UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Faculdade de Ciências e Tecnologia
Câmpus de Presidente Prudente

Data-driven mathematical models for assessing the COVID-19: SIRD-type equations

Fabio Vinicius Goes Amaral

Orientador: Prof. Dr. Cássio Machiaveli Oishi

Co-orientador: Prof. Dr. Wallace Correa de Oliveira Casaca

Financiamento: CAPES (Bolsa PICME)

Programa: Matemática Aplicada e Computacional

Presidente Prudente, August 2021

UNIVERSIDADE ESTADUAL PAULISTA

Faculdade de Ciências e Tecnologia de Presidente Prudente

Programa de Pós-Graduação em Matemática Aplicada e Computacional

Data-driven mathematical models for assessing the COVID-19: SIRD-type equations

Fabio Vinicius Goes Amaral

Orientador: Prof. Dr. Cássio Machiaveli Oishi

Dissertação apresentada ao Programa de Pós-Graduação em Matemática Aplicada e Computacional da Faculdade de Ciências e Tecnologia da UNESP para obtenção do título de Mestre em Matemática Aplicada e Computacional.

Presidente Prudente, August 2021

A485d

Amaral, Fabio Vinícius Goes
Data-driven mathematical models for assessing the COVID-19 : SIRD-type equations / Fabio Vinícius Goes Amaral. -- Presidente Prudente, 2021
79 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Cássio Machiaveli Oishi
Coorientador: Wallace Correa de Oliveira Casaca

1. Neural networks (Computer science). 2. Forecasting. 3. Mass vaccination. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Presidente Prudente

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: Data-driven mathematical models for assessing the COVID-19: SIRD-type equations

AUTOR: FABIO VINÍCIUS GÓES AMARAL
ORIENTADOR: CÁSSIO MACHIAVELI OISHI

Aprovado como parte das exigências para obtenção do Título de Mestre em MATEMÁTICA APLICADA E COMPUTACIONAL, pela Comissão Examinadora:

Prof. Dr. CÁSSIO MACHIAVELI OISHI (Participaçao Virtual)
Departamento de Matemática e Computação / Faculdade de Ciencias e Tecnologia de Presidente Prudente

VIDEOCONFERÊNCIA

Dr. ALEXANDRE LOUREIRO MADUREIRA (Participaçao Virtual)
Departamento de Matematica Aplicada e Computacional / Laboratório Nacional de Computação Científica LNCC

VIDEOCONFERÊNCIA

Prof. Dr. JOAO PAULO PAPA (Participaçao Virtual)
Departamento de Computação / Faculdade de Ciências de Bauru

Presidente Prudente, 29 de julho de 2021

Dedico este trabalho à minha família e aos meus amigos

Agradecimentos

Agradeço, primeiramente, a minha família pelo apoio e amor dado durante toda minha vida.

Aos meus amigos, pelos momentos de descontração ao longo da jornada acadêmica, em especial Alex, Diovanni, Giovana, Gustavo Gimenez, Gustavo Lopes, Isabella, João Paulo Sapia, Lucas Foltran, Wendy. Aos demais amigos não citados que sempre se uniram e se ajudaram no aprendizado e formação pessoal.

Aos meus orientadores, professor Cássio e professor Wallace, pela ajuda e conhecimento compartilhado.

Aos professores, que tive durante a vida, que ampliaram meu conhecimento e ajudaram na formação que tenho, em especial aos que se esforçam ao máximo para atrair o interesse aos estudos.

À Cinthia, por todo suporte oferecido durante o programa.

Á UNESP, por ser uma instituição pública, gratuita e de qualidade, que com o tempo se tornou uma segunda casa e ofereceu os recursos e infraestrutura para construção deste trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"Faça o teu melhor, na condição que você tem,
enquanto você não tem condições melhores, para fazer melhor ainda!"
Mário Sergio Cortella

Resumo

São Paulo é o estado mais populoso do Brasil, com de cerca de 22% da população do país. O número total de pessoas infectadas pelo COVID-19 em São Paulo ultrapassou a marca de um milhão em janeiro de 2021, enquanto o número total de óbitos é de cerca de 25% dos óbitos do país. Juntando-se aos esforços acadêmicos brasileiros na luta contra a progressão da COVID-19 no país, um novo método de projeção baseado em dados é proposto, combinando o modelo chamado Suscetível-Infectados-Recuperado-Óbitos com estratégias de aprendizado de máquina para ajustar adequadamente os coeficientes do modelo para predição de Infecções, Recuperações, Óbitos e o número de reprodução do vírus. Mostramos que o preditor resultante é capaz de lidar com dados mal-condicionados enquanto gera projeções acuradas de 10 dias. Nosso sistema integrado pode ser utilizado para guiar ações governamentais principalmente baseados em dois aspectos: avaliação em tempo real dos dados e projeções dinâmicas para as curvas da COVID-19 em diferentes regiões do estado e país. Finalmente, estendemos a metodologia para investigar os efeitos do processo de vacinação utilizando um modelo mais complexo. Em particular, estudos mostram diferentes cenários variando da taxa de vacinação e eficácias de vacinas.

Palavras-Chave: *Redes neurais (Ciência da computação), Previsão, Vacinação em massa.*

Abstract

São Paulo is the most populous state in Brazil, home to around 22% of the country's population. The total number of COVID-19-infected people in São Paulo has reached more than 1 million, while its total death toll stands at 25% of all the country's deceases. Joining the Brazilian academia efforts in the fight against COVID-19, in this work we describe a unified framework for monitoring and forecasting the COVID-19 progress. A novel forecasting data-driven method has been proposed, by combining the so-called Susceptible-Infectious-Recovered-Deceased model with machine learning strategies to properly fit the mathematical model's coefficients for predicting Infections, Recoveries, Deaths, and Viral Reproduction Numbers. We show that the resulting predictor is capable of dealing with badly conditioned data samples while still delivering accurate 10-day predictions. Our integrated computational system can be used for guiding government actions mainly in two basic aspects: real-time data assessment and dynamic predictions of COVID-19 curves for different regions of the state and country. Finally, we extend our methodology to investigate the effects of the vaccination process with a more complex model. In particular, our studies are able to predict different scenarios varying the rate of vaccination and the effectiveness of the vaccines.

Keywords: *Neural networks (Computer science), Forecasting, Mass vaccination.*

List of Figures

List of Tables

List of Simbols

Notation	Description
$S(t)$	number of susceptible at time t
$I_j(t)$	number of infected from subgroup $j = s, v_1, v_2$ at time t
$I(t)$	Sum of all subgroups I_j at time t
$R_j(t)$	number of recovered from subgroup $j = s, v_1, v_2$ at time t
$R(t)$	Sum of all subgroups R_j at time t
$D(t)$	numbers of deaths at time t
$\bar{V}_i(t)$	numbers of vaccinated but not yet immunized at time t with i = 1,2 doses
$V_i(t)$	numbers of immunized at time t with i = 1,2 doses
$\beta(t)$	transient transmission rate
$\beta_{net}(t)$	prediction for the transmission rate at time t
γ_r	rate of recovered
γ_d	rate of mortality
ν_1 and ν_2	first and second dose vaccination rate, respectively
θ_1 and θ_2	first and second dose effectiveness, respectively
α_i	time delay for vaccine dose i = 1,2 effectiveness
$R_t(t)$	time-dependent effective reproduction number
M	pre-specified training period
p	desirable forecast period
Y_i and $\dot{Y}_i$	real and predicted daily values w.r.t. a given target variable

Contents

Chapter 1

Introduction

According to the official report published by the World Health Organization (WHO) [91], up to November 2020, the novel coronavirus infected more than 6 million people in Brazil. While some countries in Europe were facing the second wave of the pandemic, Brazil was suffering the adverse impacts of a COVID-19's lasting wave while still preparing for the arrival of a new wave hitting the country at the end of the year. In particular, in the state of São Paulo, which is the most populous state, holding around 22% of the country's population, the infections reached more than 1 million people [53].

Another disconcerting fact about the COVID-19 situation in the state of São Paulo is that it was accounting for 25% of all deaths in the country [53] in 2020. As a result, the state has been the epicenter of the coronavirus outbreak in Brazil so that it can be compared to other countries, placing São Paulo (until October 2020) in the 5th and 6th positions globally with respect to the confirmed cases and deaths, respectively, thus above countries like Germany, France and UK [92], as one can see in Figure 1.1.

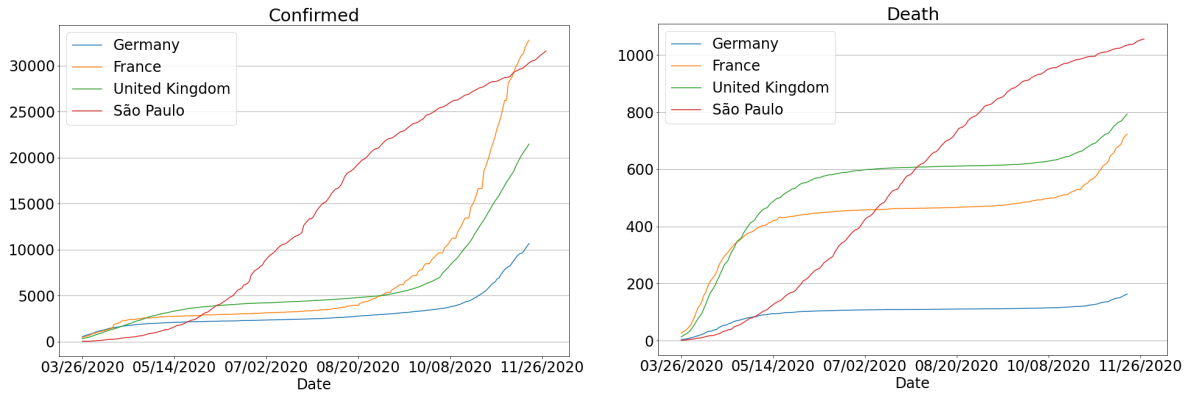


Figure 1.1: Comparison of cumulative number of cases and deaths per million population: São Paulo state, France, Germany and United Kingdom. Countries data from Johns Hopkins University [40].

Massive vaccination campaigns are one of the most effective strategies against the COVID-19 disease. Under this premise, many countries have dedicated a considerable amount of effort in negotiating and administrating different types of COVID-19 immunizers such as Pfizer/BioNTech, Oxford/AstraZeneca and CoronaVac/Sinovac vaccines in order to control the spread of new coronavirus. As a result, the nations which implemented mass vaccination early have seen a significant reduction of their SARS-CoV-2 cases and deaths, as for instance, Israel and UK [52, 25, 40]. However, due to limitations imposed

by the production of vaccine inputs and the global fluctuation of the doses distribution, it is well-known that several countries around the world have faced enormous challenges in trying to control the new waves of COVID-19. This is the case experienced in Brazil, a developing country which has continuously suffered due to the delay in negotiating deals with pharmaceutical vaccine companies [81] and because of the high social inequality currently present in the country [51]. As a consequence, the total number of new cases and deaths significantly increased in 2021, surpassing the worst scenario of previous year [52].

Due to the unclear scenario of COVID-19 pandemic in São Paulo state, the public health system has been dealing with many challenging issues, as those faced by other countries such as the availability of beds in hospitals [74, 46], monitoring of control measures [43, 83, 8], and implementation of immediate mitigation plans [26, 15] to contain the advance of coronavirus. These also include the development of effective data-driven responses such as real-time monitoring systems and forecasting models to track and predict COVID-19 advance on each region of the state, even under real-world circumstances that are hard to handle in practice. For example, a drastic reduction of data updates for a few days – because of a delay in making COVID-19 test results public – as well as a retroactive data refresh due to inconsistencies occurred when managing multiple data sources, can result in a poorly-trained model with high chances of failure when fitting COVID-19 data. Moreover, the forecasts strongly reflect the accuracy of the collected data, as notifications are usually recorded by date of disease confirmation rather than date of occurrence. In fact, more realistic predictions depend on successive updates with accurate data in order to be effective [6, 93, 48]. Therefore, the first goal of this work is to address the issue of inaccurate/delayed data for predicting 10-days COVID-19 curves with a satisfactory level of accuracy.

Oftem another issue when extrapolating epidemiological data is that the model's parameters are assumed to be constant, e.g., transmission rate and rate of recovery, as typically taken by classic Susceptible-Infectious-Recovered (SIR) based approaches [87, 11]. Despite the existence of very effective SIR variants which takes the model's parameters as constant [7], their calibration when concomitantly assessing a great variety of regions with distinct traits is not a straightforward task, since some of the tunable values depend on local government-regulated measures, difficult to estimate in practice, especially in the Brazilian context. In order to circumvent the parameter issue of classic SIR-derived methods while still allowing the mathematical model to cope with time-varying coefficients, the use of Machine Learning strategies has been a popular choice and a trend. Indeed, recent developments involving variable-parameter SIR variants to assess the course of COVID-19 can be found in [39, 20, 94, 80, 42, 89, 86, 10, 9, 88, 36], which include the use of effective Artificial Intelligence (AI) strategies, as for example, in [57, 23, 14, 39, 20, 56, 61]. Following these recent efforts in modeling COVID-19 dynamics from epidemic models tuned with learning mechanisms, in this dissertation we propose an effective, data-driven SIR model whose parameters are calibrated by temporal functions, learned from individual regressors and trained on different data sources. The predictions are obtained using a time-dependent SIR-based model [41] coupled with an intelligent architecture that learns the model's parameters for each one of the regions analyzed in our study.

A important advantage of our data-driven approach is that it only assumes as input the raw data of infected, recovered and deaths to produce the definitive forecasts. In fact, the current learning scheme does not require any prior knowledge of specific time-series such as the transmission rate curve. Another relevant aspect to be observed is that the designed technique does not impose any particular probability distribution to the epidemiological curves, thus avoiding the use of pre-fixed forms of data distribution such as exponential solutions and logistic regression-based models.

In regard of the vaccination assessment, a financially inexpensive yet effective way to measure the success of vaccination from a dynamic point-of-view is by applying SIR variation models model. Mathematically speaking, additional Ordinary Differential Equations (ODE) can be included as part of the full SIR formulation to model the number of vaccinated individuals and their roles in the resulting dynamic system. This is the way adopted by many researchers to simulate and understand the impacts of previous pandemics. For instance, Alexander et al. [2] investigated the spread of influenza from the so-called Susceptible-Vaccinated-Infected-Recovered-Susceptible (SVIRS) model, while Counotte et al. [55] studied the epidemics of Zika virus by rearranging the SIR with vaccination. Other interesting mathematical advances on SIR-based models with vaccination were also given and discussed by Sun and Hsieh [79], Sinha et al. [77], and Mathur and Narayan [50].

Most recently, modified SIR-type models with vaccination compartments have been employed for dealing with the COVID-19 pandemic. The so-called SEIRD model was employed by Roy et al. [70] to investigate the best vaccine allocation strategy in the New York State, while Fudolig and Howard [32] introduced a multi-strain version of SIR model so as to verify some variations for the reproduction number of the disease. Based on the advance of the SARS-CoV-2 vaccines, Saad-Roy et al. [73] presented a very robust investigation concerning the discussions on single or double doses of immunizers, including analyzes of different scenarios in terms of transmission and vaccine immunity. Similarly, Harizi et al. [13] applies the SIRV model to study the dynamic behavior of COVID-19 spreading using various daily vaccination rates and vaccine efficacy in Canada. An alternative SIR-type model covering impulsive vaccination strategies has been discussed by Etxeberria-Etxaniz et al. [27], while the mass vaccination in Greece was investigated by Rachaniotis et al. [64]. The vaccination process was also considered as part of a feedback immunization control rule in the model studied by De la Sen et al. [24], while Mak et al. [47] assume a two-dose vaccine model to explore various vaccine rollout policies.

In order to carry out such a complete assessment of vaccination in Brazil, we take advantage of the presented data-driven approach to model the vaccination dynamic under a two-dose regime. As shown in the experiments, the current methodology succeeds in capturing the trend of epidemiological curves in well-behaved data such as the Israel one, as well as in particular scenarios wherein the data is ill conditioned, as observed in the Brazilian context. We also assess the outcome from the CoronaVac/Sinovac vaccine in the adult population of Serrana's town, in São Paulo State – Brazil, being this numerical analysis another important contribution in the context of this work. It is worth mentioning that more recently, Serrana becomes the world's first fully vaccinated town as part of *Project S* [82], a Brazilian study conducted by *Instituto Butantan* [37] to answer important issues regarding the mass vaccination of people with two-dose shots of Coronavac/Sinovac vaccine. Finally, this work can be extended for further investigation concerning COVID-19 vaccine mixing, i.e., the proposed methodology allows to numerically explore the recent idea of mix-and-match vaccination strategies [75, 18].

Research Questions The initial questions that results this work was:

- Can an ANN estimate the unknown behavior of the transmission rate?
- How can an ANN adapt various models for assessing different aspects of the pandemics?

Contributions The main contributions of this work can be summarized as follows:

- The implementation of urgent responses, as listed below, to mitigate the progress of coronavirus in São Paulo state, which is the most populous and economically active state in Brazil, responsible for 34% of the Brazilian GDP [90].

- A novel forecasting model which combines the simplicity of SIR-based formulation with the effectiveness of data-driven learning strategies, for predicting COVID-19 cases, deaths, recoveries and the virus reproduction number. The designed method is also capable of addressing "the curse of delay", as usually observed in the Brazilian reports of cases and deaths, determining whether or not a coronavirus-related time-series period is "well-posed".

- Our predictive approach learns the epidemiological parameters as time-dependent functions, which are calibrated by a recursive training approach based on Artificial Neural Network, therefore allowing the forecaster to fit and customize COVID-19 curves for each region of the state.

- The *SP COVID-19 Info Tracker* is COVID-19 data repository and a freely-available online platform firstly developed by Wallace Casaca and Marilaine Colnago, which has been accessed by citizens, authorities and media agencies to track and inspect the COVID-19 progress in São Paulo state. New COVID-19 notifications are immediately available throughout the platform, by getting fresh data published daily by 92 city halls spread over the state (the so-called first-hand local sources), in an attempt to reduce the delay in reporting the new cases and deaths as often observed in the Brazilian government updates [68, 31]. This work helped improve the platform with forecast contributions using its data.

- A method for assessing the impact of the vaccination process, considering various scenarios, for individual regions and countries.

The dissertation is organized as follows: Chapter 2 presents a introductory explanation of neural network operation.Champter 3 presents the methodology. Section 3.1 introduces the problem description and the mathematical design of our data-driven epidemiological model, and Subsection 3.2 describes the details of the proposed training apparatus to learn the model's parameters and 3.3 extends the model for vaccination analysis. Next, Section 4 brings the validation study of our approach and numerical experiments with real data focused on São Paulo state and other Brazilian regions. Experiments involving other countries are also given. Finally, Section 5 summarizes our findings, observed conclusions and future work, while in Appendix, we present our online tracking platform used by decision-makers and other interested people in interacting, auditing and navigating coronavirus-related data in the state of São Paulo.

This work resulted in two papers, entitled "Towards Providing Effective Data-Driven Responses to Predict the COVID-19 in São Paulo and Brazil" [5] and "Simulating Immunization Campaigns and Vaccine Protection Against COVID-19 Pandemic in Brazil" [4]

Chapter 2

Basic Concepts on Neural Networks

In order to understand the method applied, we first present an introduction to machine learning and neural networks. According to [34] "Machine learning is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions". There are many learning algorithms and strategies and most of them can be divided into two categories: supervised and unsupervised learning. Those categories are defined by the experience with the dataset during the leaning process, also called training. The proposed strategy fits the supervised case, since the expected results of the training data are known

2.1 Artificial Neural Network

The deep learning development were known with other names since 1940s. Some of the earliest learning algorithms were intended to be models of biological learning, thus some common names as neurons are used. Those models are designed to, given a input $x_0, ..., x_n$, associate an output y.

2.1.1 General Idea

In an Artificial Neural Network (ANN), a neuron is a component that stores an activation value. This value comes from the interaction with others neurons. A layer is a group of neurons and in a dense feed forward network, as the one used in this work, the layers are sequential and fully connected, i.e, each neuron is connected with all neurons of the next layer.

Each connection has a weight and the activation of the neuron are given the connections with the previous layers. There are 3 types of layer: Input layer, Output layer and Hidden layer.

- **Input layer** As the name suggests, is the layer where the neuron activation comes from the input data. is where the feed forward process begin.
- **Output layer** The last layer of the network, where the results are presented. In some cases, as the classification problem for example, the result are given by the neuron with the highest activation and in our case this layer has only one neuron and its activation value is used.

- **Hidden layers** The middle part between the input and output layer. The name comes from the fact that their activation values are hidden from the user, only used as input for the next layer

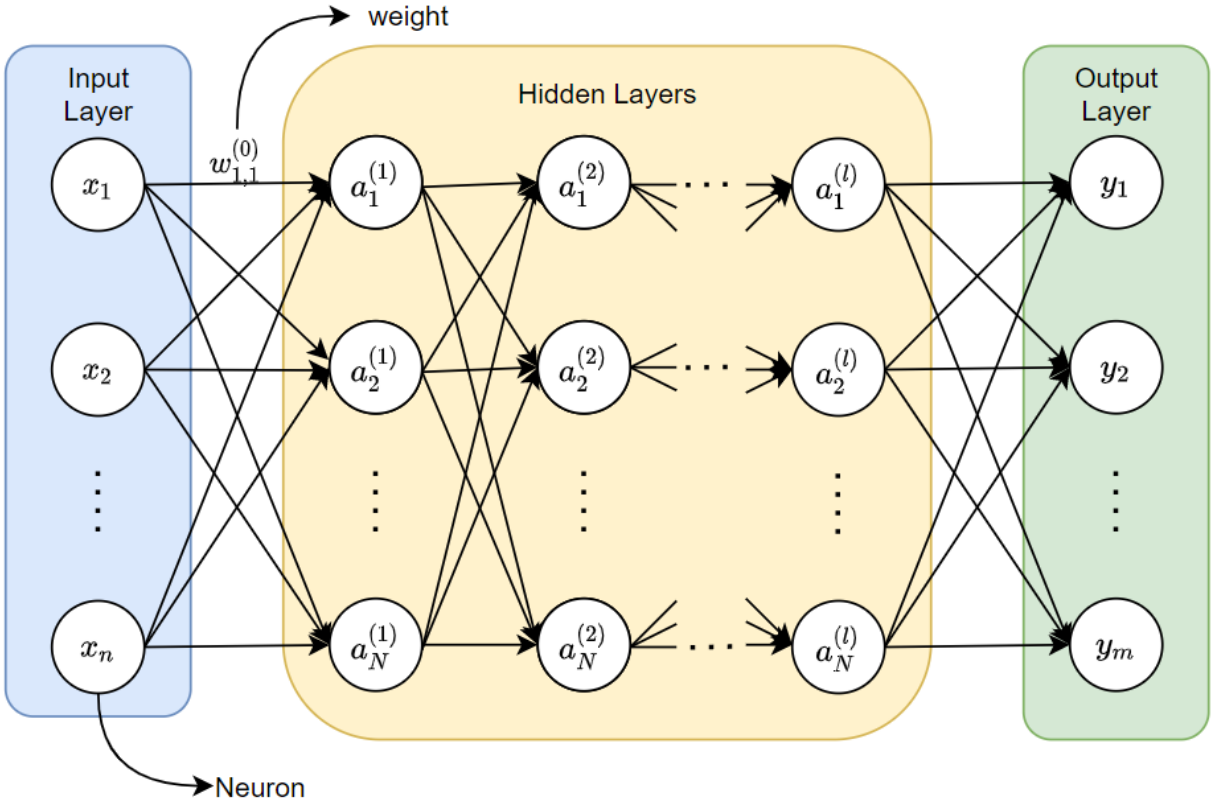


Figure 2.1: Illustration of a generic Artificial Neural Network

Figure 2.1 shows a common illustration of a feed forward ANN. The number of neurons on the input and output layers depends on the data and task intended for the network, but the number of hidden layers and the number of neurons in each layers can be chosen based the inputs, outputs and the complexity of the task [1]. Its very common to choose the structure of the network based in empirical tests.

Each connection between neurons has a weight that determines the importance of the connection. The activation of the neuron comes from a function, called activation function, where the input of this function is the weight sum of each neuron of previous layers connected to the neuron. Some popular activation functions are:

- **Sigmoid**: $\sigma(x) = \frac{1}{1+e^{-x}}$
- **Hyperbolic Tangent**: $\tanh(x) = \frac{2}{1+e^{-2x}} - 1$
- **Rectified Linear Unit**: $ReLU(x) = \max(0, x)$

The training process goal is to find the optimal values of those weights so that given a input the ANN gets a expected output value. In many cases, the neuron activation cannot be optimized only depending on the input and a value that regardless of the input. This value is called bias. The bias can be represented as an extra neuron on the previous layer that has its activation value set to be always 1. then, the bias factor on the activation is given by the weight of the connection with the bias neuron.

2.1.2 Mathematical Notation

The mathematical expression of a neural network evaluation can be written as follows: First, we define a ANN with the following hyper parameters:

- n Neurons on the input layer;
- m Neurons on the output layer;
- l Hidden Layers;
- N neurons in each hidden layer.

The same number of neurons on the hidden layers on this example was chosen for simplicity, in practice the number of neurons may change on each hidden layer.

We denote the i^{th} neuron activation value of the k^{th} hidden layer as $a_i^{(k)}$. The value $w_{i,j}^{(k)}$ represents the weight of the connection of the i^{th} neuron of the k^{th} layer with the j^{th} neuron with the $(k+1)^{\text{th}}$ layer. The input and output layers have the superscript equals 0 and $l+1$ respectively. Then we have:

$$\begin{aligned}
a_i^{(1)} &= f\left(\sum_{j=1}^{n} w_{i,j}^{(0)} x_j + b_i^{(0)}\right),\ i = 1, ...,\ N; \\
a_i^{(k+1)} &= f\left(\sum_{j=1}^{N} w_{i,j}^{(k)} a_j^{(k)} + b_i^{(k)}\right),\ i = 1, ...,\ N; k = 1, ..., l-1; \\
y_i = a_i^{(l+1)} &= f\left(\sum_{j=1}^{N} w_{i,j}^{(l)} a_j^{(l)} + b_i^{(l)}\right),\ i = 1, ..., m,
\end{aligned} \tag{2.1}$$

where f is the chosen activation function and $b_i^{(k)}$ is the neuron bias. Different layers may have different activation functions, thus one can use f_k notation for the activation function.

A compact way to write (2.1) is to adopt a matrix notation. Using a column vector $x^{(k)}$ to represent a layer, where each position has the activation value of a neuron ($x^{(k)} = \{a_i^{(k)}\}$), we have:

$$x^{(k+1)} = f_k\left(W^{(k)} x^{(k)} + b^{(k)}\right), k = 0, ..., l, \tag{2.2}$$

where $W^{(k)} = \{w_{i,j}^{(k)}\}$ is a matrix and $b^{(k)} = \{b_i^{(k)}\}$ is a column vector.

The training process to obtain the values of the weights $W^{(k)}$ and bias $b^{(k)}$ is an optimization problem. In supervised training, we set a loss function, also known as objective function, that evaluate the error of the network compared with the expected result. Then, we use an optimization process to minimize the loss function with respect to the weights and bias of a set of inputs called training dataset.

Since we expect the results to be used in different inputs of the training set (generalization of the network) some implementations use a test dataset to check for over-fitting, that is when the network results perform better for the training data but poorly for others inputs.

There are many factors that determines the network structure, such as the number of neurons, number of layers, activation functions and the need of bias. This factors are called hyper parameters. The choice of the hyper parameters are highly dependent of the problem that the network are solving.

2.1.3 Back-propagation

Most optimization process rely on the gradient of the function. Computing the analytical expression for the gradient of an ANN can be straightforward but often expensive. The back-propagation algorithm [72, 71] compute the gradient by flow the cost information backwards through the network. Once the gradient is numerically obtained other algorithm such as gradient descent or Limited Memory Broyden—Fletcher—Goldfarb–Shanno (L-BFGS), for example, is used to perform the training process using this gradient.

The back propagation algorithm is an application of the chain rule from calculus. Let x and y be real scalars and f and g be function that maps $\mathbb{R}$ to $\mathbb{R}$. Suppose that $y = g(x)$ and $z = f(y)$. If we want to calculate $\dfrac{\partial z}{\partial x}$, by the chain rule we have:

$$\frac{\partial w}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}. \tag{2.3}$$

We can generalize the rule beyond the scalar case. Let $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$, $g : \mathbb{R}^m \to \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$, then the chain rule can be written as:

$$\nabla_{\mathbf{x}} z = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^T \nabla_{\mathbf{y}} z \tag{2.4}$$

where $\nabla_{\mathbf{x}} z$ and $\nabla_{\mathbf{y}} z$ is the gradient of z with respect with $\mathbf{x}$ and $\mathbf{y}$, respectively, and $\dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}$ is the Jacobian matrix of g.

The rule can even be extended for multidimensional tensors. Like the vector notation, we can write the gradient of z with respect to the tensor $\mathbf{X}$ as $\nabla_{\mathbf{X}} z$. Using a variable i to notate the tuple index of the tensor, we can write $(\nabla_{\mathbf{X}} z)_i = \dfrac{\partial z}{\partial X_i}$ just like the vector notation $(\nabla_{\mathbf{x}} z)_i = \dfrac{\partial z}{\partial x_i}$. Then, taking the tensor $\mathbf{Y} = g(\mathbf{X})$ and $z = f(\mathbf{Y})$, the chain rule for tensors is given by:

$$\nabla_{\mathbf{X}} z = \sum_j \left(\nabla_{\mathbf{X}} Y_j\right) \frac{\partial z}{\partial Y_j}. \tag{2.5}$$

Since neurons activation values are obtained by matrix multiplication, one can apply the chain rule recursively through the network layers in order to calculate the required loss function gradient for each network neuron. For each step, the neuron just need to know how to calculate the gradient of its activation function to propagate it for the parents neurons apply the chain rule.

The Figure 2.2 illustrate the gradient computation of the loss function with respect with an output neuron. The simplified illustration shows the method via computational graphs and the process can be done recursively through the network for all neurons.

The black arrows in Figure 2.2 shows the flow of calculations on the feed forward process, until the point the loss is calculated. The red arrows represents the back-propagation of the gradient evaluation. In this step, the derivative of each operation are known. neural network libraries has the implementation of the gradient calculation for the most common activation functions or have tools for the user inform how to do this calculation.

Note that the process can be done with a mine batch of inputs or full batch, for the stochastic gradient descent method, for example, by just considering the input variable as an tensor.

The main advantage of the back-propagation algorithm is the reduced computation of repeated operation. Once a gradient is calculated, it value can be stored and used

for multiple parent nodes on the chain rule. One can exchange the storage cost with computational cost re-evaluating some nodes on the graph.

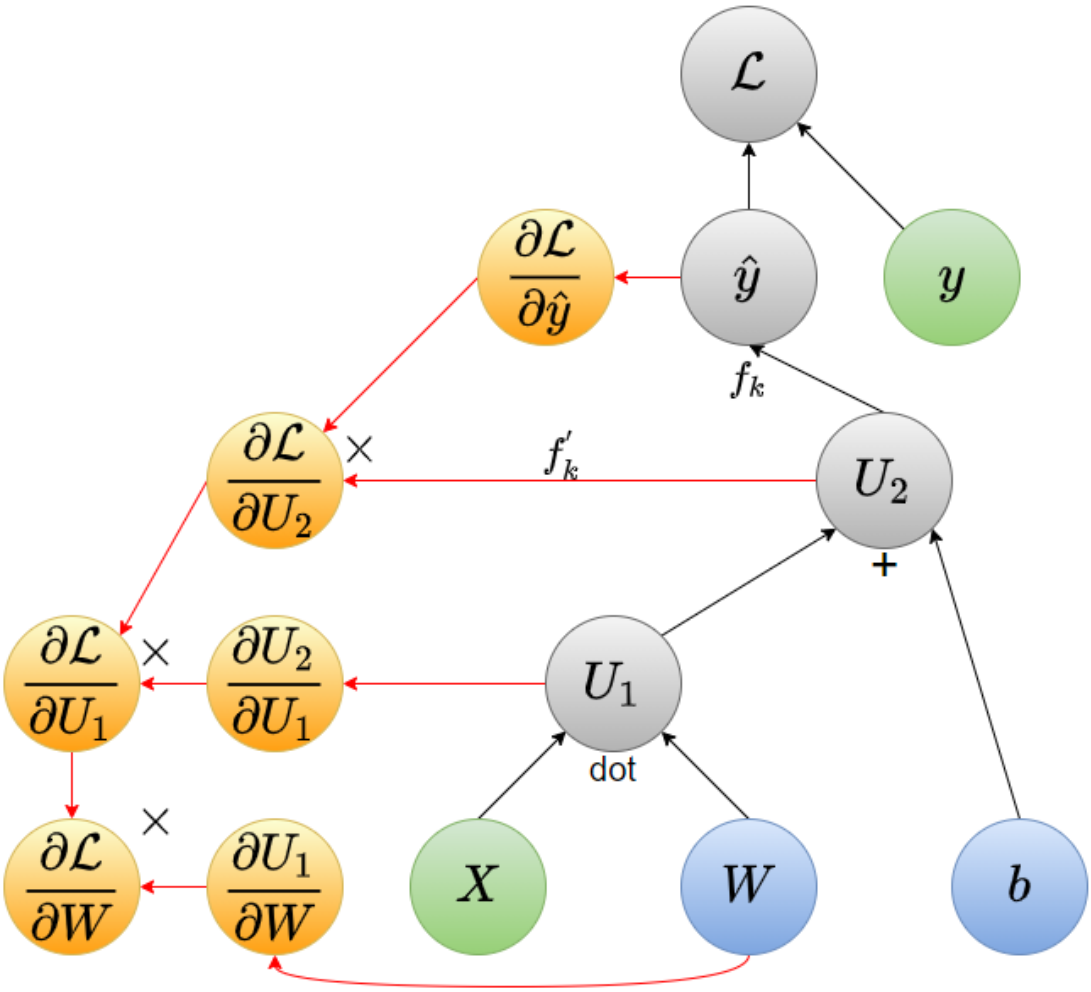


Figure 2.2: Operation graph for evaluating the activation of a neuron and the back-propagation process of evaluating $\frac{\partial \mathcal{L}}{\partial W}$. Each node represents a variable. The directed edges represents operations. The black edges represents the feed forward process and the red arrows represents the back-propagation process.The green nodes represents the input variables, which may be an vector, scalar or tensors and the blue nodes represents the network weights. The *dot* operation represents the matrix-vector multiplication. The trivial operation $\frac{\partial \mathcal{L}}{\partial \mathcal{L}}$ is omitted.

Chapter 3

Mathematical Modeling and Numerical Approach

3.1 A Time-Dependent SIR-based Model

In this section, we present the mathematical design for the proposed data-driven epidemiological model for forecasting COVID-19 trends.

Let N be the size of the total population we intend to model. The classical SIR model [41] is given by the following system of Ordinary Differential Equations (ODE):

$$\begin{aligned} \frac{dS}{dt} &= -\beta\frac{SI}{N}, \\ \frac{dI}{dt} &= \beta\frac{SI}{N} - \gamma I, \\ \frac{dR}{dt} &= \gamma I, \end{aligned} \tag{3.1}$$

where $S = S(t)$, $I = I(t)$ and $R = R(t)$ are the *numbers of susceptible*, *infected* and *recovered individuals*, respectively, as the time t varies. The canonical form of SIR modeling assumes $N = S(t) + I(t) + R(t)$, while the *transmission rate* β and the *rate of recovery* γ are taken as real constants. The so-called *basic reproduction number* R_0, which is one of the key metrics in epidemiology, is defined by $R_0 = \frac{\beta}{\gamma}$ [35, 80].

In our mathematical approach, we introduce a new population group $D(t)$, to represent the *total number of infected people who died*. A normalized total population, $N = 1$, is also taken in the ODE system (3.1) so that the resulting modified SIR model, namely *Susceptible-Infectious-Recovered-Deceased* (SIRD) [58], is derived:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI, \\ \frac{dI}{dt} &= \beta SI - (\gamma_r + \gamma_d)I, \\ \frac{dR}{dt} &= \gamma_r I, \\ \frac{dD}{dt} &= \gamma_d I. \end{aligned} \tag{3.2}$$

where the parameters γ_r and γ_d account for the *rates of recovered* and *lethality*, respectively. In our formulation, we assume that the transmission rate has a transient trajectory,

i.e., $\beta = \beta(t)$. As a consequence, we get a time-dependent reproduction number on the form:

$$R_t(t) = \frac{\beta(t)}{(\gamma_r + \gamma_d)} S. \tag{3.3}$$

The so-called *effective reproduction number* $R_t(t)$ is an important epidemiological metric that quantifies the average number of new infections arising from a primary infected individual in the population [21, 10]. In practice, R_t measures the COVID-19 spread rate, and it changes as either the individuals gain immunity or die. The ODE system (3.2) with $\beta = \beta(t)$ is also known as *variable coefficient Susceptible-Infected-Removal* (vSIR) [80], *time-varying SIR epidemic* [42], or simply as *time-dependent SIR model* [20, 89].

The Differential Equations system (3.2) is numerically solved for S, I, R and D from a given set of initial condition values, S_0, I_0, R_0 and D_0, producing the numerical solutions $\tilde{S} = S(t_n)$, $\tilde{I} = I(t_n)$, $\tilde{R} = R(t_n)$ and $\tilde{D} = D(t_n)$ for a discretized time t_n with a fixed time step Δt. To do so, we run the *Livermore Solver of Ordinary Differential Equations with Automatic Method Switching* (LSDOA) [60]. Actually, the ODE system (3.2) is recurrently solved as part of an integrated training pipeline, which learns the model's parameters according to data signature of each state region, as we will discuss below.

3.2 Learning Epidemiological Parameters: An Integrated Data-Driven Approach

In this section, we describe our hybrid machine learning pipeline to fit the epidemiological parameters $\beta(t), \gamma_r$ and γ_d by recursively refining the solution of the ODE system (3.2). The proposed learning scheme relies on the solution of an inverse problem, given in terms of the SIRD model (3.2) coupled with an Artificial Neural Network (ANN) to learn from the COVID-19 data, the infected, recovered and deceased cases, denoted here as I, R and D, respectively. Since most given data concern the total infections and total deceases, the values of I and R often need to be estimated. We use the method present in [62], where we take C the total confirmed cases data and $T = 21$ the approximated recovery period of the infection. Then we have $R(t) = C(t-T) - D(t-T)$ and $I(t) = C(t) - D(t) - R(t)$. The unified ANN architecture with SIRD model is illustrated in Figure 3.1.

We construct an ANN to predict the values of $\beta(t)$ for each discrete time t_n, generating a full time-varying curve $\beta_{net}(t)$ The proposed ANN architecture is composed of a hidden layer, containing 10 neurons, and the Sigmoid kernel as the network activation function. The output layer is fully connected to the hidden layer thorough a single neuron with no bias weights, wherein the ReLU is taken to trigger the neuron. As the loss function, we minimize the following aggregated error measure, given in terms of the model's variables I, R and D:

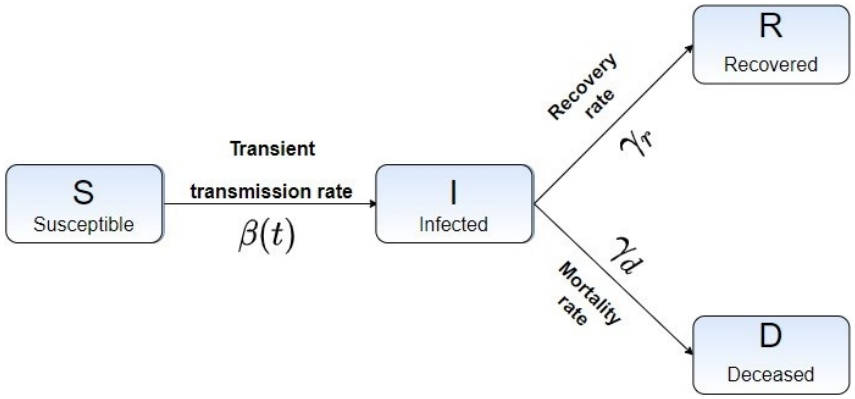
$$\mathcal{L}(\beta_{net}(t), \gamma_r, \gamma_d) = l_I + l_R + l_D \tag{3.4}$$

where:

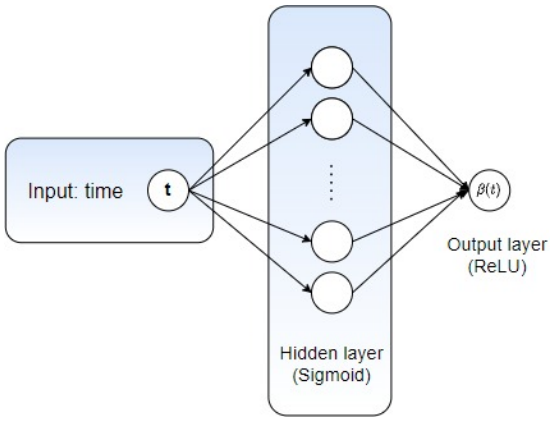
$$l_I = \frac{1}{M} \sum_{i=0}^{M} \left| \log(I_i) - \log(\tilde{I}_i) \right|^2, \tag{3.5}$$

$$l_R = \frac{1}{M} \sum_{i=0}^{M} \left| \log(R_i) - \log(\tilde{R}_i) \right|^2, \tag{3.6}$$

$$l_D = \frac{1}{M} \sum_{i=0}^{M} \left| \log(D_i) - \log(\tilde{D}_i) \right|^2. \tag{3.7}$$



(a) SIRD Model



(b) ANN for $\beta(t)$

Figure 3.1: (a) SIRD model with its corresponding parameters and (b) the ANN design for learning $\beta(t)$.

In Equations (3.5)-(3.7), M is the pre-specified training period, the notation X_i represents the value of $X(t_i)$ for the respective variable I, R or D and the log operator has the role of improving training performance regardless of data scalability, since the analyzed dataset is normalized before running the learning process.

The SIRD's parameters $\beta_{net}(t)$, γ_r and γ_d are predicted by solving the following ANN-related optimization problem:

$$\underset{W,b,\gamma_r,\gamma_d}{\arg\min}\ \mathcal{L}(\beta_{net}(t), \gamma_r, \gamma_d). \tag{3.8}$$

In the proposed learning formulation, the trained parameters are the set of the ANN weights $\{W, b\}$, and the outputs are, the time-varying function $\beta_{net}(t)$, and the epidemiological parameters γ_r and γ_d. In our tests, we solve the ANN optimization problem (3.8) by running the *Limited Memory Broyden—Fletcher—Goldfarb-–Shanno* (L-BFGS) algorithm [16] Notice that the use of ANN instead of any other particular data-driven approach relies on two basic aspects: (i) the effectiveness of neural network design in learning trends and patterns from time-series data, and (ii) the success of more recent works on applying ANN to forecast COVID-19 epidemiological curves, as for example, in [38, 3]. Algorithm 3.1 presents the implementation of the optimization process, with the required data input while Algorithm 3.2 highlights how the numerical solution of the mathematical is used in the loss function.

Once the epidemiological parameters are estimated via the neural network architecture, we solve the ODE-SIRD system (3.2) for the infected, recovered and deceased cases so that a recursive learning procedure is established. More precisely, the loss function

```
def train(recovered, death, inf,vac1, vac2):
    #Normalization
    recovered = recovered / self.norm_fat   #Cumulative recovered
    death     = death / self.norm_fat       #Cumulative deceased
    inf       = inf / self.norm_fat         #Active cases data
    vac1      = vac1 / self.norm_fat        #1 dose data
    vac2      = vac2/ self.norm_fat         #2 doses data

    # Options for the optimizer
    worked = False
    eps = 1e-9 #Grad. Precision
    options_lbfgsb={'disp': None,
    'maxcor': 10,
    'ftol': 2.220446049250313e-09,
    'gtol': 1e-05,
    'eps': eps, # Only change, other config are deafault
    'maxfun': 15000,
    'maxiter': 15000,
    'iprint': - 1,
    'maxls': 20,
    'finite_diff_rel_step': None}
    while not worked:
    # loop to train util convergence (changing grad. precision)
        optimalGamma = minimize(self.model.loss, self.params_calibration
    ,
                                args=([inf, vac1, vac2, recovered, death
    ],
                                self.net_beta, self.other_param),
                                method='L-BFGS-B', tol=1e-13,
                                options = options_lbfgsb)
        eps /= 10
        worked = optimalGamma.success
        if eps < 10e-15: #Grad precision limit
            break
    self.worked = optimalGamma.success
    self.params_calibration = optimalGamma.x

    n = self.net_beta.get_num_param()
    self.net_beta.set_weights(self.params_calibration[:n])
```

Algorithm 3.1: Train Overview

$\mathcal{L}(\beta_{net}(t), \gamma_r, \gamma_d)$ is re-evaluated for the current set of SIRD parameters, and both the numerical resolution of Equation (3.2) and the training scheme (3.8) are repeated until the loss function reaches a minimum. Both *L-BFGS* and *LSODA* are implemented in the *scipy* Python package. Figure 3.2 illustrates this step.

3.2.1 Improving Data Fitting Robustness and Accuracy

As our pipeline makes use of fresh data to learn the parameters, the ill-posedness of a few city datasets in certain time intervals may affect the training task, especially when M varies in ODE-SIRD system (3.2). A low value for M may cause the training to ignore past events, thus overfitting the most recent diseases occurrences. On the other hand, taking a large value for M can lead the mathematical model to drop fresher information. Additionally, as the data is recurrently updated, there is no straightforward way to detect these particular badly-conditioned sub-intervals over the full time-series.

```
def loss(self, point, data, net, params):
    """
    Parameters:
        point: array, list,iterable
            parameters for calibration
        data: array, list,iterable
            Training data
        net: Object
            Object to run the ANN for evaluate beta(t)
        params: array, list,iterable
            Given model parameters
    """
    size = len(data[0])
    #LSODA Solving
    solution = solve_ivp(self.model, [0, size], self.val_0,
                         t_eval=np.arange(0, size, 1), vectorized=True,
   method='LSODA', args=(point, net, params))

    try:
        t = [int(t) for t in solution.t]
        y = solution.y
        I = self.get_infected(y)
        R = self.get_rec(y)
        D = self.get_death(y)
        V1 = self.get_vac1(y)
        V2 = self.get_vac2(y)

        model_vars = [I,V1, V2, R, D]
        l = 0
        for i,d in enumerate(model_vars):
            if d is not None: # If no vacc. model check
                l += (np.mean(((np.log(d) - np.log(data[i][t].astype('
   float32')))) ** 2))

    except Exception as e: #Error check
        print(str(e))
    return l
```

Algorithm 3.2: Loss calculation

In order to improve data fitting of ill-behaved data portions while preserving the epidemiological traits of SIRD modeling, we have adopted a moving window-based strategy to balance the contributions for the forecasted variables over different training intervals. More precisely, our approach takes the following steps:

1. Compute training's outputs for several time windows, by repeatedly solving the ODE-SIRD system (3.2) for $M = M_i \in \{M_1, M_2, ..., M_n\}$, where $M_1 = 10, M_2 = 11, ..., M_n = 30$ days, calibrating the net weights, bias, and parameters γ_r and γ_d for different simulation intervals.

2. Once the set of epidemiological curves $\Lambda = \{C_i : C_i = \{I_i(t), D_i(t), R_i(t)\}\}$ is obtained, we compute the Mean Absolute Percentage Error (MAPE) (4.1) – taken here as an error assessment metric – to decide whether or not a subset of $C_i's$ from Λ is classified as "outlier", i.e., a badly-conditioned time-series period whose epidemiological variables $I_i(t)$, $D_i(t)$, $R_i(t)$ and $R_t(t)$ highly diverge from other evaluated periods. In our tests, we discard the ill-behaved C_i's whose MAPE errors are greater than 20% for any of the variables $I_i(t)$, $D_i(t)$ or $R_i(t)$.

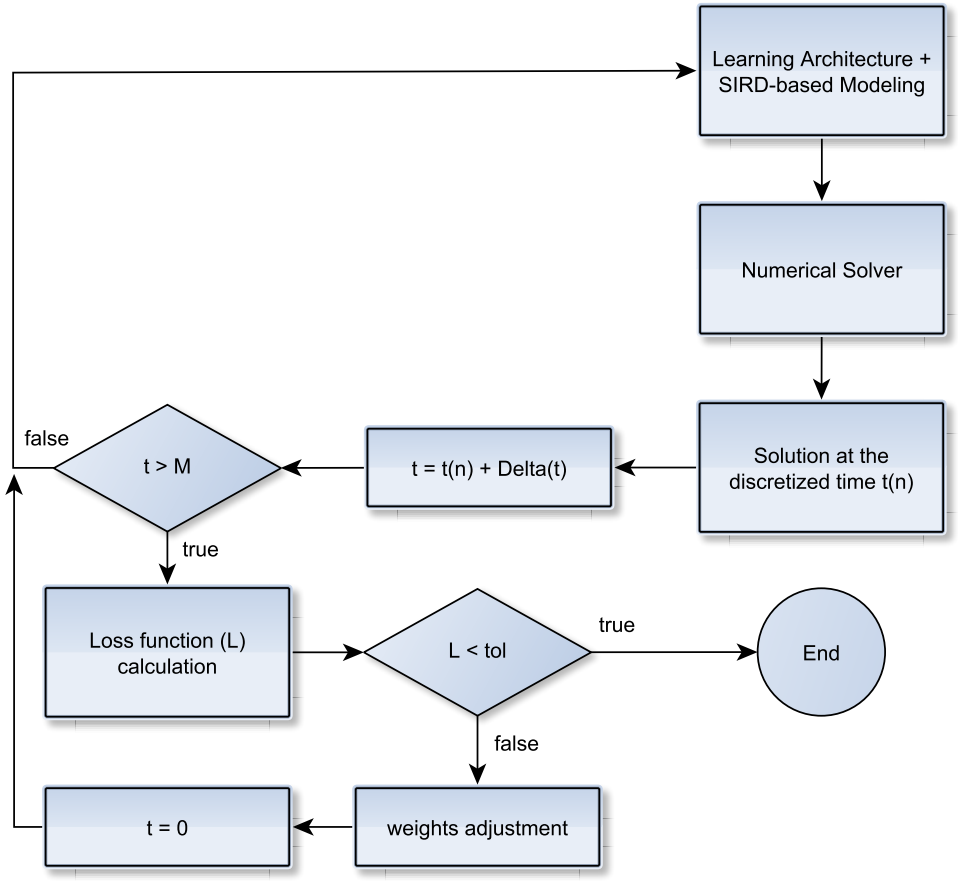
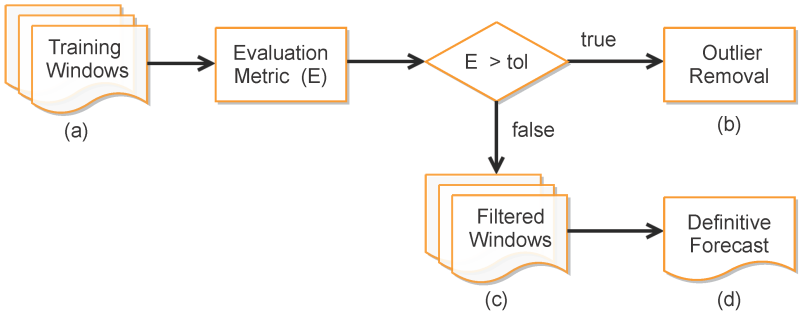


Figure 3.2: Illustration sketch for the parameter calibration step.

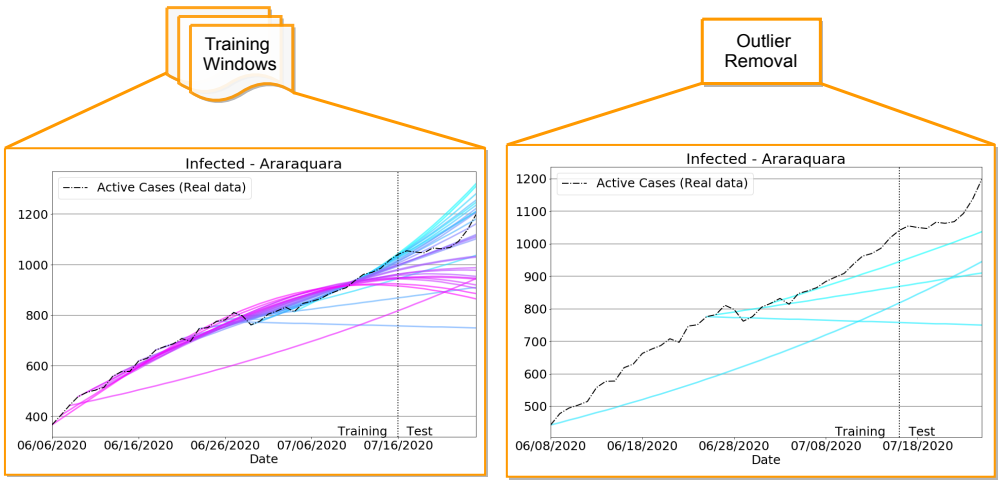
3. Finally, the remaining trained curves are used to calculate the definitive forecasts using the numerical solution of the SIRD system for $t \in [0, M + p]$, where p is the desirable forecast period. This is performed so as to balance the well-behaved contributions in the set of ODE solutions Λ, taking the mean of these outputs to determine $I_i(t)$, $D_i(t)$, $R_i(t)$ and $R_t(t)$.

The rationale behind the above-described outlier filtering scheme is that it prevents bad training results interfering with the forecast quality. Indeed, the filtering acts as an adaptive data-driven classifier, identifying badly-conditioned time window periods over the full time-series while still ensuring a better data fitting performance and smoothing. Further, as the effective reproduction number $R_t(t)$ drives the slope of the infection



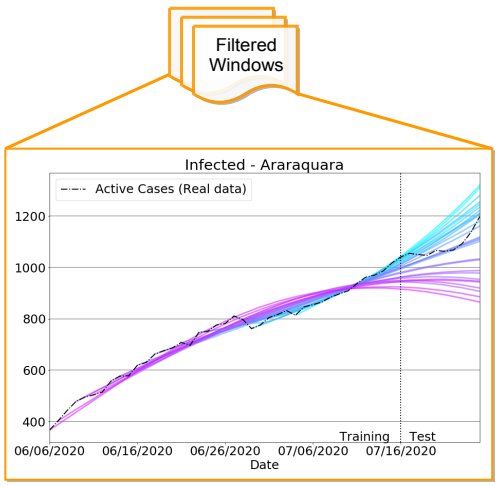
General filtering pipeline.

Figure 3.3: Cont.

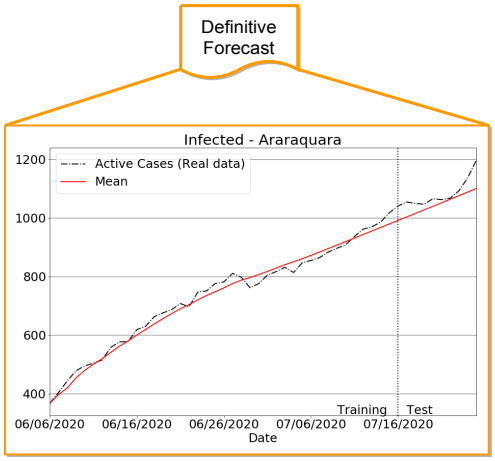


(a) Training Windows

(b) Discarded windows



(c) Acceptable matches



(d) Definitive prediction

Figure 3.3: (i) The complete filtering pipeline. (a) Training outputs for different time windows each colored line represents a training result. (b) The selected ill-behaved training periods (discarded trainings). (c) Training results which have passed the error criteria for good training. (d) Averaged results as the definitive prediction.

curve [1], it is expected that different successful training results produce similar estimations w.r.t. the true observed data of I, D and R so that the learning process will take into account only well-behaved parameters to estimate the definitive $R_t(t)$ curve. Figure 3.3 illustrates the aforementioned filtering approach when applied on a given dataset, while the implementation details are given in the Appendix section.

3.3 A SIR-Based Model Integrating Vaccination Dynamic

In order to extend the method and take vaccination process into consideration, we made a SIR variation that considers two doses and the period delay for the vaccine

[1] If $R_t(t) < 1$, the number of new infections in the next generation is reduced, while $R_t(t) > 1$ imposes the opposite situation.

take full efficacy in the human system, as shown in the following system of ODE with vaccination dynamic:

$$\begin{aligned}
\frac{dS}{dt} &= -\nu_1 S - \beta(t) IS, & \frac{dI_{v1}}{dt} &= \beta(t) I(V_1(1-\theta_1) + \bar{V}_1) - \gamma_r I_{v1},\\
\frac{d\bar{V}_1}{dt} &= \nu_1 S - \beta(t) I\bar{V}_1 - \alpha_1 \bar{V}_1, & \frac{dI_{v2}}{dt} &= \beta(t) I(V_2(1-\theta_2) + \bar{V}_2(1-\theta_1)) - \gamma_r I_{v2},\\
\frac{dV_1}{dt} &= \alpha_1 \bar{V}_1 - \beta(t) IV_1(1-\theta_1) - \nu_2 V_1, & \frac{dR_s}{dt} &= \gamma_r I_s - \nu_1 R_s,\\
\frac{d\bar{V}_2}{dt} &= \nu_2 V_1 - \beta(t) I\bar{V}_2(1-\theta_1) - \alpha_2 \bar{V}_2, & \frac{dR_{v1}}{dt} &= \gamma_r I_{v1} + \nu_1 R_s - \alpha\nu_2 R_{v1},\\
\frac{dV_2}{dt} &= \alpha_2 \bar{V}_2 - \beta(t) IV_2(1-\theta_2), & \frac{dR_{v2}}{dt} &= \alpha\nu_2 R_{v1} + \gamma_r I_{v2},\\
\frac{dI_s}{dt} &= \beta(t) IS - (\gamma_d + \gamma_r) I_s, & \frac{dD}{dt} &= \gamma_d I_s.
\end{aligned} \tag{3.9}$$

Due to the data-driven capability for parameter fitting, the model splits the population of infected and recovered into vaccinated and non-vaccinated so that we can track the vaccinated individuals over time.

An illustrative representation of the mathematical model can be seen in Fig. 3.4. Susceptible individuals are represented by S, while $\bar{V}_1$ and $\bar{V}_2$ account for the vaccinated individuals who took the first and second doses of the COVID-19 vaccine, respectively, but that are not totally immunized due to the time for it, i.e., enough time to provide effective protection. Since an individual could be infected after receiving the first dose, we have denoted this group as I_{v1}. Similarly, I_{v2} represents the group of individuals vaccinated with the second dose. Moreover, in our mathematical model, $I = I_s + I_{v1} + I_{v2}$ gives the infected individuals, where I_s accounts for the non-vaccinated individuals. After the vaccine has taken effect, vaccinated individuals are moved to V_1 and V_2 (subscript 1 for the first dose, and 2 for the second shot). Finally, R_{v1} and R_{v2} set the recovered individuals that have been vaccinated with 1 and 2 doses, respectively, while D denotes the total of deaths. Notice that we assume that the vaccine has complete effectiveness to severe cases, which means that deaths come from non-vaccinated individuals. For a list of parameters and variables, see Table 3.1.

In our approach, the LSODA solver is applied for numerically solving the mathematical model (3.9). An Artificial Neural Network (ANN) has also been used as part of our methodology, which takes a hidden layer with ten neurons, and the Sigmoid kernel as the network activation function. The output layer is fully connected to the hidden layer through a single neuron with no bias weights, wherein the ReLU has been adopted to trigger the neuron.

We provide below the main steps of our computational methodology, which include the redesign of the learning steps previously presented in [5] to cover the vaccination.

3.3.1 Obtaining the model's parameters

The epidemiological parameters $\beta(t), \gamma_r$ and γ_d are computed using a learning strategy based on the real data, which minimizes the following loss function:

$$\mathcal{L}(\beta_{net}(t), \gamma_r, \gamma_d, \nu_1, \nu_2) = \sum_{l \in L} l, \tag{3.10}$$

Table 3.1: Model Parameters

Notation	Description
$S(t)$	number of susceptible at time t
$I(t)$	Sum of all subgroups I_j at time t
$R(t)$	Sum of all subgroups R_j at time t
$D(t)$	numbers of deaths at time t
$\bar{V}_i(t)$	numbers of vaccinated but not yet immunized at time t with i = 1,2 doses
$V_i(t)$	numbers of immunized at time t with i = 1,2 doses
$\beta(t)$	transient transmission rate
γ_r	rate of recovered
γ_d	rate of mortality
ν_1 and ν_2	first and second dose vaccination rate, respectively
θ_1 and θ_2	first and second dose effectiveness, respectively
α_i	time delay for vaccine dose i = 1,2 effectiveness

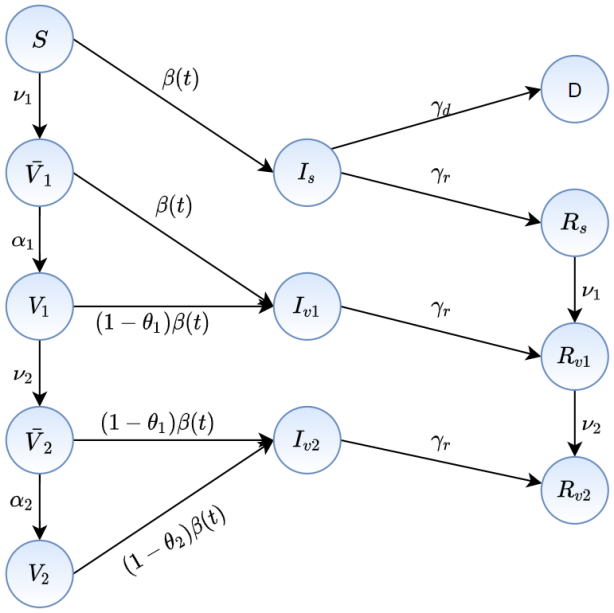


Figure 3.4: Representation of the current SIR-based model.

where

$$L = \{l_I, l_R, l_D, l_{v1}, l_{v2}, l_{sum}\},$$

$$l_I = \frac{1}{M}\sum_{i=0}^{M}\left|\log(I_i) - \log(\tilde{I}_i)\right|^2, \qquad l_{v1} = \frac{1}{M}\sum_{i=0}^{M}\left|\log(V_{1,i}) - \log(\tilde{V}_{1,i})\right|^2,$$

$$l_R = \frac{1}{M}\sum_{i=0}^{M}\left|\log(R_i) - \log(\tilde{R}_i)\right|^2, \qquad l_{v2} = \frac{1}{M}\sum_{i=0}^{M}\left|\log(V_{2,i}) - \log(\tilde{V}_{2,i})\right|^2,$$

$$l_D = \frac{1}{M}\sum_{i=0}^{M}\left|\log(D_i) - \log(\tilde{D}_i)\right|^2, \qquad l_{sum} = \frac{1}{M}\sum_{i=0}^{M}\left|\log(\tilde{T}_i)\right|^2,$$

$$\tilde{I}_i = \tilde{I}_s(t_i) + \tilde{I}_{v,1}(t_i) + \tilde{I}_{v,2}(t_i),$$
$$\tilde{R}_i = \tilde{R}_s(t_i) + \tilde{R}_{v1}(t_i) + \tilde{R}_{v2}(t_i),$$
$$\tilde{V}_{j,i} = \tilde{V}_{j,i} + \tilde{\tilde{V}}_{j,i} + \tilde{I}_{vj} + \tilde{R}_{v_j},\ j = 1, 2,$$
$$\tilde{T}_i = \tilde{S}_i + \tilde{I}_i + \tilde{R}_i + \tilde{D}_i + \tilde{V}_{1,i} + \tilde{V}_{2,i}.$$

In Equation (3.10), X_i and $\tilde{X}_i$ account for the real value and the numerical solution, respectively, at the discretized time t_i for a given target variable. Notice that, in the absence of available data of the vaccinated individuals who got infected, we take the number of vaccine shots when computing l_{v1} and l_{v2}.

In order to assign the time variation to the transmission rate, $\beta_{net}(t)$ is properly computed from a network-based architecture. As a result, the trainable parameters of the epidemic model are properly learned and computed by solving the ANN-related optimization problem (3.8). If the vaccination one wants to train the vaccination rate, the optimization problem can include those parameters as shown in equation (3.11):

$$\underset{W,b,\gamma_r,\gamma_d,\nu_1,\nu_2}{\arg\min} \quad \mathcal{L}(\beta_{net}(t), \gamma_r, \gamma_d), \tag{3.11}$$

where $\{W, b\}$ are the ANN weights.

3.3.2 Effective Reproduction Number

Since the effective reproduction number $R_t(t)$ is a very important measure used in infectious diseases, we analyze how the new variables inserted into our mathematical model influence this measure. Formally speaking, $R_t(t)$ is defined as the quotient between the transmission and recovery rates weighted by the percentage of susceptible population [5].

Notice that, in our SIR formulation, vaccinated individuals are still susceptible to contagion, but they are subject to a lower rate of infectivity due to the vaccine. Following the multi-group model of the reproduction number given by Driessche and Watmough [84], $R_t(t)$ can be computed as:

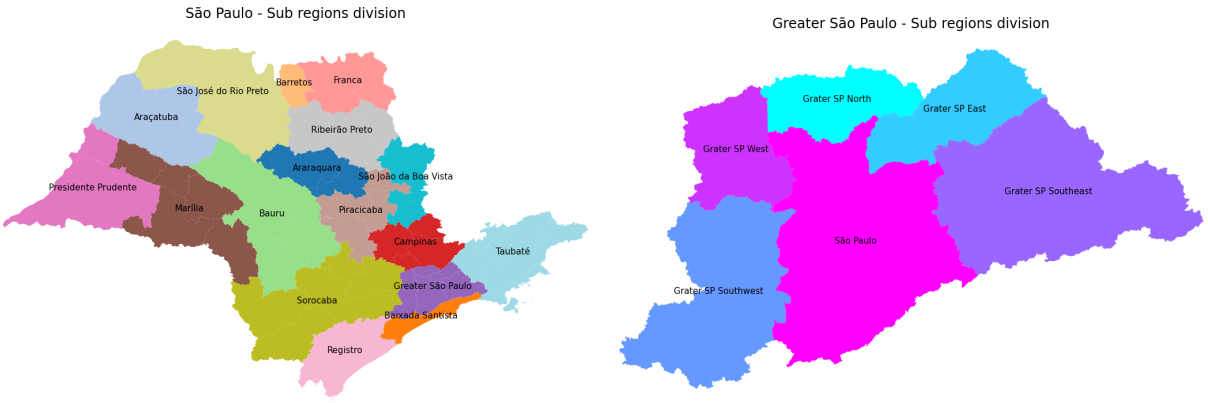
$$R_t(t) = \frac{\beta(t)}{\gamma_r}\left[\bar{V}_1 + (1-\theta_1)(V_1 + \bar{V}_2) + (1-\theta_2)V_2\right] + + \frac{\beta(t)}{\gamma_r + \gamma_d}S \tag{3.12}$$

Chapter 4

Results and Discussion

4.1 Data Organization

In order to track the daily evolution of COVID-19 while collaborating with the decision makers of Brazilian public body, we rearrange the collected data into 22 large regions corresponding to each Regional Health Department of the state (see Fig. 4.1(a)). Particularly, due to the huge urban sprawl around São Paulo city, state government has grouped the so-called *Greater São Paulo region* into 6 sub-regions (São Paulo North, São Paulo Southeast, São Paulo Southwest, São Paulo Northeast, São Paulo Metropolitan, São Paulo East, São Paulo West), as illustrated in Fig. 4.1(b). As a result, for each one of the 22 Health Departments, time-series for *confirmed cases* and *deaths* were obtained, with entries ranging from April 1 to October 31, i.e., a seven-month period of daily records.



(a) (b)

Figure 4.1: Sub-region maps of São Paulo state: (a) State map showing the 22 state sub-regions and (b) São Paulo metropolitan region.

The data collection utilized in our experiments with vaccination model comprises several data resources and public repositories, which vary according to the region/country studied. Below, we detail the data sets employed in our analysis and simulations.

- **Israel dataset**: COVID-19 confirmed cases and deaths were downloaded from the well-known *Johns Hopkins University* (JHU) data repository [25], which is available at `https://coronavirus.jhu.edu/map.html`. Concerning the vaccination data, it

Figure 4.2: Geographic location of Serrana's town.

was collected from the *Israel Ministry of Health* website, at `https://www.gov.il/en/departments/guides/information-corona`.

- **Serrana's town dataset**: Times-series of confirmed cases, recoveries, and deaths gathered from the COVID-19 health bulletins of *Serrana's local government*, hosted at `http://www.serrana.sp.gov.br/coronavirus`. The total number of vaccine doses administered per day was obtained from the *São Paulo State government* website, at `https://www.saopaulo.sp.gov.br/planosp/simi/dados-abertos`, while the vaccination time-series used to simulate a regular immunization campaign was taken from the online tracking platform *SP Covid-19 Info Tracker* [5], available at `http://www.spcovid.net.br`.

- **São Paulo State dataset**: Data acquired from the *São Paulo State government* website, including confirmed cases, recoveries, deaths, and the number of immunized people per day for each type of vaccine dose.

- **Brazil dataset**: Time-series taken from the *Brazilian Ministry of Health* website, at `https://covid.saude.gov.br`. The administered doses per day were obtained from the *Vaccinometer-SUS*, a real-time platform of Brazilian government hosted at `https://localizasus.saude.gov.br`.

4.2 Metrics

In our experiments, the forecasts are assessed by applying the *Mean Absolute Percentage Error* (MAPE), a classic evaluation metric widely used in time-series analysis [44, 45]:

$$MAPE(Y_i, \dot{Y}_i) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \dot{Y}_i}{Y_i} \right| \times 100, \tag{4.1}$$

where Y_i and $\dot{Y}_i$ account for the real and predicted daily values of any target variable as predicted by our data-driven model. In our assessments, we follow [45] so that a percentage of up to 10 % is established for MAPE in order to ensure a "satisfactory level" of accuracy w.r.t. predictive performance.

Another evaluation metric taken in our qualitative analysis is the *Normalized Root Mean Square Error* (NRMSE), computed according to the following expression [65]:

$$NRMSE(Y_i, \dot{Y}_i) = \frac{1}{n} \frac{\sqrt{\sum_{i=1}^{n} (Y_i - \dot{Y}_i)^2}}{\overline{Y}} \tag{4.2}$$

where $\overline{Y}$ determines the average of the observed data.

Finally, we also make use of the *Variance* to assess how the trained parameters can affect the reproduction number $R_t(t)$ as the training interval i in the SIRD model varies. Such statistical metric is calculated for each time t_j of the training period, by applying the following formula:

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (R_t^{(i)}(t_j) - \overline{R_t}(t_j))^2 \,, \tag{4.3}$$

where $R_t^{(i)}(t_j)$ represents the estimated value of $R_t(t)$ at the discretized time t_j.

4.3 Results for the Data-driven SIRD Model

4.3.1 Badly-Conditioned Samples × Data Fitting Robustness and Accuracy

As previously discussed in Section 3.2.1, the amount of data used to calibrate the model's parameters can impact the Covid-19 estimations such as the actual infections $I(t)$ and reproduction number $R_t(t)$, as specific time-series periods are made-up of badly-conditioned data. In order to show how such an issue can influence the forecasts, and how our moving window-based training scheme can fix it, we present in Figure 4.3 both the $R_t(t)$ and infection $I(t)$ curves in the period when there were no full updates of Covid-19 data in various São Paulo state regions, as pointed out by the Brazilian press news [30]. Notice from the results with badly-conditioned data that although the predicted values produced large peaks and valleys in both $R_t(t)$ and $I(t)$ curves, the definitive forecast values (in red) were successfully fitted, keeping very close to the true data. Indeed, even in more drastic cases involving bad behaving data (see Greater SP Southeast and Marília regions), our data fitting approach performed well, ensuring the correct tendency of the real curves. Finally, it can be seen that the reproduction number $R_t(t)$ dictates the slope of infection curves, as expected.

The forecasting results from Figure 4.3 were also assessed via quality metrics. Besides the well-established MAPE score, we take Eq. (4.3) as a popular assessment metric to gauge data variability and inconsistency level. More precisely: given a fixed point t_j in the simulation domain, we compute the variance s_j w.r.t. all i samples of $R_0^{(i)}(t_j)$, generated as the i-th training window varies during the full learning process. As a result, if the variance s_j is low, the trainable infected values for all i indices will follow the same common tendency, which means that the node t_j does not hold badly-conditioned data. On the other hand, if the variance is high, then there are training intervals i probably inconsistent and badly behaved. From Table 4.1, one can check that our training approach delivered low error measurements, producing very stable estimations with low prediction

variations, as measured by the variance. For example, the largest MAPE error was around 4%, while the highest value for variance w.r.t. the predicted effective reproduction number $R_t(t)$ was 1.478.

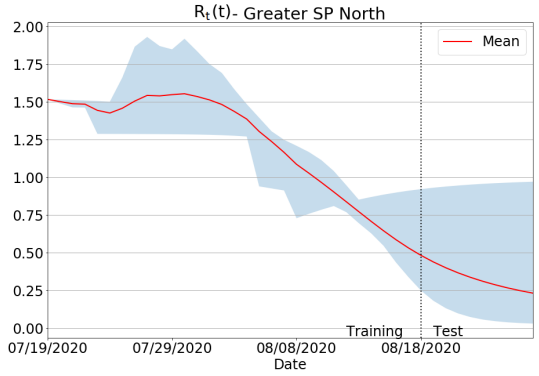
Table 4.1: Variance computed during the training process, and average MAPE for active cases (infected) w.r.t. Fig. 4.3 results.

Region	**Variance Norm** ($||s^2||_2$)	**MAPE for Active Cases (%)**
Greater São Paulo North	0.098	2.658
Greater São Paulo Southeast	1.478	4.414
Marília	0.378	1.928
Ribeirão Preto	0.063	3.894

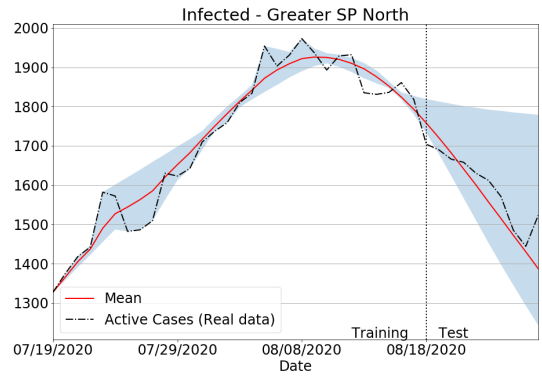
4.3.2 The Transient Behavior of Transmission Rate

As discussed in Section 3.2, an important strategy adopted in our mathematical approach is the use of a simple, but effective Artificial Neural Network (ANN) for estimating the transmission rate β in a transient context.

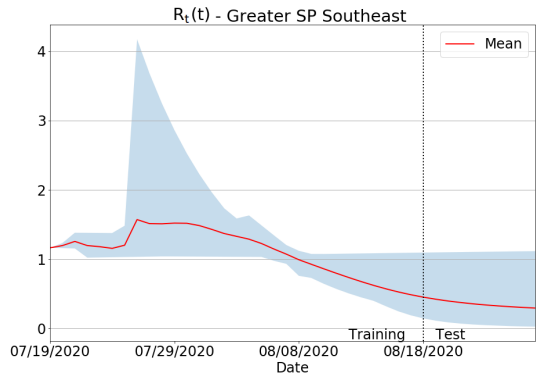
To better emphasize the neural network importance in ensuring a transient behavior to the transmission rate, we compare the results with/without the ANN so that a transient/constant behavior for β is achieved. In particular, we select two well-distinct



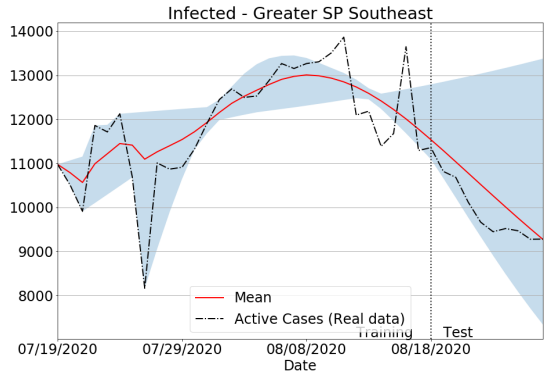
(a) $R_t(t)$ for Greater São Paulo North region.



(b) $I(t)$ for Greater São Paulo North region.

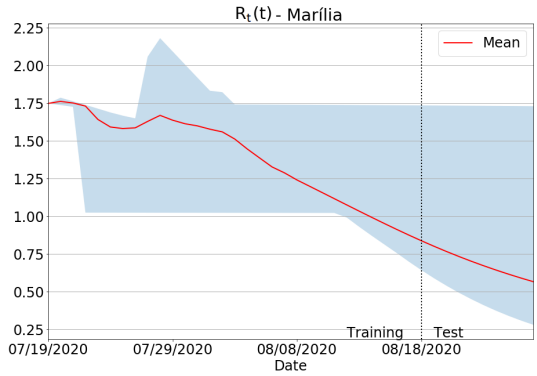


(c) $R_t(t)$ for Greater São Paulo Southeast region.

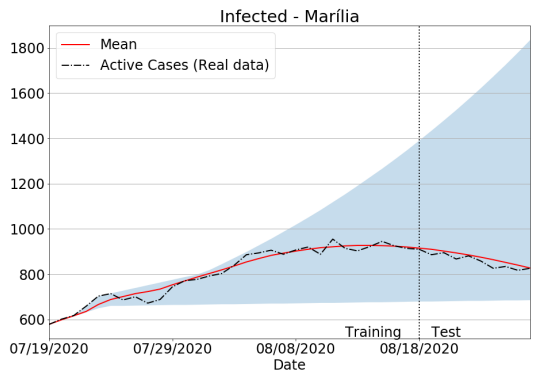


(d) $I(t)$ for Greater São Paulo southeast region.

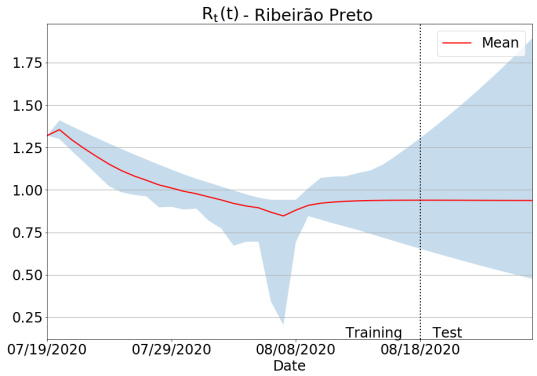
Figure 4.3: *Cont.*



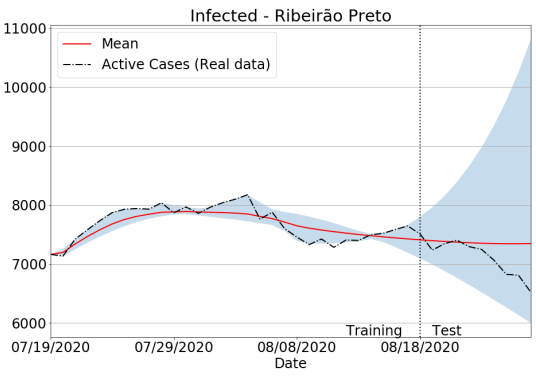
(e) $R_t(t)$ for Marília region.



(f) $I(t)$ for Marília region.



(g) $R_t(t)$ for Ribeirão Preto region.

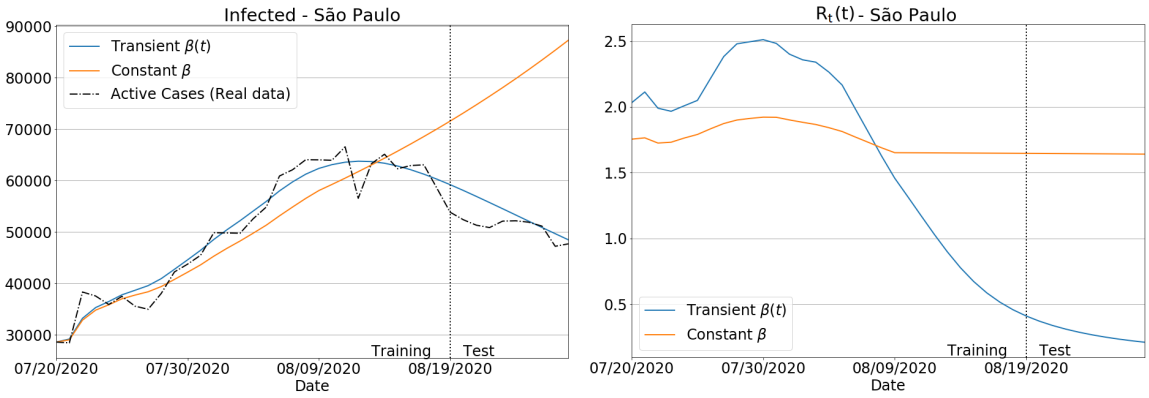


(h) $I(t)$ for Ribeirão Preto region.

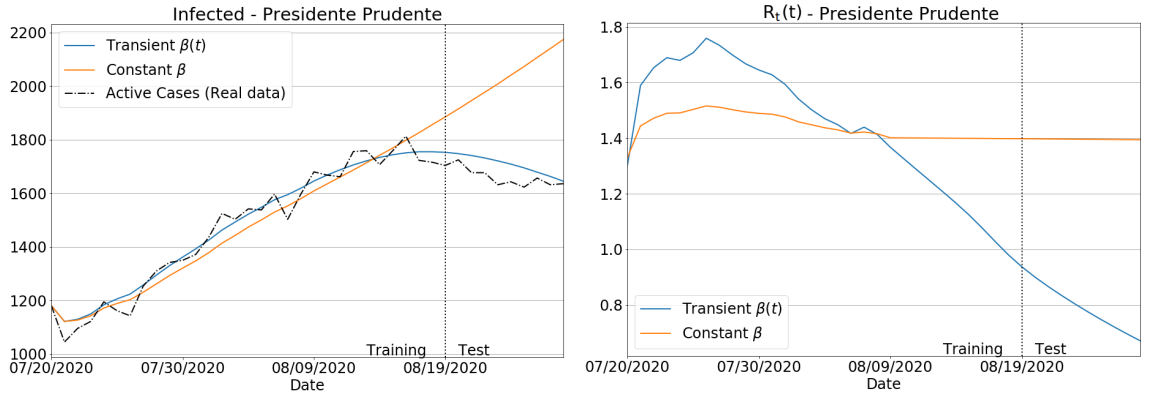
Figure 4.3: Reproduction number $R_t(t)$ and infected $I(t)$ predictions for the mean, minimum, and maximum forecasted values as the training window moves, i.e., by varying $M = 10, 11, ..., 40$ in Equation (3.2) and training the parameters in a coupled and recursive way. Red lines establish the mean predicted values after the full learning procedure is finished, while the vertical dotted lines split the training and forecasting periods.

regions: a small one (Presidente Prudente), and the biggest region of the São Paulo state (São Paulo city).

Figures 4.4(a) and 4.4(c) bring the infected curves. Blue lines give the estimation for transient β, while the orange curves represent the homogeneous behavior for β. From the plotted curves, we can confirm that the best predictions for the number of the infected are those using the $\beta_{net}(t)$ for both regions. For the effective reproduction number $R_t(t)$, Figures 4.4(b) and 4.4(d) display the importance of taking into account the transient form of β for a more accurate estimation. Notice that, by assuming a constant value for β, one can get $R_t(t)$ with a low variation and adjustability, specially in the test period, as no significant changes have been found w.r.t. the reproduction number. Finally, we measure in Figure 4.5 the effective impact of transience on β, by assessing the MAPE for the number of infections in the same time period as considered in Figure 4.4. From the reported scores, one can see that there is a substantial reduction in the MAPE errors as the transmission rate β is estimated from a transient way. Moreover, by assuming a data-driven β learned via an intelligent architecture such as ANN, one can see that $\beta = \beta_{net}(t)$ is not only suitable to improve the prediction accuracy of SIRD-based formulation, but also the well-known SIR model. In fact, both SIR and SIRD when coupled with a learned transmission rate $\beta_{net}(t)$ perform similarly, producing much lower prediction errors than the case for which β is taken as a non-learned function.



(a) Active cases (b) $R_t(t)$



(c) Active cases (d) $R_t(t)$

Figure 4.4: Infected and effective reproduction number using constant and transient values for β: São Paulo region (first row) and Presidente Prudente region (second row).

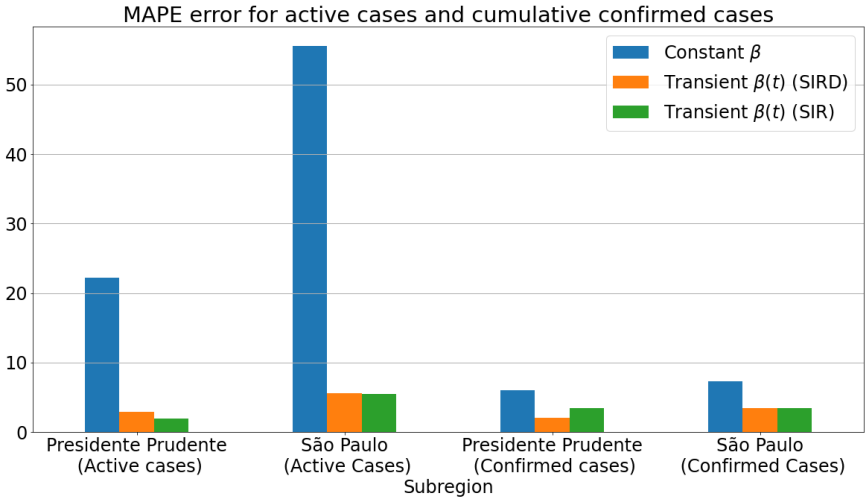


Figure 4.5: Comparison of MAPE errors for constant and transient values of β: SIR and SIRD models.

4.3.3 Invariance to Training Periods

This section is dedicated to confirming that our approach can accurately predict accumulated, recovered, and deceased cases regardless of data training period. In order to verify such a method's feature, we follow the usual subdivision of the São Paulo state to

group the whole population into four main regions: *Coastal*, *Greater São Paulo*, *West*, and *East* areas, as illustrated in Figure 4.6.

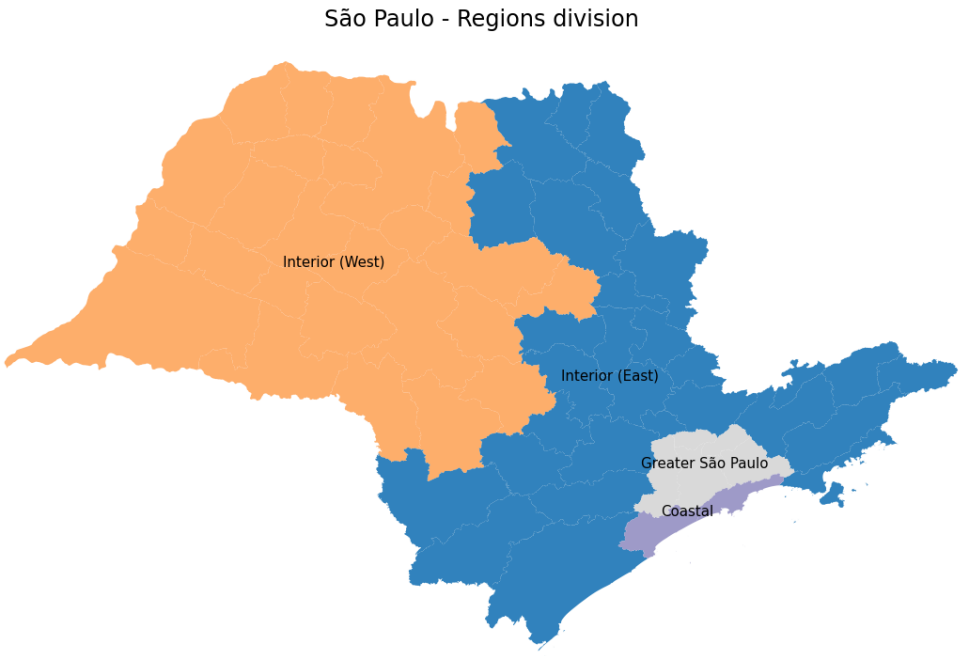


Figure 4.6: São Paulo state subregions.

As our trained model can estimate distinct epidemiological metrics, in this experiment, we focus on the pandemic parameters for which the MAPE can be properly computed, i.e., accumulated, recovered and deceased cases. In our quantitative analysis, we train our approach by taking a full period of $M = 30$ consecutive days to predict the next 10-days of the aforementioned variables over three different forecasting periods: August, September and October, as listed in Table 4.2. One can check from the tabulated scores that the predictions are quantitatively accurate and stable, since MAPE errors are substantially low for all regions. The maximum MAPE is observed for recovered cases in Greater São Paulo for the first period, while both accumulated and deceased cases delivered low errors, even the biggest measured ones, as reported to death curve of East's first period (3.465) and Coastal's second period for Covid-19 cases (1.536). Therefore, our data-driven approach turns out to be stable and robust over different prediction periods, even with a small amount of data taken to generate the training set: a 3 to 1 ratio w.r.t. the full test set, i.e., 30 past days for training the model, and 10 days for future predictions. In fact, in Table 4.3, we show that the proposed learning approach still remains unchanged and consistent as the window size of the training set changes.

4.3.4 Quantitative and Qualitative Investigations

We now discuss the forecasting results provided by the proposed methodology under 10 day-time horizons for all the São Paulo regions. Additionally, we extend our analysis to better understand and discuss both the past pandemic situation and the rise of a second wave in the whole country, as we have recently observed from the most current data. In such particular case, the Brazilian dataset has been downloaded directly from the government official source [53], and it covers all the five regions of the country (see Figure 4.7 for an illustration). From the available data, we performed our analysis in terms of the following Covid-19 indicators: accumulated, recovered and deceased cases. These data – together with new hospitalizations – have been the main pandemic metrics

Table 4.2: MAPE errors for different forecasting periods.

Region	MAPE error for accumulated cases (%)	MAPE error for recovered cases (%)	MAPE error for deceased cases (%)
August 15 - August 24			
Coastal	1.513	0.951	1.046
Greater São Paulo	0.753	3.731	1.394
Interior (East)	0.454	1.491	3.465
Interior (West)	1.085	1.826	2.618
September 15 - September 24			
Coastal	1.536	0.347	2.503
Greater São Paulo	0.598	0.344	0.926
Interior (East)	0.937	0.461	1.157
Interior (West)	1.277	0.753	0.603
October 15 - October 24			
Coastal	0.533	0.249	0.268
Greater São Paulo	0.105	0.438	0.776
Interior (East)	1.413	0.886	0.236
Interior (West)	0.832	1.097	0.881

Table 4.3: MAPE errors for Greater São Paulo region as the size of the training window varies.

Training windows	MAPE error for cumulative cases (%)	MAPE error for cumulative deceases (%)	MAPE error for cumulative recovereies (%)
10-30 days	0.285	0.753	0.293
10-40 days	0.762	0.928	0.321
10-50 days	1.179	0.894	0.592

used by public body to assess the Covid-19 scenario in São Paulo and Brazil [53]. Finally, it is worth mentioning that, in synergy with the efforts made in both state and national levels, we have continuously collaborated with different press conglomerates and public authorities, especially in the last few weeks, where a substantial increase in new cases of Covid-19 and hospitalizations have been firstly warned by our data analysis tool – *Info Tracker* (see [12, 19, 66] for a few English news published by Brazilian media agencies).

4.3.4.1 São Paulo State Regions

Firstly, we provide in Table 4.4 both MAPE and RMSE measurements for the São Paulo state regions. As one can verify, all the MAPE errors are lower than 1, except Coastal region, where a MAPE of 1.2 has been calculated. Regarding RMSE, regions have presented very low errors, whose values are of order 10^{-2} on average, thus attesting to the high quality performance of our hybrid SIRD enhanced by a Machine Learning based approach.

For completeness, we have plotted the results for the São Paulo state in Figure 4.8. In particular, comparisons between real data and our estimates for accumulated, recovered, and deceased cases are presented in first, second and third columns in Figure 4.8. To better emphasize the Covid-19 transmissibility in the state, the reproduction number was

Figure 4.7: Brazilian regions.

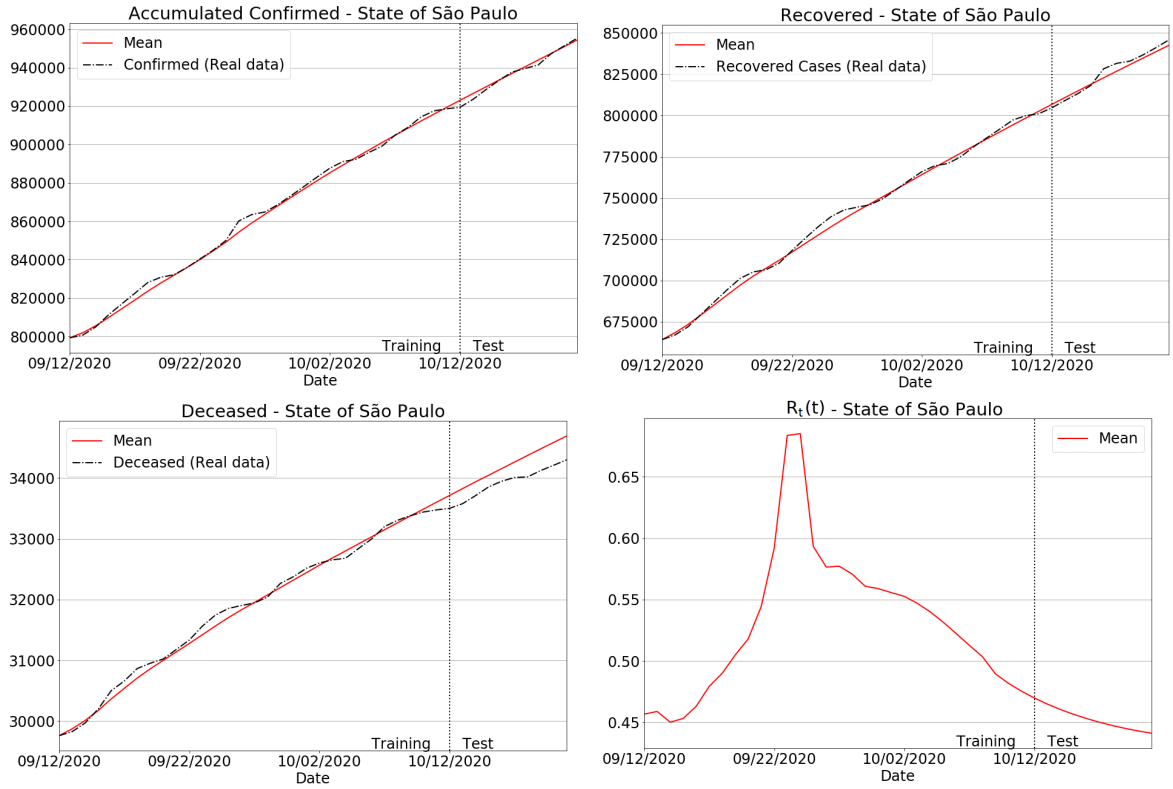


Figure 4.8: Forecasting results for São Paulo state.

also displayed in the last column. Considering the prediction intervals from test periods,

Table 4.4: Tabulated errors w.r.t. predictions depicted in Figure 4.8 (São Paulo state regions).

Region	Accumulated Cases		Recovered		Deaths	
	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE
Costal	0.325	0.004	0.907	0.010	1.200	0.012
Greater São Paulo	0.680	0.007	0.371	0.004	0.714	0.007
Interior (East)	0.818	0.010	0.592	0.007	0.312	0.004
Interior (West)	0.376	0.005	0.626	0.007	0.826	0.009
State of São Paulo	0.219	0.003	0.455	0.005	0.475	0.005

we can see that the proposed model reaches a very accurate agreement between the true data and the forecasts of accumulated, recovered, and deceased cases in São Paulo state. Another important aspect to be noted is that our model fits accurately the real data in the training interval, regardless of Covid-19 indicator.

4.3.4.2 Brazilian Regions

The next experiment comprises the Brazilian case, where predictions of confirmed, recovered and deceased cases have been delivered for all the five regions of the country. In terms of quantitative assessment, Table 4.5 reports the MAPE and RMSE, where one can verify that the predictions for the accumulated, recovered and deaths are numerically consistent and reliable. Additionally, we have plotted the results for the Brazil in Figure 4.9, including the predictions of confirmed, recovered and deceased cases. Similar to the São Paulo state case, we also provide $R_t(t)$. Particularly, well-behaved curves were produced considering the extrapolation of real data for the whole country; thus demonstrating the effectiveness of the proposed method in dealing with a huge amount of Covid-19 data.

Table 4.5: Tabulated errors w.r.t. predictions depicted in Fig. 4.9 (Brazilian regions).

Region	Accumulated Cases		Recovered		Deaths	
	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE
Midwest	1.169	0.014	0.989	0.013	0.856	0.009
North	0.889	0.010	0.282	0.003	0.173	0.003
Northeast	0.244	0.003	0.342	0.005	0.487	0.005
South	4.413	0.047	7.111	0.072	0.397	0.004
Southeast	0.815	0.009	0.675	0.009	0.427	0.005
Brazil	0.323	0.004	0.638	0.008	0.273	0.003

4.3.4.3 The Second Wave of Covid-19: investigations in Brazil and other countries

In this section, we discuss our predictions considering the rise of a second wave of coronavirus hitting Brazil. Particularly, one important aspect of our tracking platform is its capability for dealing with real data resulting possibly from a second wave of Covid-19, as already pointed out by our warnings, discussed at the beginning of Section 4.3.4.

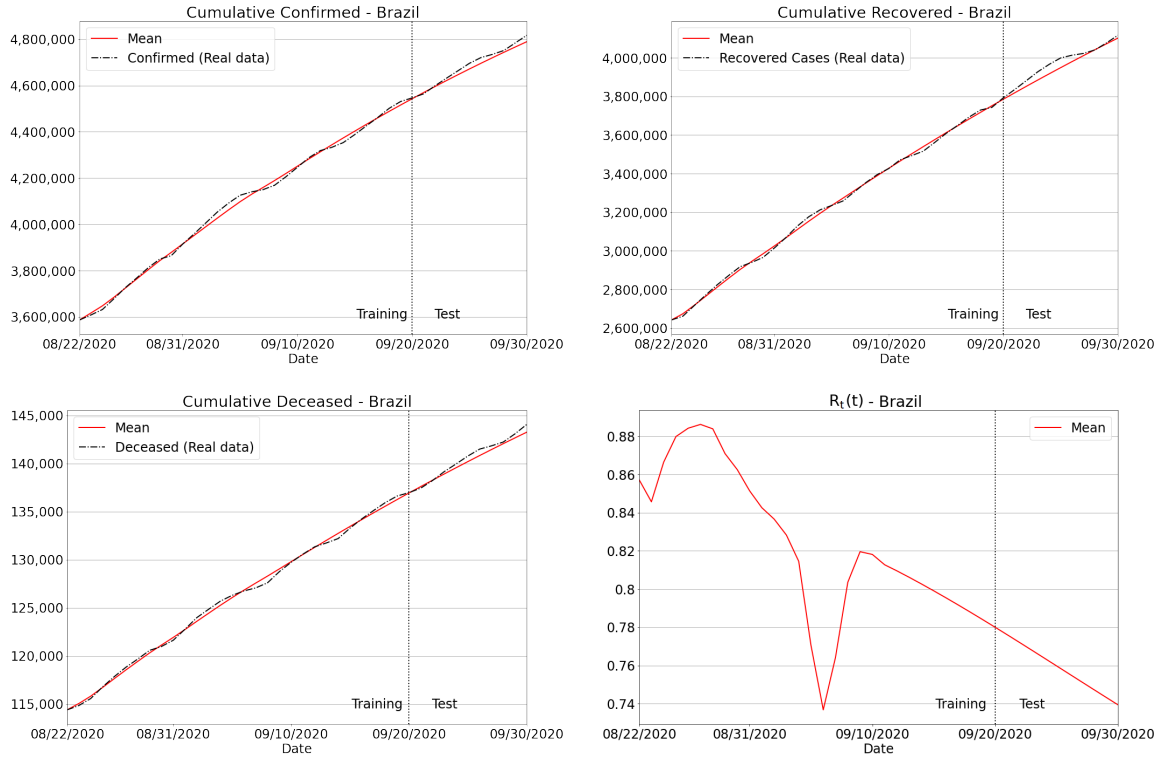
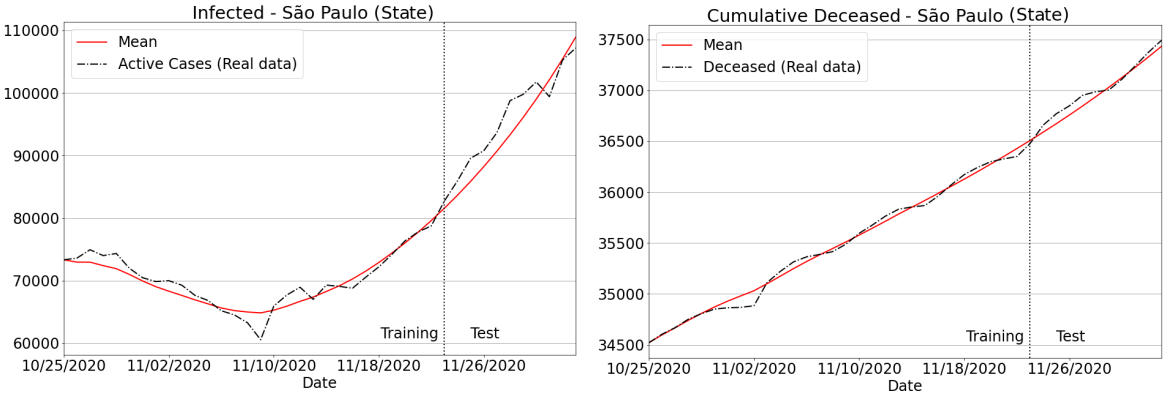


Figure 4.9: Forecasting results for Brazil.

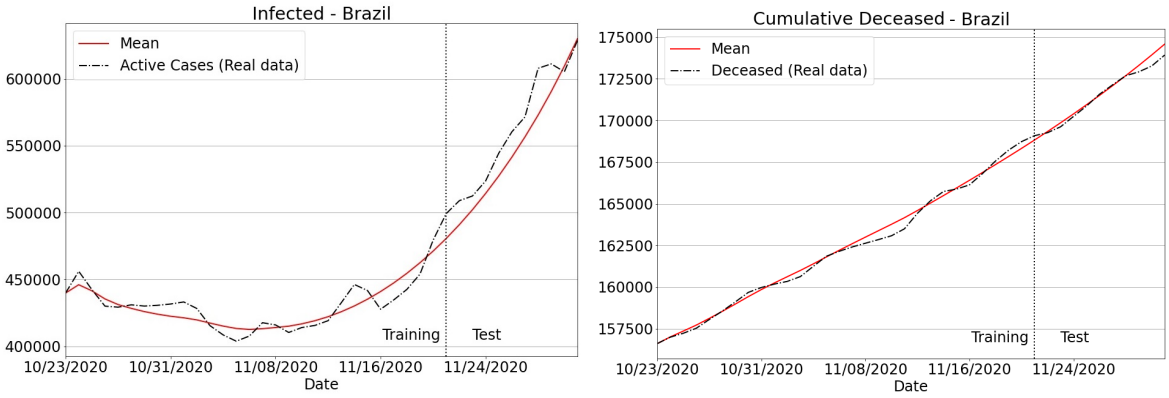
Moreover, with the eminence of a fast acceleration in new cases, as reported in European countries in the last months, very recent papers relying on SIR-based models have been presented in the literature (e.g. [17, 33, 67, 29]). Therefore, following the aforementioned works, we make use of our SIRD + machine learning methodology for both purposes: (i) analyzing the historicity of the pandemic's most recent past in Brazil, and (ii) supporting the state and federal government to implement immediate decisions in order to contain the advance of coronavirus in the country.

A warning, real case involving the predictions resultant from our approach is depicted in Figure 4.10. First, one can verify that the trajectory, as well as the real values for new infections and deaths in both training and prediction periods have been successfully captured. Second, the high upward trend of new infection curves suggest that both São Paulo and Brazil have recently suffered from a substantial growth in new cases and deaths. Note that the new infections in the state of São Paulo jumped from 70K, on November, 14 to 110k on December, 3: an increase of 57% in a short period of 19 days. When inspecting the Brazilian curves, a similar finding is observed: a raise of 415K, from November, 8 to around 630K in early December, i.e., a worrying jump of 51% in just three weeks.

To provide further evidence concerning the feasibility of the current methodology, we have investigated the spread of COVID-19 for three different data sets: Italy, Portugal and Ukraine. The analysis was conducted considering the data provided by Johns Hopkins University [40], from October, 25 to December, 3. Particularly, according to Figure 4.11, we can observe that our methodology was able to fit very well the real data from all the European countries for the total number of infected and deceased. Therefore, these results confirm that our SIRD model enhanced by a learning scheme can be successfully applied to inspect the COVID-19 spread in several regions of the world.



(a) São Paulo state

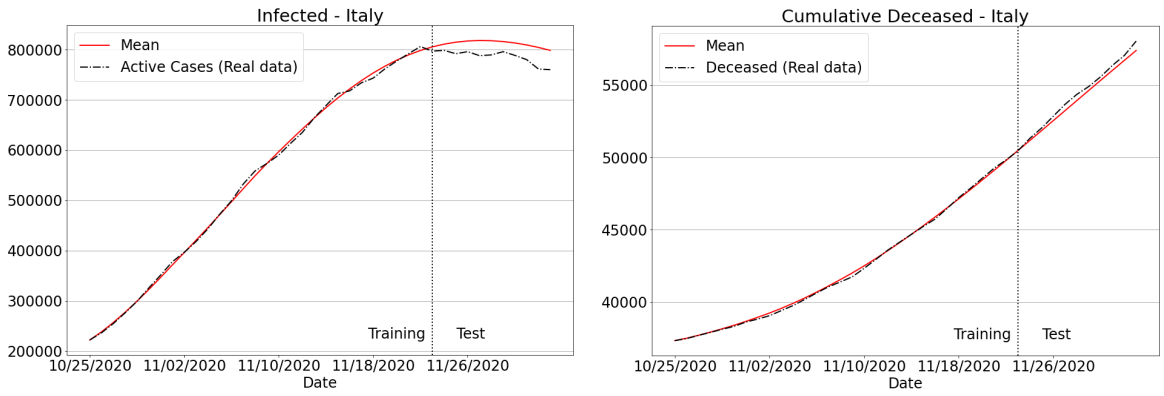


(b) Brazil

Figure 4.10: Infected and deaths for São Paulo state and Brazil. The high increase in both indicators suggests the eminence of a "second wave" of coronavirus hitting the country and starting in the second half of November of 2020.

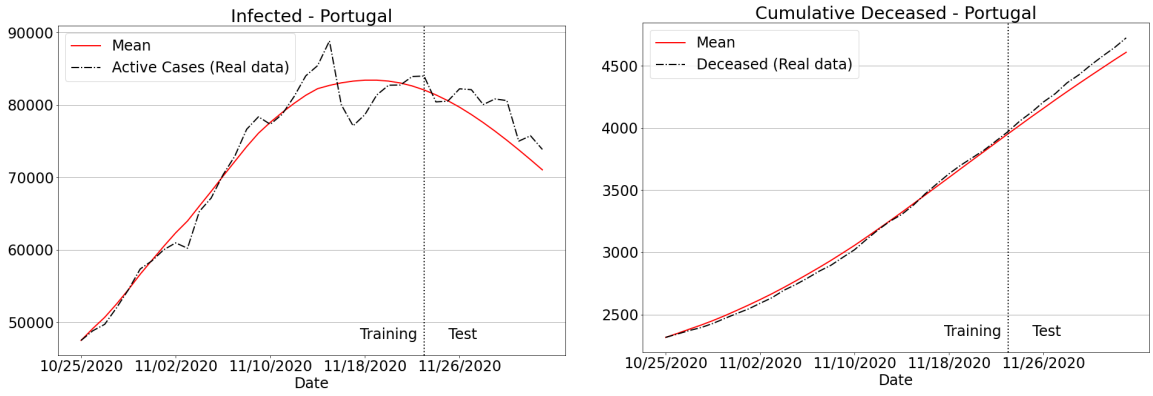
4.4 Results for the Data-driven Vaccination Model

Now, we present and discuss our forecast results, simulated scenarios, and the main findings emerging from our data-driven analysis considering the vaccination data. Parameters θ_1, θ_2 and α used to run the learning steps of the SIR-based models were taken following the reported values in Table 4.6. Although the efficacy delay α are obtained

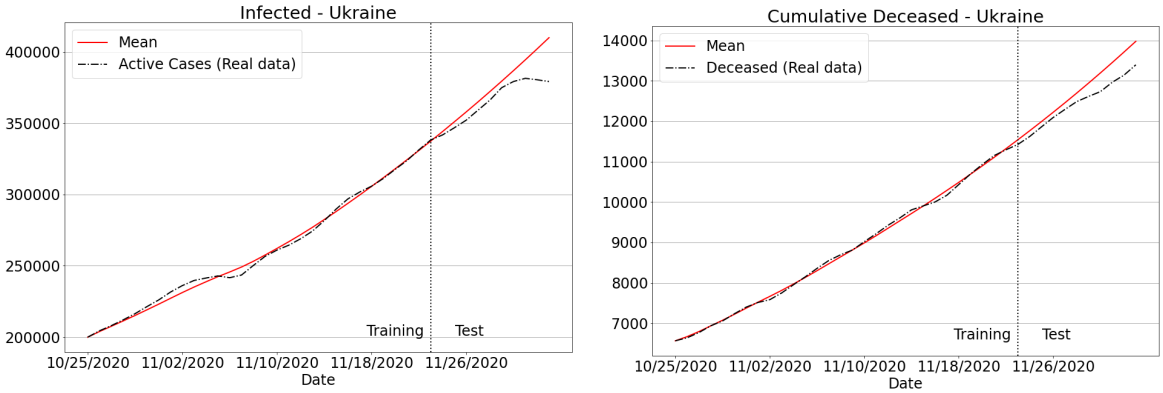


(a) Italy

Figure 4.11: *Cont.*



(b) Portugal



(c) Ukraine

Figure 4.11: Infected and deaths for Italy, Portugal and Ukraine over recent data.

on the manufacturer researches and those data are the ones used this work, there are researches [76] showing that the efficacy can change with a optimal dose interval. The vaccination rates ν_1 and ν_2 were set according to the total number of vaccinated people per day for each type of vaccine dose as observed on each region/country studied.

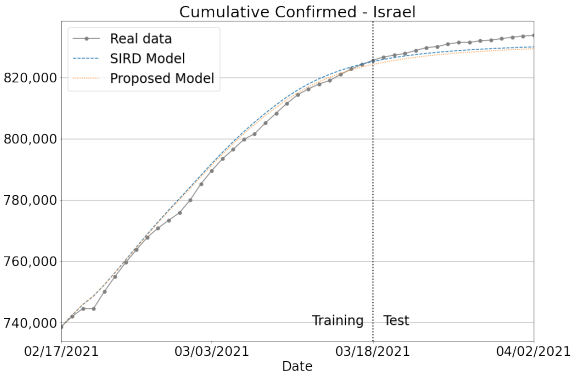
Table 4.6: Efficacy of different vaccine types.

Vaccine	**Efficacy Dose 1**	**Efficacy Dose 2**	**Efficacy Delay**	**Source**
Coronavac (Sinovac)	$\theta_1 = 5.98\%$	$\theta_2 = 66.48\%$	$\alpha = \frac{1}{14}$	[78]
Pfizer (BioNTech)	$\theta_1 = 52.00\%$	$\theta_2 = 95.00\%$	$\alpha = \frac{1}{21}$	[63]
AstraZeneca (Oxford)	$\theta_1 = 64.00\%$	$\theta_2 = 70.40\%$	$\alpha = \frac{1}{14}$	[85]

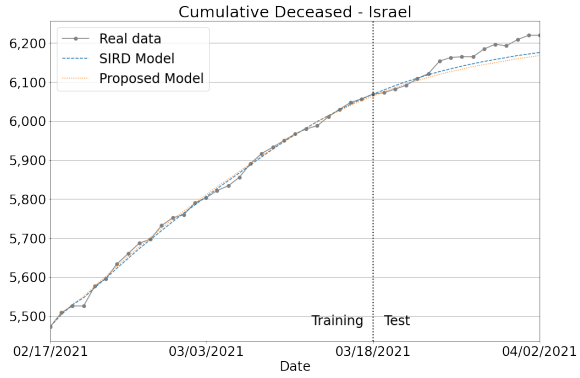
4.4.1 Validation with Real Vaccination Campaigns

In order to assess the proposed methodology, we take as a benchmark the well-established vaccination data from Israel, since the country has already reached a large number of immunized people against COVID-19 with Pfizer vaccine [63]. Also, we attest to the accuracy of our SIR-based model by analyzing the mass vaccination rollout of

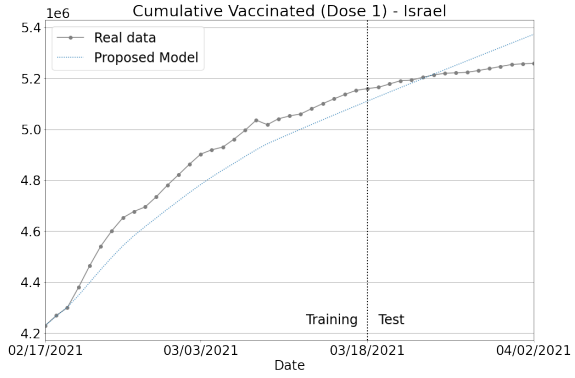
Serrana's town: a provincial city of 45,000 inhabitants located in the State of São Paulo, Brazil. Very recently, Serrana became the first fully immunized city in the World as a consequence of *Project S* [82, 69], a Brazilian pilot study conducted by the São Paulo State government in cooperation with *Instituto Butantan* [37] to address important issues regarding the mass vaccination effects with Coronavac/Sinovac immunizer. Finally, the results obtained from the recently published data-driven model [5] (namely here as "SIRD") are also given for the sake of comparison, since it is a robust SIR-based approach that does not include any vaccination dynamic into its formulation.



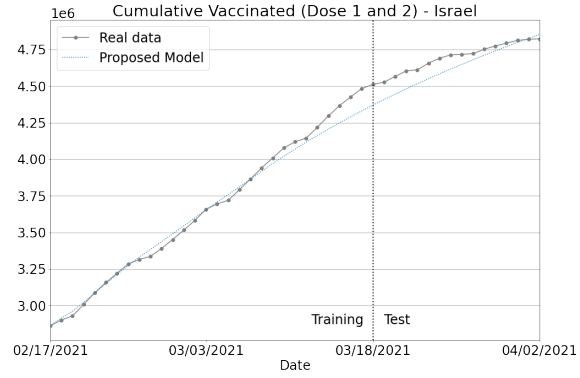
(a) Cases



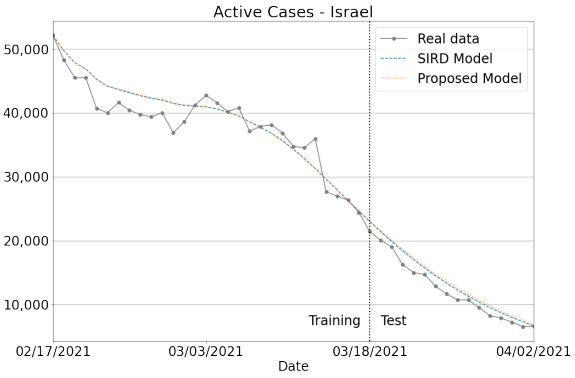
(b) Deaths



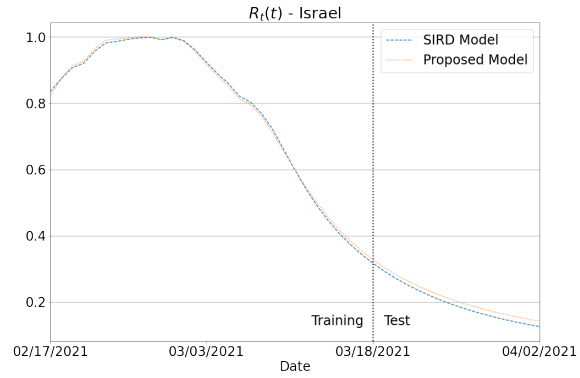
(c) Vaccinated people (dose 1).



(d) Vaccinated people (doses 1 and 2).



(e) Active cases



(f) $R_t(t)$

Figure 4.12: Israel validation results. (a) Confirmed cases, (b) deaths, (c) vaccinated people who received at least one vaccine dose, (d) two doses, (e) active cases, and (f) effective reproduction number. Training period from February 17$^{\text{th}}$ 2021 to March 18$^{\text{th}}$ 2021.

4.4.1.1 Israel vaccination campaign

Figure 4.12 shows the forecasting results, under a 14 days' time horizon, for several Israeli COVID-19 curves. One can verify that our model fits the real scenario as precisely as the SIRD method [5]. However, since the SIRD approach does not take into account a data-driven vaccine compartment, the effects of transmission rate are only assigned to the learned parameter $\beta(t)$, which attempts to capture the different stages of the virus spreading. On the other hand, our model allows to properly deal with an immunization campaign against the coronavirus, including the tuning of the vaccination parameters as well as understanding the role of the vaccinated individuals on the full dynamic system of the disease. The reproduction number also follows similar trajectories for both predictors, indicating that active cases and the effective reproduction number $R_t(t)$ significantly decrease when a massive vaccination rollout is combined with a high-efficacy vaccine. Finally, concerning the quantitative verification listed in Table 4.7, once again both models produce very similar results, as the highest computed MAPE is on the order of less than 1%, i.e., a very small prediction error.

Table 4.7: Average MAPE and NRMSE for the two-week forecasting period plotted in Figure 4.12.

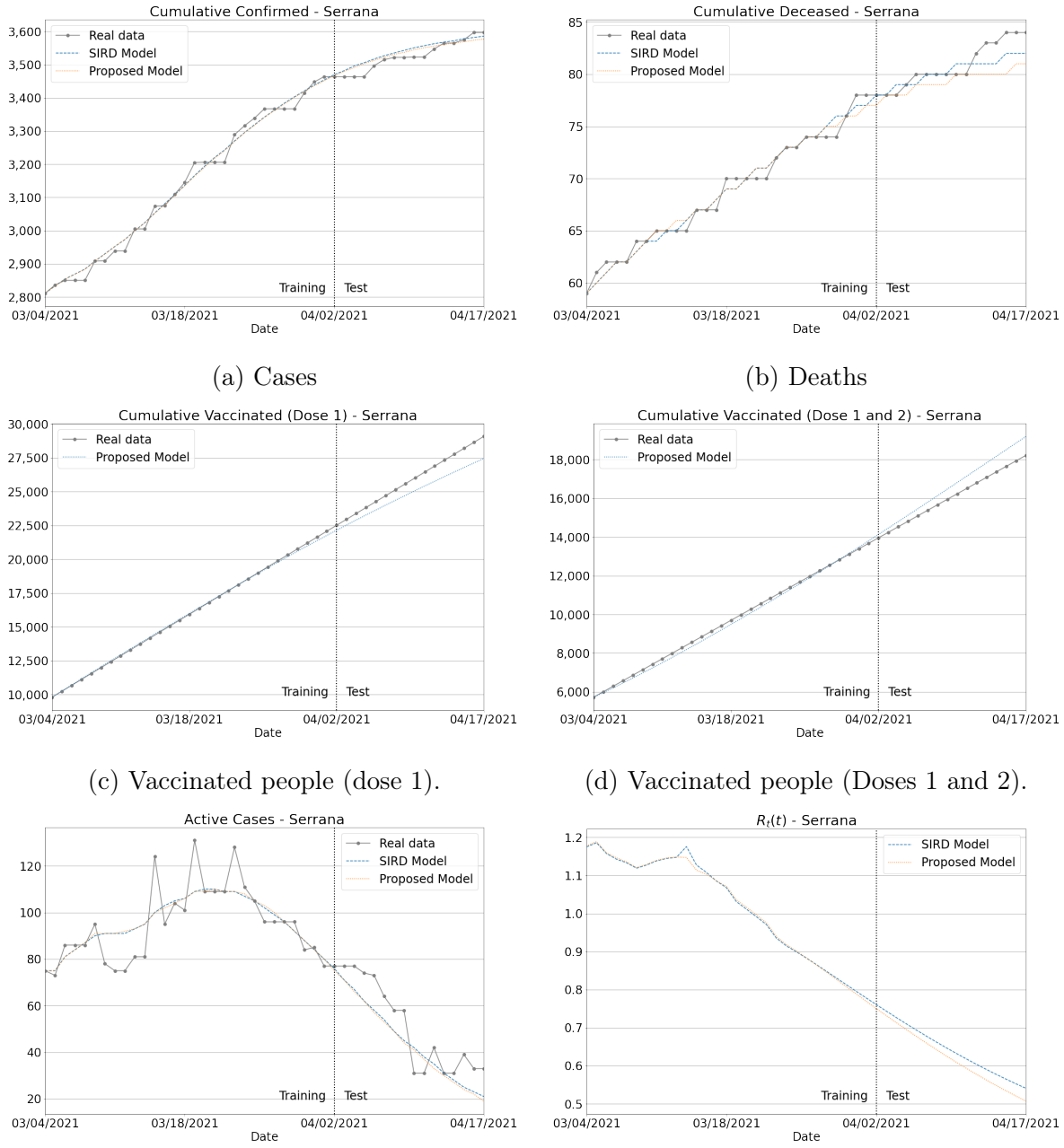
Variable	**Metric**	**Proposed Model**	**Data-Driven SIRD**
Cases	MAPE	0.370	0.276
	NRMSE	0.004	0.003
Deaths	MAPE	0.467	0.389
	NRMSE	0.006	0.005
Vaccinations (Dose 1)	MAPE	0.888	-
	NRMSE	0.011	-
Vaccinations (Doses 1 and 2)	MAPE	1.476	-
	NRMSE	0.009	-

4.4.1.2 Serrana's town vaccination campaign

Figure 4.13 presents the fitting and forecasting results concerning Serrana's massive vaccination campaign. Since the adult residents in Serrana have been immunized with Sinovac vaccine (Coronavac), we assume θ_1 and θ_2 as in Table 4.6. By visually inspecting the results, both methods follow the trajectories of real data. Moreover, the models have learned the changing trend of active cases, which is reflected by the decay of $R_t(t)$ to below 1.0 in the second half of March. The vaccination curves were also captured by our approach, even at high immunization rates as promoted by *Project S* [82]: over 67% of Serrana's adult population has been vaccinated with one shot of Coronavac until March 3rd. Finally, the quality metrics listed in Table 4.7 indicate a good agreement between the real and estimated data, as the highest MAPE and NRMSE values are much smaller than the acceptable values of 10% and 0.2, respectively.

4.4.2 Simulation-Based Vaccination Scenarios

We now assess the impact of different vaccination scenarios for Serrana's town, São Paulo State, and Brazil. More precisely, we simulate a variety of potentially realistic scenarios of vaccinations against new coronavirus, by varying in our data-driven SIR technique both the vaccination rates and immunization efficacies.



(a) Cases

(b) Deaths

(c) Vaccinated people (dose 1).

(d) Vaccinated people (Doses 1 and 2).

(e) Active cases

(f) $R_t(t)$

Figure 4.13: Serrana's town validation results. (a) Confirmed cases, (b) deaths, (c) vaccinated people who received at least one vaccine dose, (d) two doses, (e) active cases, and (f) effective reproduction number. Training period from March 4$^{\text{th}}$ 2021 to April 2$^{\text{nd}}$ 2021.

4.4.2.1 Assessing the impact of vaccination speed and vaccine efficacy on a fully vaccinated population

We start by investigating how the vaccination speed can influence the epidemic curves in a population almost fully immunized as the one living in Serrana's town. For this purpose, we take two evaluation scenarios: (i) the real data from Serrana, including its genuine vaccination rates as quickly leveraged by *Project S* with Coronavac vaccine, and (ii) the same data as (i), but now replacing the Serrana's vaccination rates for another time-series following the "standard" vaccination rollout as observed in Dracena, which is another small town in the state of São Paulo holding the same number of inhabitants as

Table 4.8: Average MAPE and NRMSE for the two-week forecasting period plotted in Figure 4.13.

Variable	Metric	Proposed Model	Data-Driven SIRD
Cases	MAPE	0.449	0.537
	NRMSE	0.005	0.006
Deaths	MAPE	1.733	1.131
	NRMSE	0.023	0.015
Vaccinations (Dose 1)	MAPE	3.520	-
	NRMSE	0.038	-
Vaccinations (Doses 1 and 2)	MAPE	3.404	-
	NRMSE	0.037	-

Serrana. We also simulate the use of several vaccine efficacies according to those reported by popular companies of anti-COVID-19 immunizers (see Table 4.6).

The blue and orange lines in Figures 4.14(a)-(c) give the learned data and the future estimates w.r.t. a time horizon of two-months for scenarios (i) and (ii), respectively. The vaccination speed has positively affected both confirmed cases and deaths. Indeed, after two months, it is expected that the "orange campaign" reaches around 140 deaths against 89 from the blue one representing the real immunization program of Serrana. Therefore, the reduction between the two vaccination scenarios in terms of the total number of COVID-19 fatalities is around 38%. Concerning Figure 4.14(c), immunizing faster causes $R_t(t)$ to fall 29% after two months, which is another plus when accelerating the vaccination.

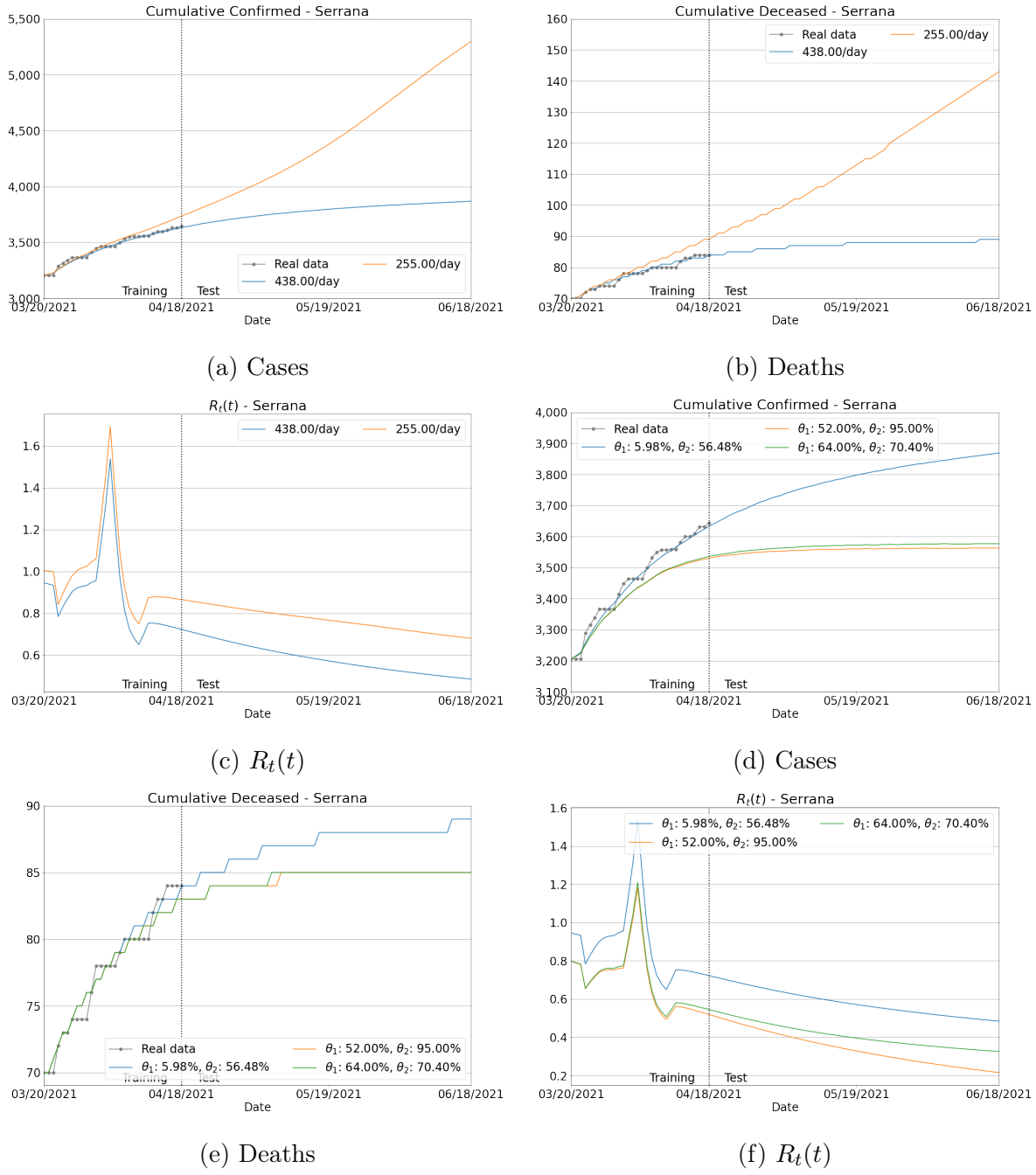
The performance of three vaccine efficacies, Coronavac (blue), AstraZeneca (orange) and Pfizer (green), are assessed in Figures 4.14(d)-(f). Due to the moderate efficacy of Coronavac, it achieves the highest number of cases, while AstraZeneca and Pfizer retain the virus spreading more efficiently. At the end of the forecasting interval, the estimates from orange and green lines produce 292 and 306 cases less than the blue one. In contrast to the reduction of cases, deaths avoided by all three immunizers remain at the same level as time advances, thus indicating they are capable of ensuring high protection against COVID-19 mortality. Finally, $R_t(t)$ assigned to Coronavac decreases by 31% and 54% when it is compared with AstraZeneca and Pfizer, respectively.

4.4.2.2 Assessing the impact of vaccination speed on a partially immunized population

We now evaluate the impact of different immunization rates concerning a much bigger population: São Paulo, which is the most populous state in Brazil, home to around 46 million people, i.e., the same as Spain. Notice that the vaccination rollout in São Paulo is still in progress so that the number of vaccinated people with at least one vaccine dose only reached 20% in recent days. To design this experiment, in Figure 4.15 we take the real data from March 3rd to April 1nd, to train the model, and then simulate the next three months of the pandemic. Four vaccination speed rates were simulated: the blue one, as taken from real data, and the rates set to 0.5x (orange), 2x (green), and 5x (red) w.r.t. the actual speed. Confirmed cases slowly decrease as the vaccination advances, except if one compares the actual speed against the 5x campaign: the cases drop from around 3.9 to 3.4 million, a reduction of 13%. Similarly, deaths are substantially mitigated as more people are vaccinated in an increasingly short period, dropping from $141,000$ (blue) to $131,000$

(green) and 111,000 (red). Finally, by accelerating the vaccination, the transmission rate is also pulled down. Thus, this experiment confirms how important it is to speed up the vaccination, as properly done in Israel, UK, and USA.

4.4.2.3 Assessing the impact of different COVID-19 vaccines



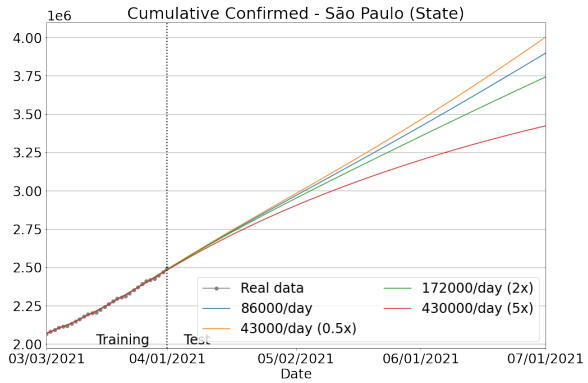
(a) Cases (b) Deaths

(c) $R_t(t)$ (d) Cases

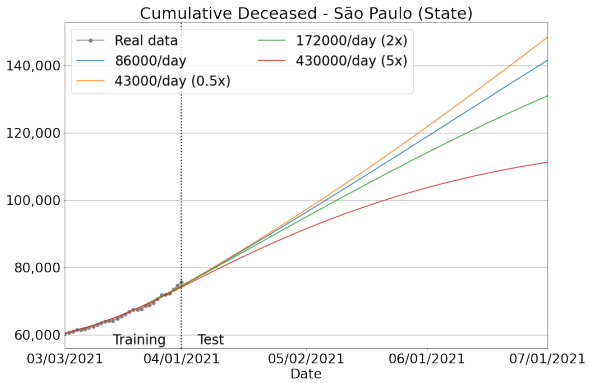
(e) Deaths (f) $R_t(t)$

Figure 4.14: Serrana's town scenarios results. (a)-(d) Confirmed cases, (b)-(e) deaths, and (c)-(f) effective reproduction number. Training period from March 20$^{\text{th}}$ 2021 to April 18$^{\text{th}}$ 2021.

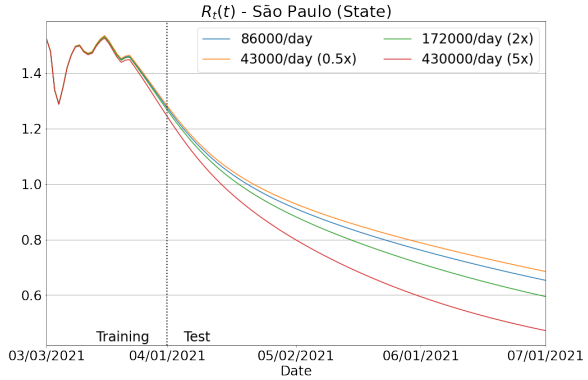
In Figure 4.16, we assess how the vaccine efficacies can dictate the infectivity of SARS-CoV-2 in São Paulo and Brazil, as the country holds the second-highest number of COVID-19 deaths globally. Together with Coronavac (purple), AstraZeneca (red) and Pfizer (green) efficacies, we take two combinations of mixed efficacies (blue and or-



(a) Cases



(b) Deaths



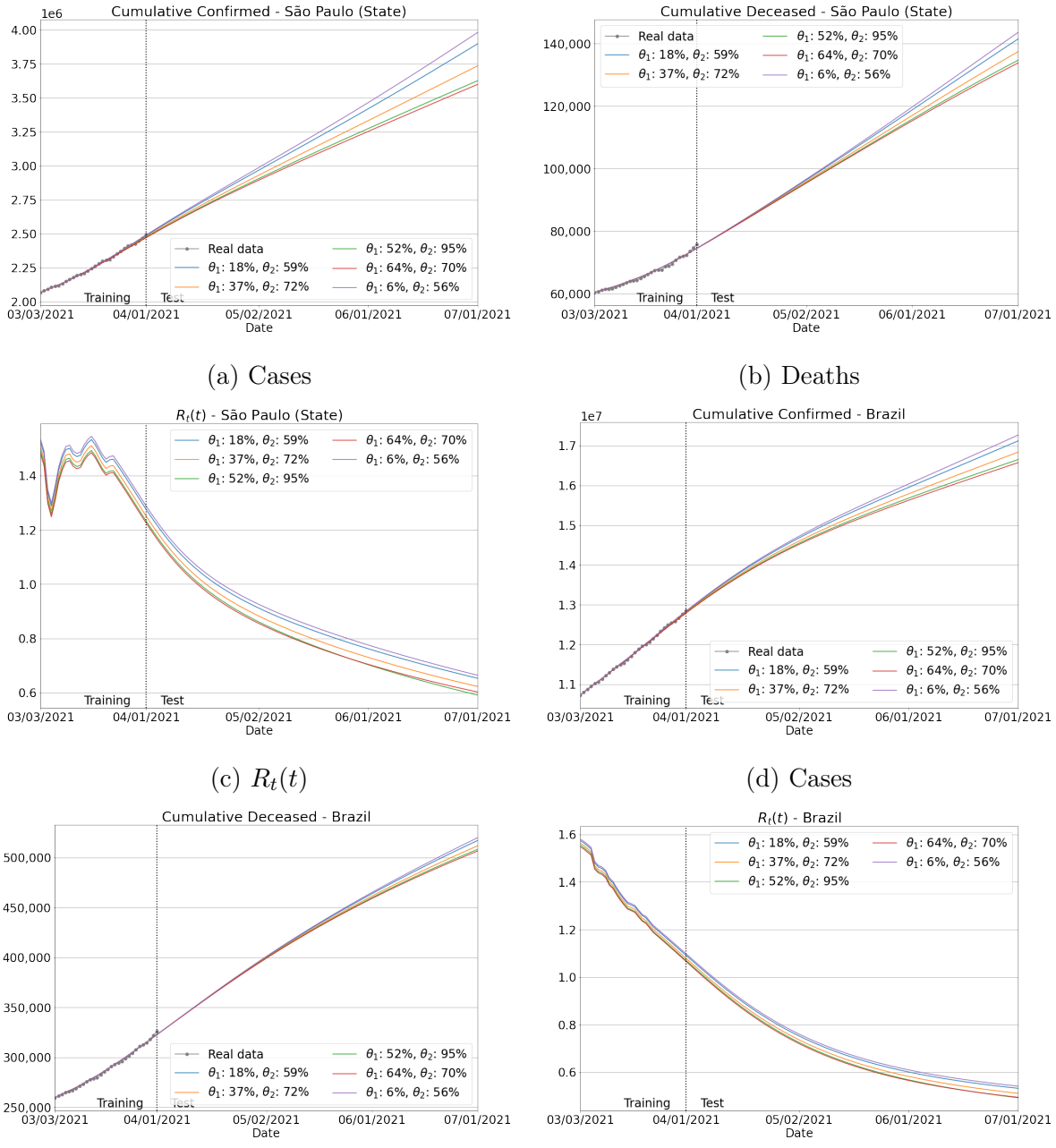
(c) $R_t(t)$

Figure 4.15: São Paulo state vaccination velocity scenarios. (a) Confirmed cases, (b) deaths, and (c) effective reproduction number. Training period from March 3rd 2021 to April 1st 2021.

ange), to simulate the vaccination rollouts with multiple types of vaccines in São Paulo and Brazil. Table 4.9 lists several combinations of vaccines, by computing the weighted average between their real efficacies and the number of administrated doses.

Table 4.9: Mixed vaccine proportions and their resultant efficacies used to run the experiments in Figures 4.16 and 4.17. Brazilian vaccine distribution (in blue) taken from *Ministry of Health - Brazil* [54] on March 31, 2021.

Coronavac	**AstraZeneca**	**Pfizer**	**Resultant Efficacy**	**Color**
80%	20%	0%	$\theta_1 = 18\%$ $\theta_2 = 59\%$	blue
40%	30%	30%	$\theta_1 = 37\%$ $\theta_2 = 72\%$	orange
100%	0%	0%	$\theta_1 = 6\%$ $\theta_2 = 56\%$	purple
0%	100%	0%	$\theta_1 = 64\%$ $\theta_2 = 70\%$	red
0%	0%	100%	$\theta_1 = 52\%$ $\theta_2 = 95\%$	green



(a) Cases (b) Deaths

(c) $R_t(t)$ (d) Cases

(e) Deaths (f) $R_t(t)$

Figure 4.16: São Paulo state and Brazil different COVID-19 vaccines scenarios. (a)-(d) Confirmed cases, (b)-(e) deaths, and (c)-(f) effective reproduction number. Training period from March 3rd 2021 to April 1st 2021.

First, considering São Paulo's case study, efficacies related to AstraZeneca, Pfizer, and the combination $\theta_1 = 37\%$ with $\theta_2 = 72\%$ make the confirmed cases down more quickly compared to the purple and blue curves. Indeed, green and red campaigns perform similarly, producing the lowest number of cases, i.e., 362k and 360k against 398k from Coronavac. Moreover, the first dose efficacy indicates that it plays a key role in the immunization process. Deaths are also mitigated when taking vaccines of greater efficacy. For example, a campaign purely based on Coronavac vaccine reaches around $1.43K$ deaths at the end of the period against $1.34k$ and $1.33k$ deaths from campaigns with Pfizer and AstraZeneca. A similar pandemic signature is observed in Brazil: if one assumes the current scenario (blue line), from Figure 4.16 and Table 4.10, the highest reduction of cases

and deaths are delivered by AstraZeneca vaccine with the current immunization velocity, and with Pfizer efficacies when boosting 5x the current vaccination speed. Finally, $R_t(t)$ significantly drops more than 50% if speed is increased by 5x with $\theta_1 = 52\%$ and $\theta_2 = 95\%$.

Table 4.10: Vaccine efficacy comparison from the current scenario ($\theta_1 = 18\%$, $\theta_2 = 59\%$ - Blue line from Fig. 4.16).

Efficacy	**Cases**		**Deaths**	
	São Paulo	**Brazil**	**São Paulo**	**Brazil**
Baseline	1,395,106	4,217,642	66,339	191,676
$\theta_1 = 37\%$ $\theta_2 = 72\%$ (Orange)	1,241,105 (-11.0%)	3,955,485 (-6.2%)	62,326 (-6.0%)	186,396 (-2.8%)
$\theta_1 = 52\%$ $\theta_2 = 95\%$ (Green)	1,137,064 (-18.5%)	3,783,811 (-10.3%)	59,605 (-10.2%)	182,920 (-4.6%)
$\theta_1 = 64\%$ $\theta_2 = 70\%$ (Red)	1,113,243 (-20.2%)	3,713,921 (-11.9%)	58,743 (-11.5%)	181,273 (-5.4%)
$\theta_1 = 6\%$ $\theta_2 = 56\%$ (Purple)	1,474,225 (+5.7%)	4,353,357 (+3.2%)	68,407 (+3.1%)	194,424 (+1.4%)

4.4.2.4 Assessing the impact on increasing the vaccination speed as the vaccine efficacy changes

Next, we evaluate the combined impact of accelerating the vaccination speed as θ_1 and θ_2 vary. To design this experiment, we assume in Figure 4.17 that the vaccination advances 5x faster than the real immunization campaign in São Paulo and Brazil. In contrast to the results taking the current vaccination rates as previously discussed in curves of Figure 4.16 (associated with Table 4.10), the efficacies in Figure 4.17 (associated with Table 4.11) significantly contribute to the flattening of curves, making the gaps between them much more prominent with the increased vaccination. In addition, from Table 4.11 one can check that the reduction in cases and $R_t(t)$ are more influenced than deaths, as expected given the COVID-19 vaccines are highly effective against deaths.

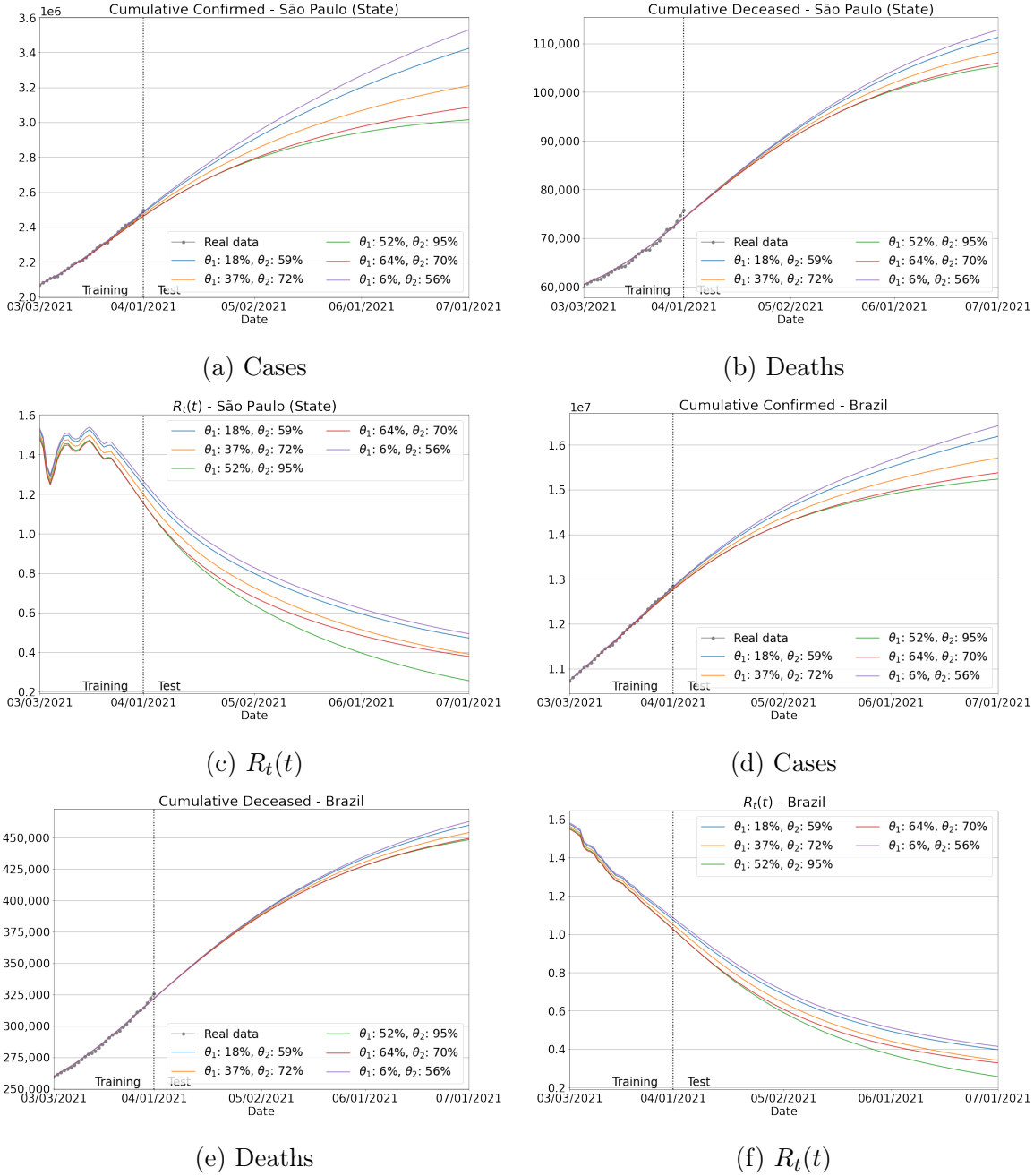
Finally, in Table 4.12 we discuss the reduction of COVID-19 rates in São Paulo and Brazil for a three-month forecasting period with their actual immunization campaigns against the hypothetical scenarios with the vaccination speed increased by 5x. Notice that the total number of cases and deaths are significantly attenuated as more people are vaccinated over the period. In fact, even though the most significant falls have been found with the Pfizer vaccine for confirmed cases, all immunizers clearly reduce deaths to similar levels, especially in Brazil (see the last column in Table 4.12).

Table 4.11: Vaccine efficacy comparison by taking as baseline the blue line from Fig. 4.17.

Efficacy	**Cases**		**Deaths**	
	São Paulo	**Brazil**	**São Paulo**	**Brazil**
Baseline	925,808	3,303,194	36,511	135,479
$\theta_1 = 37\%$ $\theta_2 = 72\%$ (Orange)	722,785 (-21.9%)	2,848,722 (-13.8%)	33,476 (-8.3%)	129,701 (-4.3%)
$\theta_1 = 52\%$ $\theta_2 = 95\%$ (Green)	538,536 (-41.8%)	2,404,922 (-27.2%)	30,648 (-16.1%)	124,115 (-8.4%)
$\theta_1 = 64\%$ $\theta_2 = 70\%$ (Red)	610,734 (-34.0%)	2,549,897 (-22.8%)	31,348 (-14.1%)	125,126 (-7.6%)
$\theta_1 = 6\%$ $\theta_2 = 56\%$ (Purple)	1,026,201 (+10.8%)	3,524,649 (+6.7%)	38,043 (+4.2%)	138,380 (+2.1%)

Table 4.12: Vaccine efficacy comparison by taking as baseline the current scenarios in São Paulo and Brazil.

Efficacy	**Cases**		**Deaths**	
	São Paulo	**Brazil**	**São Paulo**	**Brazil**
Baseline	1,395,106	4,217,642	66,339	191,676
$\theta_1 = 18\%$ $\theta_2 = 59\%$ (Blue)	925,808 (-33.6%)	3,303,194 (-21.7%)	36,511 (-45.0%)	135,479 (-29.3%)
$\theta_1 = 37\%$ $\theta_2 = 72\%$ (Orange)	722,785 (-48.2%)	2,848,722 (-32.5%)	33,476 (-49.5%)	129,701 (-32.3%)
$\theta_1 = 52\%$ $\theta_2 = 95\%$ (Green)	538,536 (-61.4%)	2,404,922 (-43.0%)	30,648 (-53.8%)	124,115 (-35.2%)
$\theta_1 = 64\%$ $\theta_2 = 70\%$ (Red)	610,734 (-56.2%)	2,549,897 (-39.5%)	31,348 (-52.7%)	125,126 (-34.7%)
$\theta_1 = 6\%$ $\theta_2 = 56\%$ (Purple)	1,026,201 (-26.4%)	3,524,649 (-16.4%)	38,043 (-42.7%)	138,380 (-27.8%)



(a) Cases

(b) Deaths

(c) $R_t(t)$

(d) Cases

(e) Deaths

(f) $R_t(t)$

Figure 4.17: São Paulo state and Brazil different COVID-19 vaccines and vaccination velocity scenarios. (a)-(d) Confirmed cases, (b)-(e) deaths, and (c)-(f) effective reproduction number. Training period from March 3$^{\text{rd}}$ 2021 to April 1$^{\text{st}}$ 2021.

Chapter

5

Conclusion

In this work, we propose different data-driven responses against COVID-19 outbreak for São Paulo state and Brazil. The investigations include a totally free interactive platform for tracking coronavirus-related data, a novel SIRD-based mathematical model which learns epidemiological parameters to best fit the corresponding data of each analyzed region, and a comprehensive experimental evaluation of both past and the current situation of the pandemic in Brazil and state of São Paulo.

As discussed in Section 4.3.4 and in Appendix A, the tracking platform – *Info Tracker*, Firstly developed by Wallace Casaca and Marilaine Colnago, – has supported public authorities, society and press agencies in better understanding and exploiting COVID-19 data, by intuitively interacting with them through a simple and easy-to-communicate interface. This work helped to further improve the platform with the inclusion of predictions of some COVID-19 data. We have found that the predictions matched the true data both qualitatively and quantitatively. As shown in the battery of tests, our unifying SIRD + machine learning approach produced considerably low MAPE and RMSE errors, as shown in Table 4.4 and 4.5. Indeed, MAPEs were less than 1 in almost all the measurements. Another important aspect noted in the experiments is that the trained forecaster turned out to be very effective and robust when dealing with badly-conditioned data, as shown in Section 4.3.1 (see the listed variances in Table 4.1).

We discussed in Section 4.3.4.3 how the predictions can be successfully used to assess the impact of a second COVID-19 wave starting in São Paulo and Brazil, warning about the sudden growth of new cases of coronavirus so as to put health authorities and the country's population on alert for the coming weeks. Additionally, we have discussed in Section 4.3.4.3 the application of the data-driven model for predicting the COVID-19 spread on Italy, Portugal and Ukraine.

We provided several simulation-based evaluations concerning the pros and cons of COVID-19 vaccination campaigns in Brazil, São Paulo State, and Serrana's town. The analysis focused on assessing the impact of the immunization speed and vaccine efficacy for at least three types of immunizers in the epidemic curves of confirmed cases, deaths and infectivity rate.

As discussed in Section 4.4, the use of different vaccine types indicated that the reduction in confirmed cases is more pronounced than in the deceased. In fact, we found that the protection against SARS-CoV-2 deaths is similar among all immunizers, in line with published clinical studies. Another finding is that the speed in administrating new shots of vaccine is of paramount importance to pull down the deaths and the infectivity levels of the disease, even with COVID-19 immunizers holding moderate efficacy. For example, we have found that confirmed cases and deaths in Brazil may be pruned to around 16%

and 27% by adopting an immunization campaign purely carried out with a vaccine of moderate efficacy as long as the speed of vaccination is accelerated.

As shown in Fig. 4.16 and 4.17, our methodology can be successfully used to perform numerical investigation concerning the recent strategy of mix-and-match vaccination, as the one in progress in the UK according to the Com-COV Study Team [75].

This work resulted in two papers, entitled "Towards Providing Effective Data-Driven Responses to Predict the COVID-19 in São Paulo and Brazil" [5] and "Simulating Immunization Campaigns and Vaccine Protection Against COVID-19 Pandemic in Brazil" [4]

Finally, we show that the proposed method can be adapted to many mathematical models if there are enough data to support the parameters fitting. This opens a hall of possibilities to analyse more complex models, like reinfection models, in order to further improve forecasting and parameter calibration.

Bibliography

[1] A.D.Dongare, R.R.Kharde, and Amit D.Kachare. Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, 2(1), 7 2012.

[2] M. E. Alexander, C. Bowman, S. M. Moghadas, R. Summers, A. B. Gumel, and B. M. Sahai. A vaccination model for transmission dynamics of influenza. *SIAM Journal on Applied Dynamical Systems*, 3(4):503–524, 2004. doi: 10.1137/030600370. URL `https://doi.org/10.1137/030600370`.

[3] A. Alsayed, H. Sadir, R. Kamil, and H. Sari. Prediction of epidemic peak and infected cases for COVID-19 disease in malaysia, 2020. *International Journal of Environmental Research and Public Health*, 17:4076, 2020.

[4] Fabio Amaral, Wallace Casaca, Cassio M. Oishi, and José A. Cuminato. Simulating immunization campaigns and vaccine protection against covid-19 pandemic in brazil. *IEEE Access*, 2021. Submitted.

[5] Fabio Amaral, Wallace Casaca, Cassio M. Oishi, and José A. Cuminato. Towards providing effective data-driven responses to predict the COVID-19 in São Paulo and Brazil. *Sensors*, 21(2), 2021.

[6] Cleo Anastassopoulou, Lucia Russo, Athanasios Tsakris, and Constantinos Siettos. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLOS ONE*, 15(3):1–21, 2020.

[7] Dario Ortega Anderez, Eiman Kanjo, Ganna Pogrebna, Omprakash Kaiwartya, Shane D. Johnson, and John Alan Hunt. A COVID-19-based modified epidemiological model and technological approaches to help vulnerable individuals emerge from the lockdown in the uk. *Sensors*, 20(17), 2020.

[8] Estela M. L. Aquino, Ismael Henrique Silveira, Julia Moreira Pescarini, Rosana Aquino, Jaime Almeida de Souza-Filho, Aline dos Santos Rocha, Andrea Ferreira, Audêncio Victor, Camila Teixeira, Daiane Borges Machado, Enny Paixao, Flavia Jase Oliveira Alves, Flávia Pilecco, Greice Menezes, Ligia Gabrielli, Luciana Leite, Maria da Conceição Chagas de Almeida, NaiÃ¡ Ortelan, Qeren Hapuk R. Ferreira Fernandes, Renzo Joel Flores Ortiz, Raquel Nunes Palmeira, Elzo Pereira Pinto Junior, Erika Aragao, Luis Eugenio Portela Fernandes de Souza, Manoel Barral Netto, Maria Glória Teixeira, Mauricio Lima Barreto, Maria Yury Ichihara, and RaÃza Tourinho dos Reis Silva Lima. Social distancing measures to control the COVID-19 pandemic: potential impacts and challenges in brazil. *Ciencia & Saude Coletiva*, 25: 2423–2446, 2020.

[9] Andrew Atkeson, Karen Kopecky, and Tao Zha. *Estimating and Forecasting Disease Scenarios for COVID-19 with an SIR Model*. NBER Working Paper, Cambridge, Massachusetts, 2020.

[10] Andrew G. Atkeson. On using sir models to model disease scenarios for COVID-19. *Quarterly Review*, 41(1):1–33, 2020.

[11] Saulo B. Bastos and Daniel O. Cajueiro. Modeling and forecasting the early evolution of the COVID-19 pandemic in brazil, 2020.

[12] BBC News Brazil. Hospital doctors in SP raise the alarm for new wave of COVID-19, 2020. [Online] Available `https://newsbeezer.com/brazileng/hospital-doctors-in-sp-raise-the-alarm-for-new-wave-of-{COVID-19}/`.

[13] Soulaimane Berkane, Intissar Harizi, and Abdelhamid Tayebi. Modeling the effect of population-wide vaccination on the evolution of COVID-19 epidemic in canada. *medRxiv*, 2021. doi: 10.1101/2021.02.05.21250572. URL `https://www.medrxiv.org/content/early/2021/04/05/2021.02.05.21250572`.

[14] Kathakali Biswas, Abdul Khaleque, and Parongama Sen. COVID-19 spread: Reproduction of data and prediction using a sir model on euclidean network, 2020.

[15] Yuri Bruinen de Bruin, Anne-Sophie Lequarre, Josephine McCourt, Peter Clevestig, Filippo Pigazzani, Maryam Zare Jeddi, Claudio Colosio, and Margarida Goulart. Initial impacts of global risk mitigation measures taken during the combatting of the COVID-19 pandemic. *Safety Science*, 128:104773, 2020.

[16] R. H. Byrd, S. L. Hansen, Jorge Nocedal, and Y. Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

[17] Giacomo Cacciapaglia, Corentin Cot, and Francesco Sannino. Second wave COVID-19 pandemics in europe: a temporal playbook. *Scientific Reports*, 10(1):15514, 2020.

[18] Ewen Callaway. Mix-and-match COVID vaccines trigger potent immune response, 5 2021. URL `https://www.nature.com/articles/d41586-021-01359-3`.

[19] CanalTech Press Agency (Translated by Time 24 News). Public hospitals face increase in admissions for COVID-19 in sp, 2020. [Online] Available `https://www.time24.news/2020/11/public-hospitals-face-increase-in-admissions-for-{COVID-19}-in-sp.html`.

[20] Yi-Cheng Chen, Ping-En Lu, Cheng-Shang Chang, and Tzu-Hsuan Liu. A time-dependent sir model for COVID-19 with undetectable infected persons. *IEEE Transactions on Network Science and Engineering*, pages 1—-1, 2020.

[21] Jonathan Cohen, William G. Powderly, and Steven M. Opal. *Infectious Diseases (Fourth Edition)*. Elsevier, Amsterdam, The Netherlands, 2017.

[22] Hector Cornide-Reyes, Fabián Riquelme, Diego Monsalves, Rene Noel, Cristian Cechinel, Rodolfo Villarroel, Francisco Ponce, and Roberto Munoz. A multimodal real-time feedback platform based on spoken interactions for remote active learning support. *Sensors*, 20(21):6337, Nov 2020.

[23] Raj Dandekar and George Barbastathis. Neural network aided quarantine control model estimation of covid spread in wuhan, china, 2020.

[24] Manuel De la Sen, Santiago Alonso-Quesada, and Asier Ibeas. On a discrete seir epidemic model with exposed infectivity, feedback vaccination and partial delayed resusceptibility. *Mathematics*, 9(5), 2021. ISSN 2227-7390. doi: 10.3390/math9050520. URL `https://www.mdpi.com/2227-7390/9/5/520`.

[25] Dong E and Du Hand Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Inf Dis.*, 2 2020. doi: 10.1016/S1473-3099(20)30120-1.

[26] Shahul H Ebrahim, Qanta A Ahmed, Ernesto Gozzer, Patricia Schlagenhauf, and Ziad A Memish. COVID-19 and community mitigation strategies in a pandemic. *BMJ*, 368:1–2, 2020.

[27] Malen Etxeberria-Etxaniz, Santiago Alonso-Quesada, and Manuel De la Sen. On an seir epidemic model with vaccination of newborns and periodic impulsive vaccination with eventual on-line adapted vaccination strategies to the varying levels of the susceptible subpopulation. *Applied Sciences*, 10(22), 2020. ISSN 2076-3417. doi: 10.3390/app10228296. URL `https://www.mdpi.com/2076-3417/10/22/8296`.

[28] FAPESP - São Paulo Research Foundation. A dimensão da pandemia (in portuguese), 2020. [Online] Available `https://revistapesquisa.fapesp.br/a-dimensao-da-pandemia`.

[29] Davide Faranda and Tommaso Alberti. Modeling the second wave of COVID-19 infections in france and italy via a stochastic SEIR model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11):111101, 2020.

[30] Folha de S. Paulo. Problems in the ministry of health's system hamper analysis of COVID-19 data, 2020. [Online] Available (In Portuguese) `https://www1.folha.uol.com.br/equilibrioesaude/2020/08/problemas-em-sistema-do-ministerio-da-saude-prejudicam-analise-de-dados-da-{COVID-19}.shtml`.

[31] Elisabeth Barboza França, Lenice Harumi Ishitani, Renato Azeredo Teixeira, Daisy Maria Xavier de Abreu, Paulo Roberto Lopes Corrêa, Fatima Marinho, and Ana Maria Nogales Vasconcelos. Deaths due to COVID-19 in brazil: how many are there and which are being identified? *Revista Brasileira de Epidemiologia*, 23:1–7, 2020.

[32] Miguel Fudolig and Reka Howard. The local stability of a modified multi-strain sir model for emerging viral strains. *PLOS ONE*, 15(12):1–27, 12 2020. doi: 10.1371/journal.pone.0243408. URL `https://doi.org/10.1371/journal.pone.0243408`.

[33] Behzad Ghanbari. On forecasting the spread of the COVID-19 in iran: The second wave. *Chaos, Solitons & Fractals*, 140:110176, 2020.

[34] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[35] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4): 599—-653, 2000.

[36] H. G. Hong and Y. Li. Estimation of time-varying reproduction numbers underlying epidemiological processes: A new statistical tool for the COVID-19 pandemic. *PLos ONE*, 15:e0236464, 2020.

[37] Instituto Butantan. Instituto butantan, 2021. [Online] Available `https://butantan.gov.br/instituto-butantan/about-us`.

[38] O. Istaiteh, T. Owais, N. Al-Madi, and S. Abu-Soud. Machine learning approaches for COVID-19 forecasting. In *International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 50–57, 2020.

[39] Hyeontae Jo, Hwijae Son, Se Young Jung, and Hyung Ju Hwang. Analysis of COVID-19 spread in south korea using the sir model with time-dependent parameters and deep learning. *medRxiv*, 2020. doi: 10.1101/2020.04.13.20063412. URL `https://www.medrxiv.org/content/early/2020/04/17/2020.04.13.20063412`.

[40] Johns Hopkins University. COVID-19 data repository by the center for systems science and engineering (csse) at johns hopkins university, 2020. [Online] Available `https://github.com/CSSEGISandData/{COVID-19}`.

[41] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A*, 115(772):700–721, 1927.

[42] Mehrdad Kiamari, Gowri Ramachandran, Quynh Nguyen, Eva Pereira, Jeanne Holm, and Bhaskar Krishnamachari. COVID-19 risk estimation using a time-varying sir-model, 2020.

[43] Shengjie Lai, Nick W. Ruktanonchai, Liangcai Zhou, Olivia Prosper, Wei Luo, Jessica R. Floyd, Amy Wesolowski, Mauricio Santillana, Chi Zhang, Xiangjun Du, Hongjie Yu, and Andrew J. Tatem. Effect of non-pharmaceutical interventions to contain COVID-19 in china. *Nature*, 585:410–413, 2020.

[44] Joao Vitor Leme, Wallace Casaca, Marilaine Colnago, and Maurício A. Dias. Towards assessing the electricity demand in brazil: Data-driven analysis and ensemble learning models. *Energies*, 13(6):1407, 2020.

[45] Lianyi Liu, Yan Chen, and Lifeng Wu. Forecasting confirmed cases, deaths, and recoveries from COVID-19 in china during the early stage. *Mathematical Problems in Engineering*, 2020(4):1405764, 2020.

[46] Xiya Ma and Dominique Vervoort. Critical care capacity during the COVID-19 pandemic: Global availability of intensive care beds. *Journal of Critical Care*, 58: 96–97, 2020.

[47] Ho-Yin Mak, Tinglong Dai, and Christopher S. Tang. Managing two-dose COVID-19 vaccine rollouts with limited supply. *SSRN*, 2 2021. doi: http://dx.doi.org/10.2139/ssrn.3790836. URL `https://ssrn.com/abstract=3790836`.

[48] Battegay Manuel, Kuehl Richard, Tschudin-Sutter Sarah, Hirsch Hans H, Widmer Andreas F., and Neher Richard A. 2019-novel coronavirus (2019-ncov): estimating the case fatality rate – a word of caution. *Swiss Medical Weekly*, 150(w20203): 1–3, 2020.

[49] Manuel Martinez, Kailun Yang, Angela Constantinescu, and Rainer Stiefelhagen. Helping the blind to get through COVID-19: Social distancing assistant using real-time semantic segmentation on rgb-d video. *Sensors*, 20(18):5202, Sep 2020.

[50] Kunwer Singh Mathur and Prakash Narayan. Dynamics of an SVEIRS epidemic model with vaccination and saturated incidence rate. *International Journal of Applied and Computational Mathematics*, 4(5):118, Aug 2018. ISSN 2199-5796. doi: 10.1007/s40819-018-0548-0. URL https://doi.org/10.1007/s40819-018-0548-0.

[51] Camila Carvalho de Souza Amorim Matos, Carolina Luísa Alves Barbieri, and Marcia Thereza Couto. COVID-19 and its impact on immunization programs: reflections from brazil. *Revista de Saúde Pública*, 54:114, 12 2020.

[52] Esteban Ortiz-Ospina Max Roser, Hannah Ritchie and Joe Hasell. Coronavirus pandemic (COVID-19). *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

[53] Ministry of Health (Brazil). Brazilian coronavirus disease (COVID-19) dashboard, 2020. [Online] Available https://covid.saude.gov.br.

[54] Ministry of Health (Brazil). COVID-19 vaccination: Applied doses dashboard, 2021. [Online] Available https://qsprod.saude.gov.br/extensions/DEMAS_C19Vacina/DEMAS_C19Vacina.html.

[55] Counotte MJ, Althaus CL, Low N, and Riou J. Impact of age-specific immunity on the timing and burden of the next zika virus outbreak. *PLoS Negl Trop Dis.*, 12 2019. doi: 10.1371/journal.pntd.0007978.

[56] Youssoufa Mohamadou, Aminou Halidou, and Pascalin Tiam Kapen. A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. *Applied Intelligence*, 50:3913–3925, 2020.

[57] Babacar Mbaye Ndiaye, Lena Tendeng, and Diaraf Seck. Analysis of the COVID-19 pandemic by sir model and machine learning technics for forecasting, 2020.

[58] Yutaka Okabe and Akira Shudo. A mathematical model of epidemics — a tutorial for students. *Mathematics*, 8(7):1174, 2020.

[59] Olha Digital. Platform publishes real-time data on COVID-19's progress in sp, 2020. [Online] Available https://olhardigital.com.br/en/2020/06/22/news/platform-publishes-real-time-data-on-{COVID-19}-progress-in-sp/amp/.

[60] Linda Petzold. Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM Journal on Scientific and Statistical Computing*, 4(1):136–148, 1983.

[61] Q. Pham, D. C. Nguyen, T. Huynh-The, W. Hwang, and P. N. Pathirana. Artificial intelligence (ai) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts. *IEEE Access*, 8:130820–130839, 2020.

[62] Ben Phillips, Melissa Makin, and Uli Felzmann. Coronavirus 10-day forecast, 2020. URL http://covid19forecast.science.unimelb.edu.au/. accessed in August, 2021.

[63] Fernando P. Polack, Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez, Gonzalo Pérez Marc, Edson D. Moreira, Cristiano Zerbini, Ruth Bailey, Kena A. Swanson, Satrajit Roychoudhury, Kenneth Koury, Ping Li, Warren V. Kalina, David Cooper, Robert W. Frenck, Laura L.

Hammitt, Özlem Türeci, Haylene Nell, Axel Schaefer, Serhat Ünal, Dina B. Tresnan, Susan Mather, Philip R. Dormitzer, Uğur Şahin, Kathrin U. Jansen, and William C. Gruber. Safety and efficacy of the bnt162b2 mrna COVID-19 vaccine. *New England Journal of Medicine*, 383(27):2603–2615, 2020. doi: 10.1056/NEJMoa2034577. URL https://doi.org/10.1056/NEJMoa2034577. PMID: 33301246.

[64] Nikolaos P. Rachaniotis, Thomas K. Dasaklis, Filippos Fotopoulos, and Platon Tinios. A two-phase stochastic dynamic model for COVID-19 mid-term policy recommendations in greece: A pathway towards mass vaccination. *International Journal of Environmental Research and Public Health*, 18(5), 2021. ISSN 1660-4601. doi: 10.3390/ijerph18052497. URL https://www.mdpi.com/1660-4601/18/5/2497.

[65] Karthikeyan Rajagopal, Navid Hasanzadeh, Fatemeh Parastesh, Ibrahim Ismael Hamarash, Sajad Jafari, and Iqtadar Hussain. A fractional-order model for the novel coronavirus (COVID-19) outbreak. *Nonlinear Dynamics*, 101(1):711–718, 2020.

[66] RecordTV (R7) Television Network (Translated by Time 24 News). State of SP has 15 regions with accelerated transmission of COVID-19, 2020. [Online] Available https://www.time24.news/2020/12/state-of-sp-has-15-regions-with-accelerated-transmission-of-{COVID-19}-news.html.

[67] Marissa Renardy, Marisa Eisenberg, and Denise Kirschner. Predicting the second wave of COVID-19 in washtenaw county, MI. *Journal of Theoretical Biology*, 507: 110461, 2020.

[68] Reuters. Brazil takes down COVID-19 data, hiding soaring death toll, 2020. [Online] Available https://www.reuters.com/article/us-health-coronavirus-brazil-idUSKBN23D0PW.

[69] Reuters. As COVID explodes in brazil, serrana becomes world's first fully vaccinated city, 2021. [Online] Available https://worldcrunch.com/coronavirus/as-covid-explodes-in-brazil-serrana-becomes-world39s-first-fully-vaccinated-city.

[70] Satyaki Roy, Ronojoy Dutta, and Preetam Ghosh. Optimal time-varying vaccine allocation amid pandemics with uncertain immunity ratios. *IEEE Access*, 9:15110–15121, 2021.

[71] David E. Rumelhart and James L. McClelland. *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press, 1987.

[72] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct 1986. ISSN 1476-4687. doi: 10.1038/323533a0. URL https://doi.org/10.1038/323533a0.

[73] Chadi M. Saad-Roy, Sinead E. Morris, C. Jessica E. Metcalf, Michael J. Mina, Rachel E. Baker, Jeremy Farrar, Edward C. Holmes, Oliver G. Pybus, Andrea L. Graham, Simon A. Levin, Bryan T. Grenfell, and Caroline E. Wagner. Epidemiological and evolutionary considerations of sars-cov-2 vaccine dosing regimes. *Science*, 2021. ISSN 0036-8075. doi: 10.1126/science.abg8663. URL https://science.sciencemag.org/content/early/2021/03/08/science.abg8663.

[74] Luiz Leduíno de Salles Neto, Camila Bertini Martins, Antônio Augusto Chaves, Thais Cláudia Roma de Oliveira Konstantyner, Horacio Hideki Yanasse, Claudia Barbosa Ladeira de Campos, Ana Júlia de Oliveira Bellini, Renan Brito Butkeraites, Leonardo Correia, Igor Luciano Magro, and Fernando dos Santos Soares. Forecast UTI: application for predicting intensive care unit beds in the context of the COVID-19 pandemic. *Epidemiologia e Serviços de Saúde*, 29, 2020.

[75] Robert H. Shaw, Arabella Stuart, Melanie Greenland, Xinxue Liu, Jonathan S. Nguyen Van-Tam, and Matthew D. Snape. Heterologous prime-boost COVID-19 vaccination: initial reactogenicity data. *The Lancet*, 397(10289):2043–2046, 5 2021. ISSN 0140-6736. doi: 10.1016/S0140-6736(21)01115-6. URL `https://doi.org/10.1016/S0140-6736(21)01115-6`.

[76] Paulo J. S. Silva, Claudia Sagastizábal, Luís Gustavo Nonato, Claudio José Struchiner, and Tiago Pereira. Optimized delay of the second covid-19 vaccine dose reduces icu admissions. *Proceedings of the National Academy of Sciences*, 118(35), 2021. ISSN 0027-8424. doi: 10.1073/pnas.2104640118. URL `https://www.pnas.org/content/118/35/e2104640118`.

[77] Sudipa Sinha, O. Misra, and Joydip Dhar. Stability analysis of sir model with vaccination. *American Journal of Computational and Applied Mathematics*, 4:17–23, 01 2014.

[78] Strategic Advisory Group of Experts (SAGE) Working Group on COVID-19 vaccine. Evidence assessment: Coronavac COVID-19 vaccine, 2021. URL `https://cdn.who.int/media/docs/default-source/immunization/sage/2021/april/5_sage29apr2021_critical-evidence_sinovac.pdf`.

[79] Chengjun Sun and Ying-Hen Hsieh. Global analysis of an seir model with varying population size and vaccination. *Applied Mathematical Modelling*, 34(10):2685–2697, 2010. ISSN 0307-904X. doi: https://doi.org/10.1016/j.apm.2009.12.005. URL `https://www.sciencedirect.com/science/article/pii/S0307904X09004016`.

[80] Haoxuan Sun, Yumou Qiu, Han Yan, Yaxuan Huang, Yuru Zhu, Jia Gu, and Song Xi Chen. Tracking reproductivity of COVID-19 epidemic in china with varying coefficient sir model. *Journal of Data Science*, 18(3):455–472, 2020.

[81] The New York Times. Bolsonaro talked vaccines down. now brazil has too few doses, 2021. [Online] Available `https://www.nytimes.com/2021/01/18/world/americas/brazil-covid-variants-vaccinations.html`.

[82] The Telegraph. Project s: Entire town vaccinated as experts race to understand if covid jabs halt transmission, 2021. [Online] Available `https://www.telegraph.co.uk/global-health/science-and-disease/project-s-entire-town-vaccinated-experts-race-understand-covid/`.

[83] Huaiyu Tian, Yonghong Liu, Yidan Li, Chieh-Hsi Wu, Bin Chen, Moritz U. G. Kraemer, Bingying Li, Jun Cai, Bo Xu, Qiqi Yang, Ben Wang, Peng Yang, Yujun Cui, Yimeng Song, Pai Zheng, Quanyi Wang, Ottar N. Bjornstad, Ruifu Yang, Bryan T. Grenfell, Oliver G. Pybus, and Christopher Dye. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in china. *Science*, 368(6491):638–642, 2020.

[84] P. van den Driessche and James Watmough. Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180(1):29–48, 2002.

[85] Merryn Voysey, Sue Ann Costa Clemens, Shabir A. Madhi, Lily Y. Weckx, Pedro M. Folegatti, Parvinder K. Aley, and et al. Safety and efficacy of the chadox1 ncov-19 vaccine (azd1222) against sars-cov-2: an interim analysis of four randomised controlled trials in brazil, south africa, and the uk. *The Lancet*, 397(10269):99–111, 2021.

[86] Lili Wang, Yiwang Zhou, Jie He, Bin Zhu, Fei Wang, Lu Tang, Michael Kleinsasser, Daniel Barker, Marisa C. Eisenberg, and Peter X.K. Song. Rejoinder: An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in china. *Journal of Data Science*, 18(3):446–454, 2020.

[87] Ning Wang, Yuting Fu, Hu Zhang, and Huipeng Shi. An evaluation of mathematical models for the outbreak of COVID-19. *Precision Clinical Medicine*, 3(2):85–93, 2020.

[88] Z. Wang, X. Zhang, G. Teichert, M. Carrasco-Teja, and K. Garikipati. System inference for the spatio-temporal evolution of infectious diseases: Michigan in the time of COVID-19. *Computational Mechanics*, 1:1–24, 2020.

[89] Jia Wangping, Han Ke, Song Yang, Cao Wenzhe, Wang Shengshu, Yang Shanshan, Wang Jianwei, Kou Fuyin, Tai Penggang, Li Jing, Liu Miao, and He Yao. Extended sir prediction of the epidemics trend of COVID-19 in italy and compared with hunan, china. *medRxiv*, 2020. doi: 10.1101/2020.03.18.20038570. URL `https://www.medrxiv.org/content/10.1101/2020.03.18.20038570`.

[90] Wikipedia. São Paulo (state), 2020. [Online] Available `https://en.wikipedia.org/wiki/S%C3%A3o_Paulo_(state)`.

[91] World Health Organization. WHO coronavirus disease (COVID-19) dashboard, 2020 (accessed on 2020-11-08). [Online] Available `https://covid19.who.int/region/amro/country/br`.

[92] Worldometers. COVID-19 coronavirus pandemic, 2020. [Online] Available `https://www.worldometers.info/coronavirus`.

[93] Peng Wu, Xinxin Hao, Eric H Y Lau, Jessica Y Wong, Kathy S M Leung, Joseph T Wu, Benjamin J Cowling, and Gabriel M Leung. Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in wuhan, china. *Eurosurveillance*, 25(3), 2020.

[94] Hou-Cheng Yang, Yishu Xue, Yuqing Pan, Qingyang Liu, and Guanyu Hu. Time fused coefficient sir model with application to COVID-19 epidemic in the united states, 2020.

Appendix

A

SP COVID-19 Info Tracker

Aiming at facing the pandemic scenario in São Paulo state, Wallace Casaca and Marilaine Colnago designed an interactive platform for real-time monitoring and predictive analysis of COVID-19 data. The tracking system, called *SP COVID-19 Info Tracker* [1], makes available to civil society, press agencies, government policymakers and scientific community accurate information and detailed reports about the daily progress of coronavirus in more than 90 cities spread in the state. Figure A.1 illustrates the released platform.

The data is collected at municipal level, i.e., it is daily taken from the epidemiological bulletins as provided by the 92 city halls monitored by the project [59], which are the primary sources of case notifications in Brazil [31]. In terms of representativeness, these cities together comprise "an universe" of 35 million people, i.e., the same as Poland's population, or twice as large as the Netherlands' residents. Another important aspect to be noted is that the São Paulo state is the most populous in Brazil, and it has been the epicenter of the pandemic in the country. Until the end of November 2020, the state corresponded to more than 20% of total confirmed cases in Brazil [53].

The motivation to design a new data repository from first-hand sources comes from the necessity of delivering rapid responses against COVID-19, providing not only more accurate records of new cases and deaths, but also the hospitalizations, suspected cases, testing levels, social isolation rates, deaths under investigation, among other pandemic-related indicators that have not been made available by both state and federal bodies. Moreover, the use of "fresh" data as promptly published by municipal sources allows us to anticipate the virus spread estimation in the state, thus mitigating the Brazilian government delay in updating their reports, as the notifications can take several days or even weeks to be inserted into the central repository, as reported by the Brazilian media [31, 68, 28]. Once the data is collected, it is made available day after day on the platform, including holidays and weekends. This process has been carried out since March 20, 2020 in an effort to keep the data as accurate as possible.

Similar to the efforts made by scientific community to overcome the COVID-19 in other social contexts of technology use, such as collaborative learning platforms [22] and real-time social-distancing detection systems [49],*Info Tracker* focus on promoting digital inclusion through the interactive platform, bringing up coronavirus-related issues like data transparency and epidemiological statistics to those interested in understanding the pandemic's course in São Paulo and, consequently, in Brazil. The *Info Tracker* platform had over 120k visits in 6 months (since June 2020). Finally, the online system has also

[1] **Platform website (in Portuguese):** http://www.spcovid.net.br

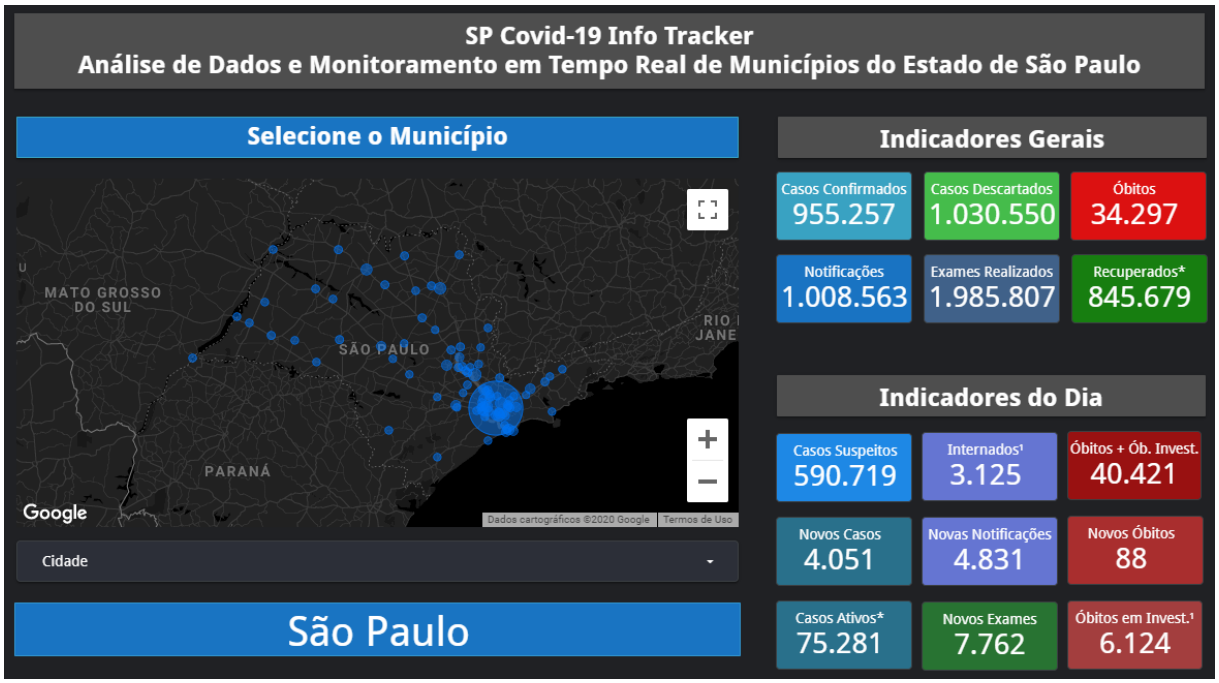


Figure A.1: SP COVID-19 Info Tracker Platform (http://www.spcovid.net.br): First page view.

been used by media agencies to audit municipal governments and other data transparency issues [2].

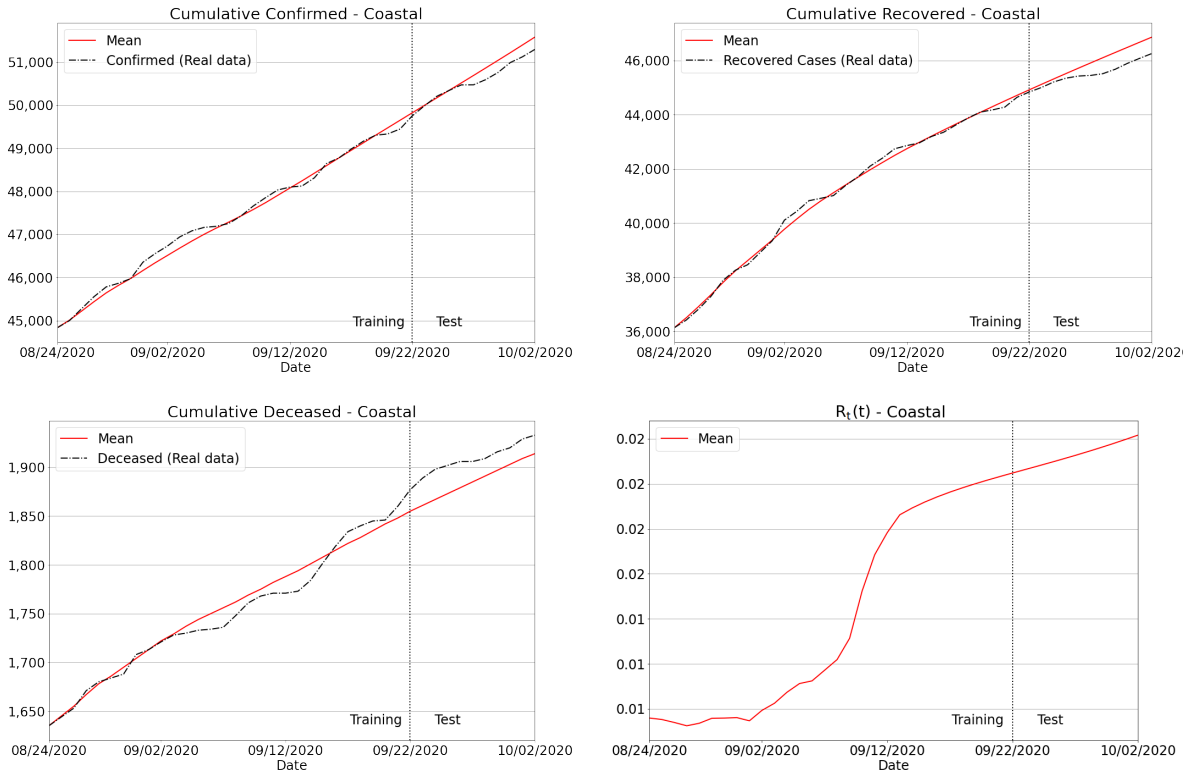
[2] **List of news: http://www.spcovid.net.br/notícias**

Appendix

B

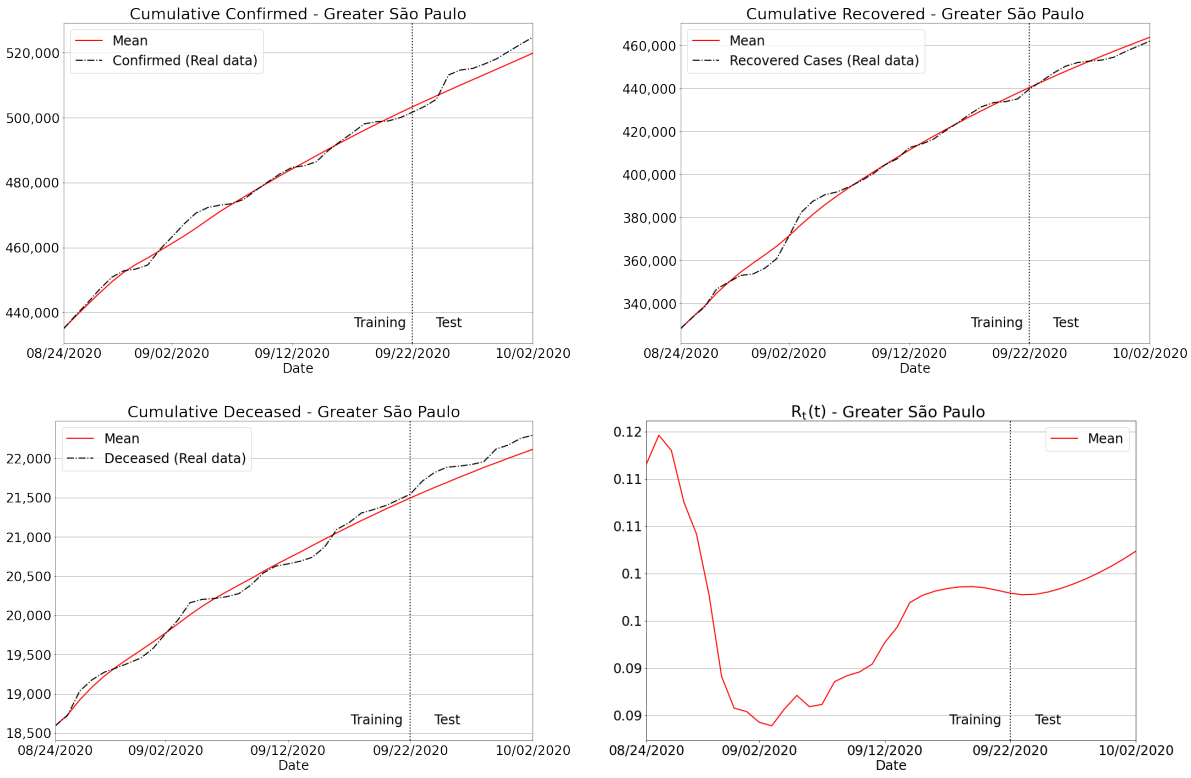
Qualitative results for São Paulo state and Brazilian regions

As previously observed from the results of São Paulo state (see Figure 4.8), the predictions can be considered appropriate for Coastal, Greater São Paulo, Interior East and West regions, as depicted in Figure B.1. Even in the more drastic case of deaths as shown in Coastal region, where the reproduction number jumped from 0.5 to 0.9, the method estimates satisfactorily the total number of fatalities. Finally, we present a similar analysis for the five regions of Brazil in Figure B.2. From the plotted curves, it can be seen that no matter what kind of signature the reproduction number has (increasing or decreasing, lower or higher 1.0), the approach correctly captured the true data in almost all the cases.

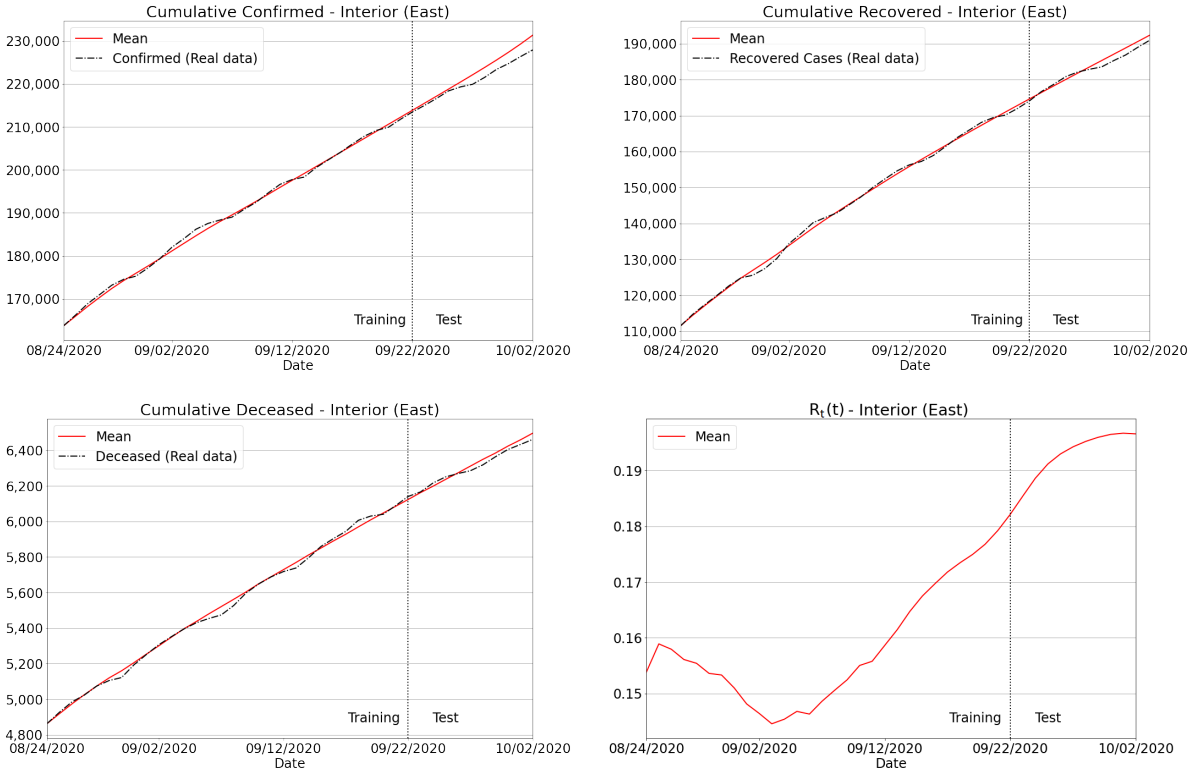


(a) Coastal

Figure B.1: *Cont.*

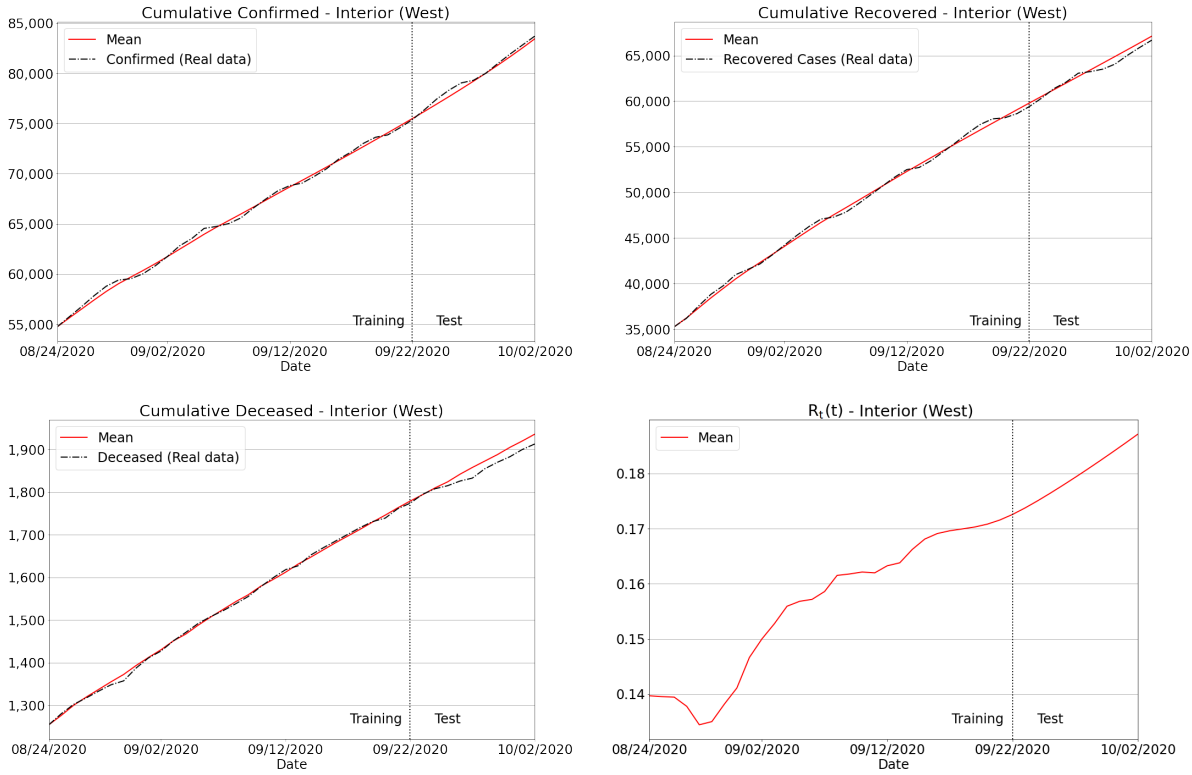


(b) Greater São Paulo



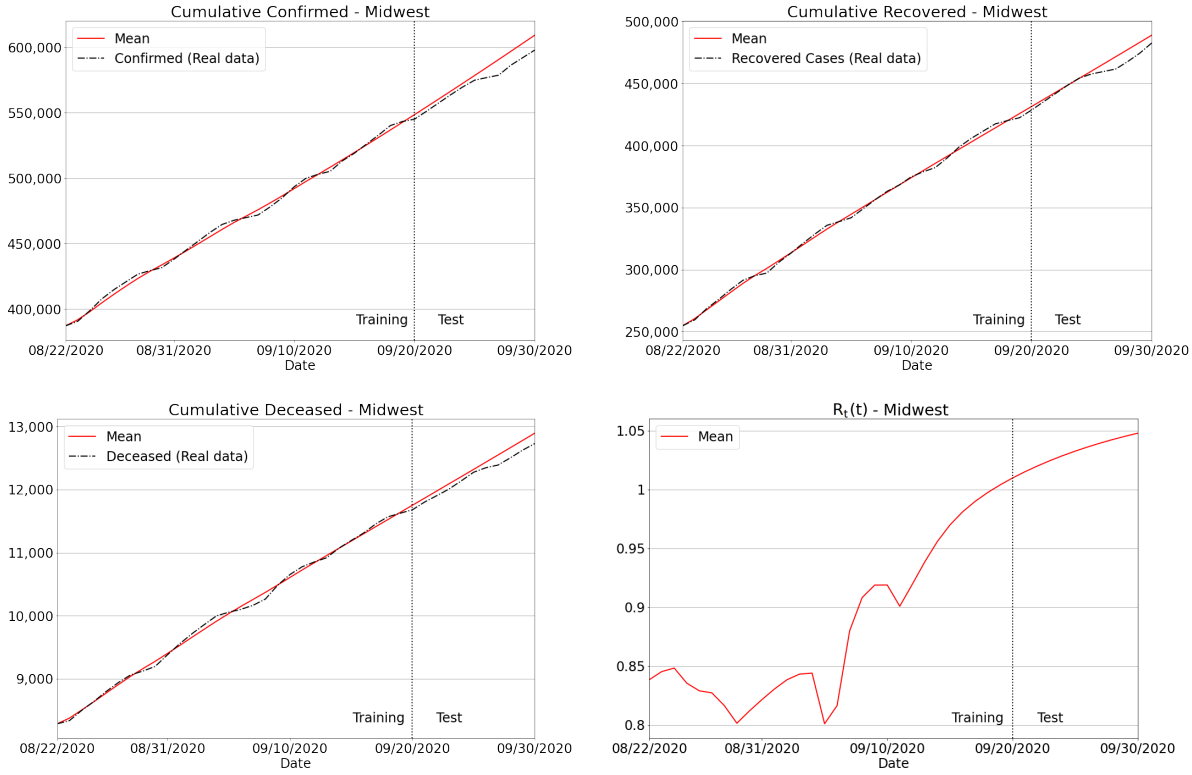
(c) Interior (East)

Figure B.1: *Cont.*



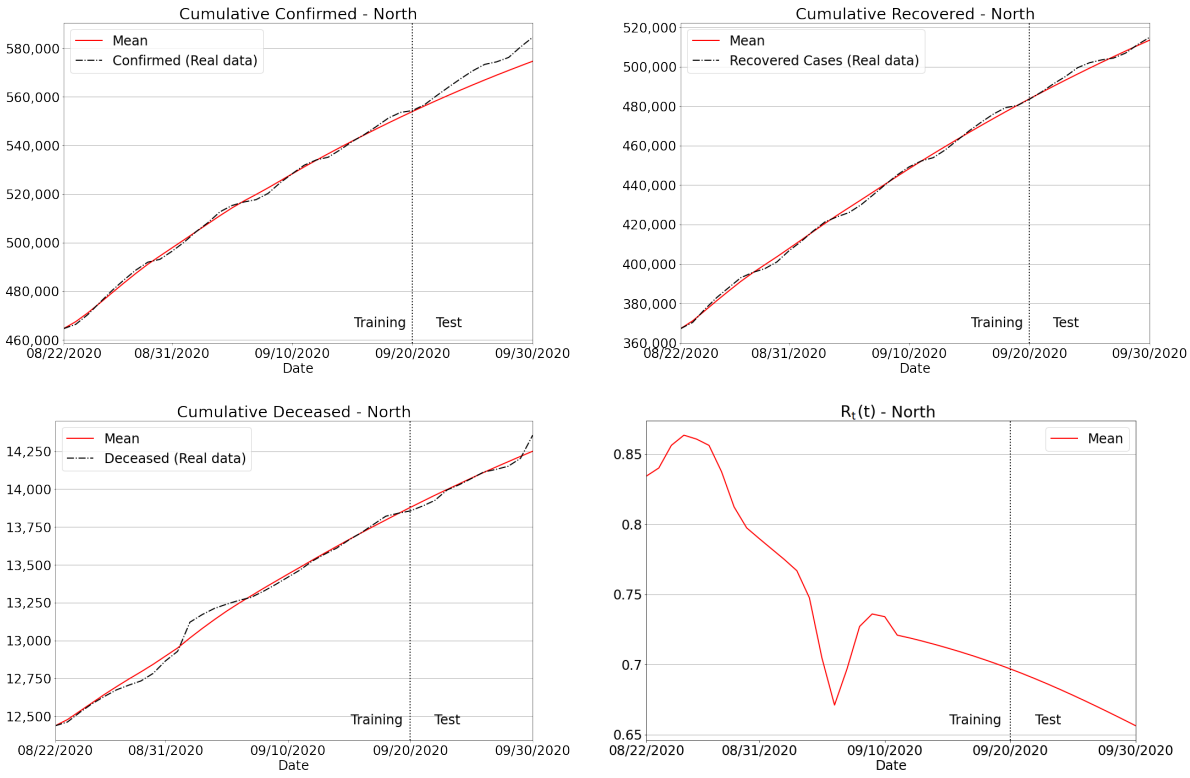
(d) Interior (West)

Figure B.1: Forecasting results covering all the main regions of São Paulo.

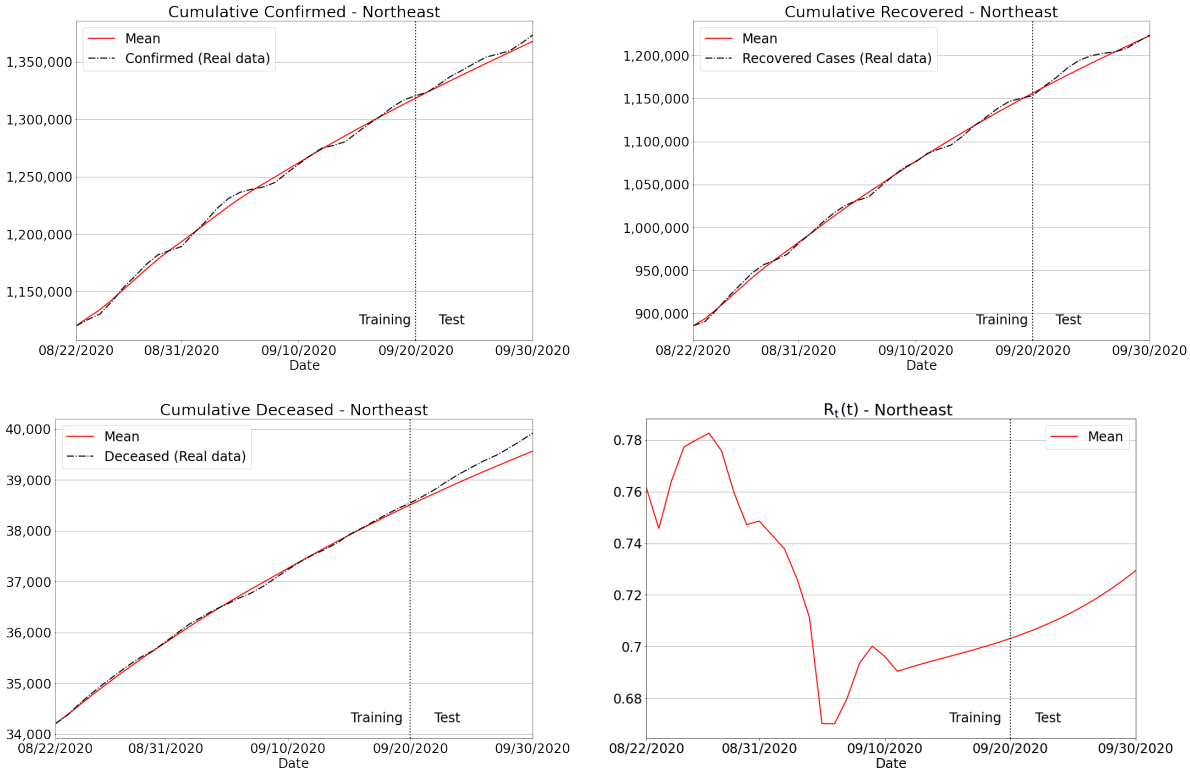


(a) Midwest region

Figure B.2: *Cont.*

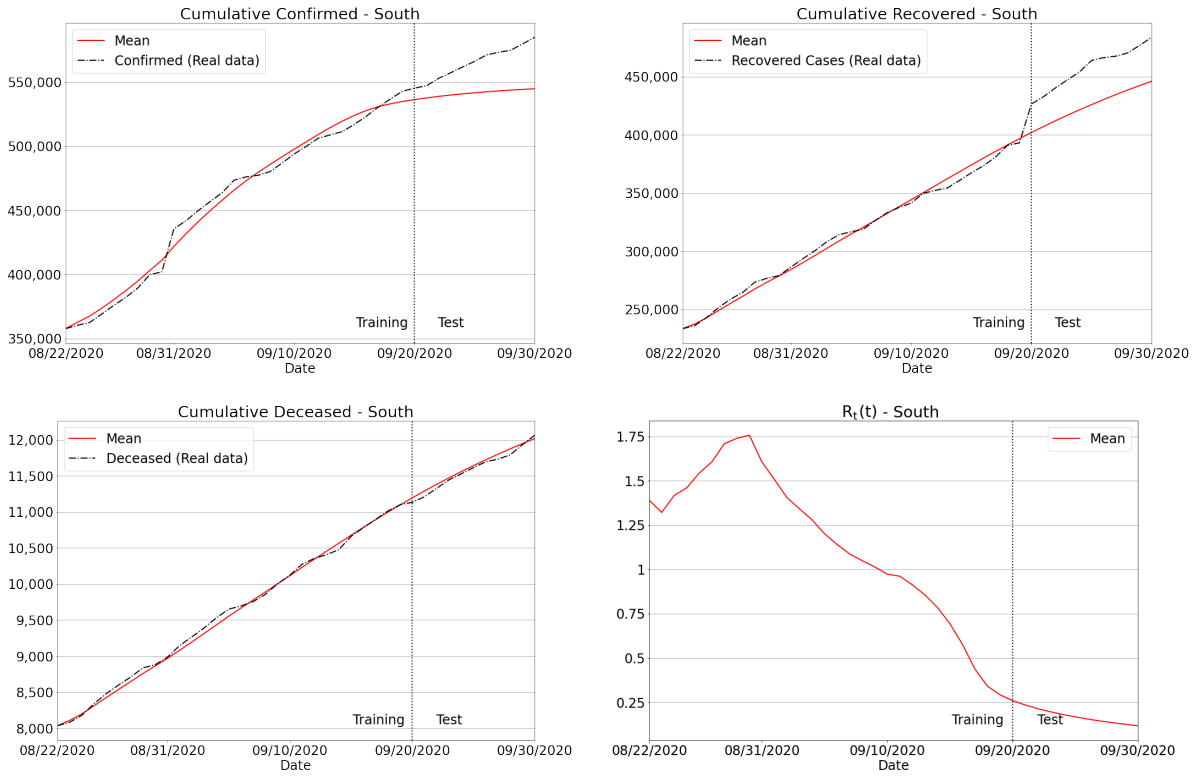


(b) North region

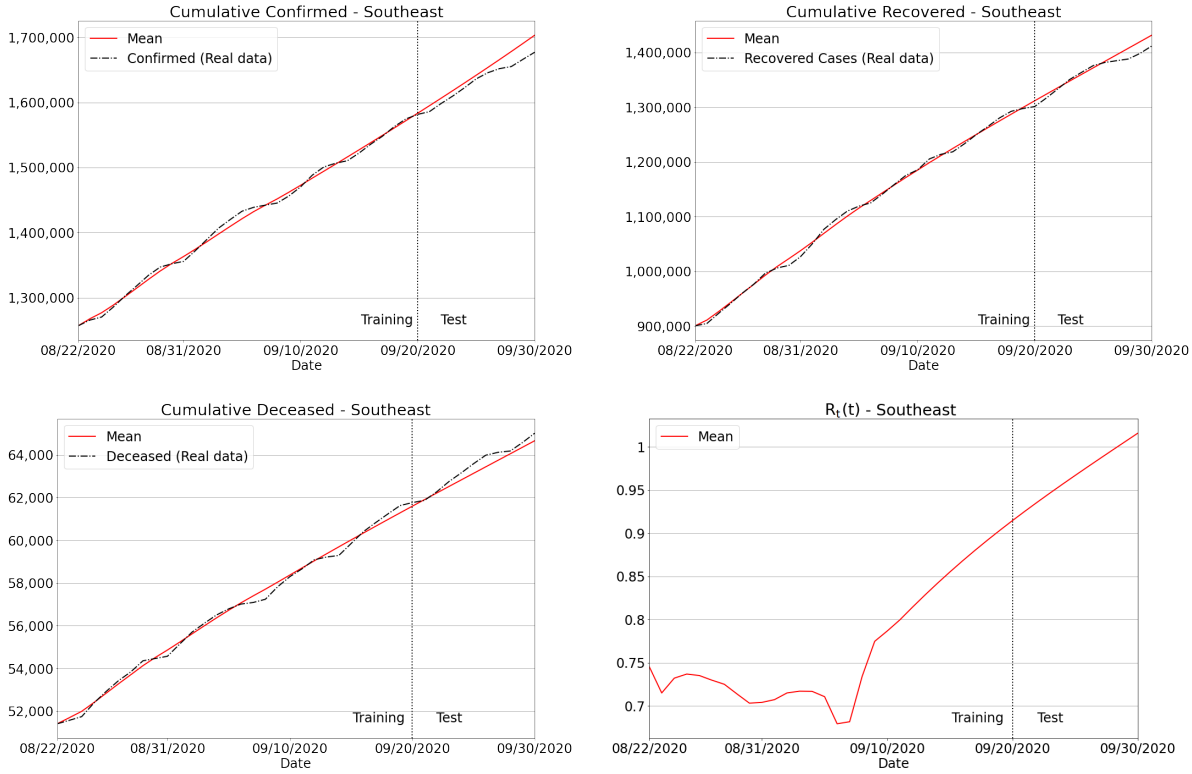


(c) Northeast region

Figure B.2: *Cont.*



(d) South region



(e) Southeast region

Figure B.2: Forecasting results for Brazil's regions.