreak of novel coronavirus 2019 (n-cov) on 23 january 2020.

[2] Gobierno de España.Centro de Coordinación de Alertas y Emergencias Sanitarias. Enfermedad por el coronavirus (covid-19). 28/03/2020.

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[4] Tom Schaul, Dan Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1312–1320. JMLR.org, 2015.

[5] Axel Gandy et. al Seth Flaxman, Swapnil Mishra. Estimating the number of infections and the impact of nonpharmaceutical interventions on COVID-19 in 11 European countries. `https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-Europe-estimates-and-NPI-impact-30-03-2020.pdf`. [Online; accessed 30-March-2020].

[6] Howard Weiss. The sir model and the foundations of public health. 2013.

[7] Eric W. Weisstein. Kermack-McKendrick Model. From MathWorld–A Wolfram Web Resource. `https://mathworld.wolfram.com/Kermack-McKendrickModel.html`. [Online; accessed 30-March-2020].