An approach to COVID-19 incidence forecasting based on mixture of non-central beta models

Paulo Renato Alves Firmino[a,*], Jair Paulino de Sales[b], Paulo S. G. de Mattos Neto[b]

[a] *Center for Science and Technology, Federal University of Cariri, Juazeiro do Norte-CE, Brazil.*
[b] *Centro de Informática, Universidade Federal de Pernambuco, Recife-PE, Brazil.*

Abstract

Keywords: Time series, Forecasting, Pandemic models, COVID-19, Optimisation

1. Introduction

COVID-19 pandemic have harshly impacted the social behaviour around the world since December 2019 [1, 2]. Regarding the global official numbers, until April 27, there have been 148.128.030 confirmed cases of COVID-19, including 3.124.905 deaths [3]. Moreover, the high widespread transmission of SARS-CoV-2 can lead to further mutations that can affect both transmissibility and effectiveness of countermeasures [4]. In this case, although rare, mutations can negatively affect the vaccine effectiveness [5]. This concern has led to the creation of specialised centres for continuous monitoring [4].

Besides mutation concerns, differences in development levels across the nations makes the impact of COVID-19 distinct to each region [6]. Commonly, variables such as population pyramid, scholar levels, sanitation conditions, number hospital beds, and intensive care units (ICUs) diverge among countries (and even within countries) [7, 8]. So, the need to develop systems useful for accompanying, predicting, and controlling the pandemic is clear, considering the particularities of each region. In this sense, time series modelling and forecasting are paramount. These scientific approaches are useful for understanding pandemic patterns over time and supporting strategic, tactic, and operational decision-making processes. However, modelling time series and performing accurate forecasts is not an easy task, mainly when middle- or long-term time windows are taken into account [9, 10]. These exercises require the adequate definition of the phenomenon to be studied and reliable historical data sets. Further, the development of new predictive methods according to the particularities of each problem is usually necessary.

Concerning COVID-19 pandemic, the usual forecasting approaches may not be useful [11], since they lack accurate middle- or long-term forecasts [12]. In turn, regardless of the mortality and transmissibility levels, degree of socioeconomic development, or epoch

*Corresponding author. Tel.:+55(88)9 9330 6093
Email address: paulo.firmino@ufca.edu.br (Paulo Renato Alves Firmino)

in which the pandemic happens in a given territory, it is expected that the shape of the time incidence involves cycles of three consecutive phases: (i) exponential increasing, (ii) plateau, and then (iii) decreasing [13]. Thus, this mixture of behaviours can involve a single or multiple peaks, making a challenge the properly modelling via a single forecasting model [11]. This dynamic cycle might reflect viruses mutation, variations on public health intervention policies, immunisation strategies, technological improvements, and so on (e.g. Spanish flu [14], H1N1 [15], Zika fever [16], Dengue and Chikungunya [16] and COVID-19 [17]). This diversity of factors increases the complexity associated with the COVID-19 forecasts, and their impacts. With this regard, Ioannidis, Cripps, and Tanner [18] present a discussion about the problems of forecasting for COVID-19. According to the authors, poor data input, wrong modelling assumptions, and poor past evidence on effects of available interventions are some of the causes that make models fail.

In literature, in a general way, the modelling strategy used to forecast pandemics can be divided into the following classes: statistical, epidemiological, machine learning, and combination of the aforementioned approaches. Hence, a statistical model can be described as a set of probability distributions considering a sample space [19] [1]. This approach commonly focus on fit a line or curve using a training data set and extrapolating it [12, 20, 13, 21, 22]. Generally, these models are useful to short-term forecasts and does not require much processing time [12]. However, they generally are not useful in long-term forecasting due to their difficulty in predicting second waves, thus, these approach fails to predict when new peaks will occur [12, 23]. In turn, epidemiological (or mechanistic) models aim to mimic the transmission scenario of a particular pathogen based on assumptions about parameters such as transmission rate, immunity, social behaviour, and so on [12, 24, 25, 26]. This class of models assumes that the disease spread becomes faster as more people become infected, decreasing when an increasing group of individuals gains immunity [27]. Although epidemiological models can normally contribute to long-term forecasting, the lack of knowledge about the current pandemic makes that the parameters are not well estimated, negatively impacting the quality of the forecasts [12, 11].

Machine learning models have been a highlighted approach due to their accuracy, flexibility, generalisation ability and generally do not requires prior knowledge of the phenomenon under study [28, 29, 30, 31] [2]. Multilayer Perceptron (MLP) [32, 33], Long Short-Term Memory (LSTM) [34, 35], Fuzzy based models [36, 37], and Support Vector Regression (SVR) [38, 39] are some of the applied methods considering the task of modelling and predicting COVID-19 time series. However, these approaches have low interpretability, making it difficult to understand the problems studied fully. Besides, ML models' training and correct design to deal with real complex problems are challenging due to misspecification, overfitting, and underfitting issues.

Alternatively, it is assumed that policymakers will be able to take more effective actions when they consider projections from multiple models [23]. In a larger context, combination

[1] jps: professores, peço ajuda nesta definição

[2] psgmn: Mais REFS — jps: feito, optei por trazer textos que tratam da aplicação de ML para outros fins além de ST

approaches [40, 41, 42] have been proposed aiming to attain higher accuracy and overcome issues regarding forecast biases [36, 43, 44]. [3] The idea of ensemble models has been applied for both regression [45, 46] and classification problems [47, 48]. Multiple predictor systems are usually built considering two phases: the generation of a pool of individual predictors and the combination based on some logical rule [49]. Thus, if a pool of low accuracy predictors is generated, the ensemble forecast also will.

Despite the numerous and varied approaches to forecasting the COVID-19 (i.e., incidence, deaths, ICU beds), the long-term estimate has been not commonly addressed.In fact, the long-term forecast is a challenging [12], mainly in pandemics time series modelling since the temporal phenomenon is composed of distinct phases as the exponential increasing, plateau, and decreasing [13]. [4] In fact, one of the main challenges in the COVID-19 related time series forecasting is to appropriately estimate the future peaks of the pandemic in long-term. In this sense, the difficulty of elaborating models capable of predicting multiple peaks impacts negatively on the quality of the forecasts generated. [5] [6]

This paper proposes a framework to model and forecast a pandemic time series, in the short- as well as middle-long terms. The proposed approach employs a mixture of non-central beta (MNCB) probability models and can fit pandemic time series with multiple peaks. Thus, a number of increasing-plateau-decreasing cycles of the pandemic time series can be enveloped and predicted. Relevant statistics, such as peak starting and end dates, the total number of infected cases, and the transmission velocity, can be estimated and analysed. Further, the future pandemic trajectory can be estimated by using only the previous univariate pandemic time series of a given territory. [7]

2. Background

An important discussion about the dynamics of the contagious diseases is the basic reproduction number (R_0), an mathematical indicator of transmissibility which refers to the number of secondary cases caused by a single infectious case considering a susceptible population [50]. In an attempt to understating the ongoing pandemic, researched have estimated the r_0 of SARS-CoV-2 varying between 2.2 and 5.7 [50, 51, 52], indicating the need for complementary studies. Further, understating the acquired immunity mechanism to SARS-CoV-2 is crucial to explain the long-term immunity [53, 54]. Regarding COVID-19, Figure 1 sketches the daily incidence in countries around the world until 2021-03-24. The

[3] psgmn: Qual o objetivo de falar de combinacao? Para mim nao ta claro — achei muito interessante o paper [23], tentei fazer uma síntese para justificar o pq da combinação.

[4] jps: apresentamos esta ideia no nosso primeiro artigo sobre covid publicado na chaos, neste caso podemos o citar? Ou seria melhor buscarmos outra referência?

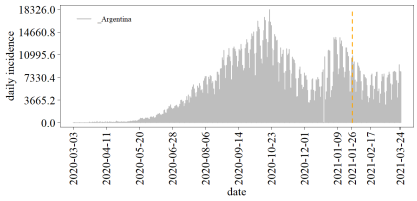
[5] praf: procurar por abordagens sobre o problema dos múltiplos picos

[6] psgmn: Enfatizar que na literatura nao temos modelos que fazem previsao de varios picos

[7] praf: destacar que prever picos de pandemia é dificil, mas que dá suporte a decisão para o 'hoje'. Esse suporte pode ser até mais importante que acertar/errar no longo prazo...

valorizar a ideia de que a serie absorve a dinamica do que ocorre na pandemia (vacinação, isolamento social, lock-down, etc)...

time series are provided by the Johns Hopkins University [55]. One can see that all countries are phasing multiple peaks, with more or less time incidence variations.



(a) Argentina

Images/_Brazil_SingleModels_Series/_Brazil_SingleModels_Series

(b) Brazil

Images/_China_SingleModels_Series/_China_SingleModels_Series

(c) China

Images/_Germany_SingleModels_Series/_Germany_SingleModels_Series

(d) Germany

Images/_India_SingleModels_Series/_India_SingleModels_Series

(e) India

Images/_Iran_SingleModels_Series/_Iran_SingleModels_Series

(f) Iran

Images/_France_SingleModels_Series/_France_SingleModels_Series

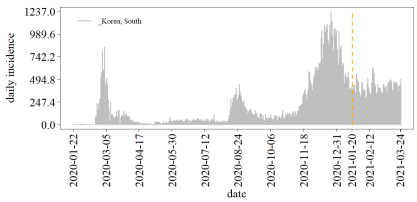
(g) France

Images/_Japan_SingleModels_Series/_Japan_SingleModels_Series

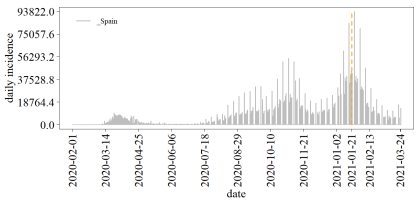
(h) Japan

Images/_Italy_SingleModels_Series/_Italy_SingleModels_Series

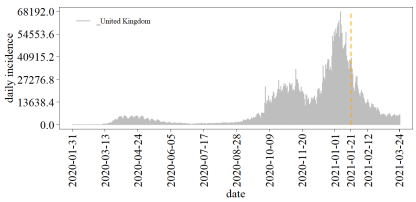
(i) Italy



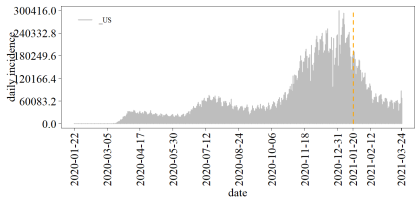
(j) South Korea



(k) Spain



(l) United Kingdom



(m) United States of America

The prediction of the periods in which the disease incidence time series will transit from one phase to the next one, i.e. from Phase (i) to (ii), from (ii) to (iii), and from (iii) to (i), is a hard task. Common time series formalisms, like ARIMA, ETS, support vector regression, and artificial neural networks [56] are not usually able to predict these variations. Further, these approaches suffer when performing relatively middle-long-step-ahead predictions in the light of small sized training data sets. The present paper aims to address these issues by adapting mixtures of probability density functions that can shape and predict multiple cycles of Phases (i), (ii), and (iii).

3. Proposed framework

[8]

Sugestao: The proposal aims to fit multiple cycles of exponential increasing-plateau-decreasing shapes. So, the proposed model (Eq (1)) combines different (n_{NCBs}) NCB probability density functions (PDFs):

$$f(x|\alpha,\beta,\lambda,\omega) = \sum_{i=1}^{n_{NCBs}} \omega_i g(x|\alpha_i,\beta_i,\lambda_i), \tag{1}$$

in which $g(x|\alpha_i,\beta_i,\lambda_i)$ is given by Eq (2), customised with the triple $(\alpha_i,\beta_i,\lambda_i)$, and ω_i is the weight of the i^{th} NCB PDF, determined in such a way that $\sum_{i=1}^{n_{NCBs}} \omega_i = 1$. From Eq (1), one can see that the parameter set of the mixture of NCBs (MNCB) involves $4 \times n_{NCBs}$ coefficients. In fact, α, β, λ, and ω are vectors of size n_{NCBs}, each one.

The single NCB probability density function (PDF), Eq (2), was used successfully to describe one-cycle exponential increasing-plateau-decreasing shape of pandemics evolution [13]:

$$g(x|\alpha,\beta,\lambda) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\beta)} e^{-\frac{\lambda}{2}} \sum_{k=0}^{\infty} \frac{\Gamma(\alpha+\beta+k)(\lambda x)^k}{\Gamma(\alpha+k)2^k k!}, \tag{2}$$

in which $x \in [0,1]$ is an instance of the NCB-distributed random variable taken into account (say X), α (>0) and β (>0) are shape parameters, λ (≥ 0) is the non-centrality parameter, and $\Gamma(y) = \int_0^\infty z^{y-1}e^{-z}dz$ is known as gamma function [57].

The MNCB is proposed to fit the time trajectory of the pandemic incidence, and it is able to model situations involving multiple peaks. In fact, MNCB is a general model since if one assumes $n_{NCBs} = 1$, then it coincides with the NCB predictor introduced by [13].

Figure 2 summarises the three steps of the proposed framewrok: Pre-processing, Modelling, and Forecasting. In the pre-processing step, the available incidence time series (of size N), say $\mathbf{u} = (u_1,\ldots,u_t,\ldots,u_n,\ldots,u_N)$, is firstly partitioned in two sets. The training set involves the first n points, $n < N$. For instance, based on the training time series, one can compute the cumulative pandemic incidence until instant n, say $Cum_n = \sum_{t=1}^{n} u_t$. On

[8] PSGMN: Temos que decidir se eh um approach, model ou framework.
praf: podemos chamar de framework :)

the other hand, the remaining $(N - n)$ points are left for evaluating the performance of the prediction model. Besides n, the analyst must determine the time horizon of the study, say $TH(> N)$, and the maximum number of NCB models, say Mn_{NCBs} $(\geq n_{NCBs})$. The resulting MNCB model will then forecast the incidence time series from instant 1 to instant TH. Based on TH, the time indexes are normalised in order to allow the use of the MNCB PDF. Let the normalised time indexes set be given by $\mathbf{x} = (x_1, \ldots, x_t, \ldots, x_n, \ldots, x_{TH})$, in which

$$x_t = \frac{t-1}{TH-1},\ t = 1,\ \ldots,\ TH. \tag{3}$$

Thus, $x_i \in [0, 1]$.

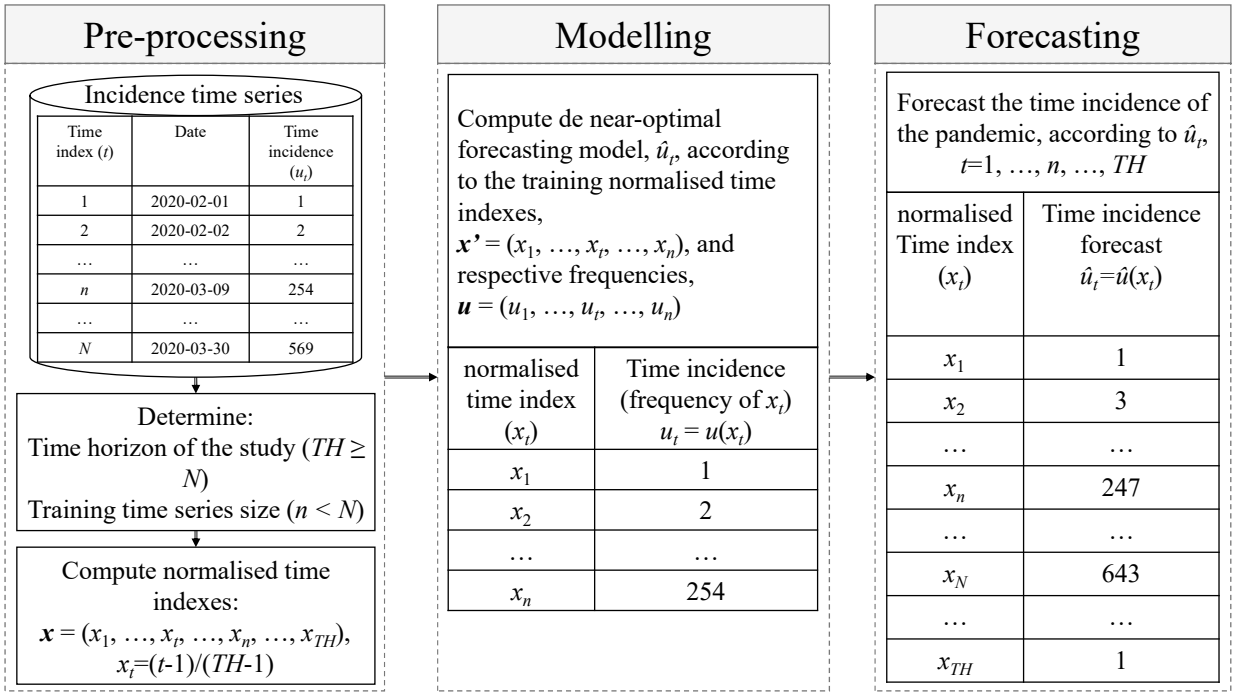


Figure 2: The proposed framework, based on [13]. In the pre-processing step the time indexes (t) are normalised. It allows one to fit the corresponding incidence time series (u_t) via a NCP-based model, say $\hat{u}_t$, in the modelling step. In the forecasting step, the model $\hat{u}_t$ is used to predict the time series of the pandemic incidence through the time horizon determined by the part of the analyst.

In turn, in the Modelling step, the training set is used to compute the near-optimal MNCB model, $\hat{u}_t$. Here, each observed time series value u_t (with $t = 1, \ldots, n$) is approached

by

$$\hat{u}_t = \hat{u}(x_t) = Int\left[f(x_t|\hat{\alpha}_{obs}, \hat{\beta}_{obs}, \hat{\lambda}_{obs}, \hat{\omega}_{obs}) \times \Delta \times T\hat{I}P_{obs}\right], \tag{4}$$

in which $\hat{\theta}_{obs}$ represents the estimates of the parameter set θ in the light of the training set, $\Delta = x_2 - x_1$ reflects the length of the interval involving each MNCB PDF evaluation, and

$$T\hat{I}P_{obs} = Int\left[\frac{Cum_n}{F_X(x_n|\hat{\alpha}_{obs}, \hat{\beta}_{obs}, \hat{\lambda}_{obs}, \hat{\omega}_{obs})}\right] \tag{5}$$

is the estimate of the total number of confirmed infected cases in the territory during TH, with $F_X(x_n|\hat{\alpha}_{obs}, \hat{\beta}_{obs}, \hat{\lambda}_{obs}, \hat{\omega}_{obs}) = \sum_{t=1}^{n} f(x_t|\hat{\alpha}_{obs}, \hat{\beta}_{obs}, \hat{\lambda}_{obs}, \hat{\omega}_{obs}) \times \Delta$. In this way, $F_X(x_n|\hat{\alpha}_{obs}, \hat{\beta}_{obs}, \hat{\lambda}_{obs}, \hat{\omega}_{obs})$ is the estimate of the cumulative probability $P(X \leq x_n)$. From [13], Eq. (5) is based on the idea that X reflects the normalised time until contamination. Thus, if Cum_n involves the proportion $100 \times F_X(x_n|\cdot)\%$, then $\frac{Cum_n}{F_X(x_n|\cdot)}$ will involve 100%. Finally, $Int[z]$ rounds z to its nearest integer number.

Following [13], Eq (4) infers the expected value of the pandemic incidence in instant t, considering that the total number of cases, $T\hat{I}P_{obs}$, occurs until TH. Therefore, once one fixes $n(< N)$, $TH(> N)$, and Mn_{NCBs}, an optimisation method must be adopted in order to achieve the best estimates of the MNCB-based predictor parameters $(\alpha, \beta, \lambda, \omega)$, $(\hat{\alpha}_{obs}, \hat{\beta}_{obs}, \hat{\lambda}_{obs}, \hat{\omega}_{obs})$. The mathematical optimisation problem in this way has named the mean square weighted error (MSWE) as fitness function to be minimised:

$$MSWE(\hat{\alpha}_{obs}, \hat{\beta}_{obs}, \hat{\lambda}_{obs}, \hat{\omega}_{obs}) = \frac{\sum_{t=1}^{n}\left[x_t(\hat{u}_t - u_t)\right]^2}{n}. \tag{6}$$

The MSWE[9] coincides with the well-known mean square error (MSE) if $x_t = 1$, $t = 1, \ldots, n$. Thus, in essence, the MSE considers that every residuals are equally important. In turn, one can see that the MSWE weights each residual $(\hat{u}_t - u_t)$ by the respective normalised time (x_t). Thus, the measure penalises more recent errors of the model than previous ones. This rule seems attractive for better forecasting future pandemics cycles via MNCB. In the present work, the probabilistic optimisation method named generalised simulated annealing (GenSA) [58] is taken into account.

4. COVID-19 Experimental Evaluation

Explicar para o leitor o que vai ser encontrado nas proximas subsecoes

[9]praf: Salgado, vc já viu essa abordagem alguma vez? Jair, vc pode fazer uma varredura na internet buscando algo semelhante? — jps: não encontrei abordagens similares, professores.

4.1. Data Set

The MNCB-based framework has been considered to model, forecast, and compare COVID-19 daily time series incidence from thirteen countries (Argentina, Brazil, China, Germany, India, Iran, Italy, Japan, France, South Korea, Spain, United Kingdom, and US). The time series have been maintained and daily updated by Johns Hopkins university collaborators [55][10].

Country	a	b	c	d	e
Argentina	2,879,677	62,087	6,693,438	2020-03-03	2021-01-26
Brazil	14,369,423	391,936	35,525,209	2020-02-26	2021-01-25
China	103,529	4,857	211,220,000	2020-01-22	2021-01-16
France	5,417,903	102,575	17,102,592	2020-01-24	2021-01-12
Germany	3,310,301	81,968	21,997,335	2020-01-27	2021-01-20
India	17,997,267	201,187	130,027,370	2020-01-30	2021-01-21
Iran	2,438,193	70,532	835,586	2020-02-19	2021-01-24
Italy	3,971,114	119,539	15,166,198	2020-01-31	2021-01-20
Japan	575,563	10,055	2,517,045	2020-01-22	2021-01-16
South Korea	120,673	1,821	2,365,177	2020-01-22	2021-01-16
Spain	3,488,469	77,738	12,865,032	2020-02-01	2021-01-19
United Kingdom	4,406,950	127,434	43,084,487	2020-01-31	2021-01-21
United States of America	31,783,375	567,327	223,293,713	2020-01-22	2021-01-16

Table 1: Confirmed cases (a), deaths (b), vaccine doses administered (c) until 27 April 2021 [3], start date of the training series (d), and start date of the test series by country.

Table 1 shows the information of each time series considered in this work and the scenario of each country studied. Each time series was divided into training and test sets. It was assumed $TH = 1000$, allowing one to predict the daily incidence during 1000 days since the first infection. The size of the training set is

$$n = Int(N \times 0.85), \tag{7}$$

in which the time series size (N) is country-dependent.

4.2. Experimental Setup

Colocar onde: The values of $g(\cdot)$ can be easily computed via statistical software like `dbeta` function of `R` [59]. The proposed framework (MNCB) is compared to NCB, epidemic models [60] (SIR, SEIR, SIS, SIR.sin), and established time series formalisms (ARIMA [61]

[10] Pq essas series foram selecionadas?
praf: não tem uma razão específica... mas elas varrem bem os continentes e trazem alguns dos países que trazem maiores preocupações... talvez seja interessane pormos um texto nesse sentido... Vc pode fazer isso, Jair? jps: Feito.

and ETS [62]). The experimental simulation is divided into two parts: training and test. For all models, the goodness of fit is evaluated according to a set of performance metrics in the training and test phases (see Section 4.3). The proposed and literature models are then considered for comparing the level of difficulty imposed by COVID-19 to the countries.

Table 2 summarises the tuning parameters adopted for achieving the near-optimal MNCB, NCB, SEIR, SIS, SIR, and SIR.sin forecasting models. It must be highlighted that the framework introduced in Section 3 has been adapted (from Eq (1) to Eq (5)) to SEIR, SIS, SIR, and SIR.sin equations. Thus, the MNCB models were adjusted via least squared estimation method, by minimising Eq (6). In turn, NCB, SEIR, SIR, SIS, and SIR.sin approaches were based on the MSE minimisation. With respect to the search spaces of the models parameters sets, the MNCB has considered $Mn_{NCBs} = 50$ and $(n_{NCBs} \in [1, Mn_{NCBs}], \hat{\alpha}_{obs} \in [1, 100]^{Mn_{NCBs}}, \hat{\beta}_{obs} \in [1, 1000]^{Mn_{NCBs}}, \hat{\lambda}_{obs} \in [0, 100]^{Mn_{NCBs}}, \hat{\omega}_{obs} \in [0, 1]^{Mn_{NCBs}})$. The NCB parameters were in the set $(\hat{\alpha}_{obs} \in [1, 100], \hat{\beta}_{obs} \in [1, 1000], \hat{\lambda}_{obs} \in [0, 100])$. In turn, the search space for SIR and SIS parameters was ([1E-10, 100], [1E-06, 100]) for the transmission and removed rates, say $(\hat{\beta}_{obs}, \hat{\gamma}_{obs})$. Besides $\hat{\beta}_{obs}$ and $\hat{\gamma}_{obs}$, SEIR has also involved the per capita death rate and transition rate from exposed to infectious: ($\hat{\mu}_{obs} \in$ [1E-10, 100], $\hat{\sigma}_{obs} \in$ [1E-10, 100]). Further, the SIR model with sinusoidal forcing of the transmission rate (SIR.sin) has considered the following intervals for the death rate (μ), the mean transmission rate (β_0), the amplitude of sinusoidal forcing (β_1), the frequency of oscillations (ω), and the recovery rate (γ): ($\mu \in$ [1E-10, 50], $\beta_0 \in$ [1E-10, 50], $\beta_1 \in$ [1E-10, 50], $\omega \in$ [1E-10, 50], $\gamma \in$ [1E-10, 50])

This optimisation phase has been implemented according to the `GenSA` package [63] of `R`. Regarding SEIR, SIR, SIR.sin, and SIS, the `EpiDynamics` package of `R` [64] has been used. Table 2 summarises the tuning parameters of the optimisation algorithm. It was considered that the maximum number of calls of each MSE-based fitness function was (GSA.max.call=) 2E+06, the maximum running time was (GSA.max.time=) 600 seconds, the maximum number of iterations of the algorithm was (GSA.max.it=) 1E+06, the initial value for temperature was (GSA.temperature=) 1E+06, and the algorithm would stop when there were no improvement after (GSA.nb.stop.improvement=) 20 steps.

Regarding ARIMA and ETS, the `forecast` package of `R` [65] was considered. The respective `auto.arima` and `ets` functions have also promoted near-optimal ARIMA and ETS models. The maximum number of models considered in the stepwise search was (nmodels=) 5E+03. The computer used to execute the modelling and forecasting exercises is a notebook with Windows 10 Home (64 bits) operational system, Intel i7 processor with 2.6GHz, and 8GB RAM memory.

characteristic	value
data percentage for training set	0.85
TH	1000
GSA.max.call	2E+06
GSA.max.time	600
GSA.max.it	1E+06
GSA.temperature	1E+06
GSA.nb.stop.improvement	20
nmodels	5E+03

Table 2: Tuning parameters of the COVID-19 daily incidence models for each country taken into account (Argentina, Brazil, China, Germany, India, Iran, Italy, Japan, France, South Korea, Spain, United Kingdom, and US).

4.3. Performance measures

The following metrics are considered for evaluating the quality of the predictors: Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE); Average Relative Variance (ARV); Index of Disagreement (ID); Theil's U (Theil); Wrong Prediction on Change of Direction (WPOCID); Intercept of the linear fit between $\hat{u}_{t,i}$ and u_t (Reg_Intercept); Slope Coefficient of the linear fit between $\hat{u}_t$ and u_t (Reg_Slope); Indeterminacy Coefficient of the linear fit between $\hat{u}_{t,i}$ and u_t (WR2) [66, 67]; and an Aggregate Performance Metric (APM). See Eqs (9)-(16). The greater the value of a given metric, the worse the model is.

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(u_t - \hat{u}_t)^2}, \tag{8}$$

in which u_t is the incidence of COVID-19 at day t, $\hat{u}_t$ the respective forecast for u_t, and N is the number of observations of the incidence time series taken into account.

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{u_t - \hat{u}_t}{u_t}\right|, \tag{9}$$

$$ARV = \frac{\sum_{t=1}^{N}(u_t - \hat{u}_t)^2}{\sum_{t=1}^{N}(\hat{u}_t - \bar{u}_t)^2}, \tag{10}$$

in which $\bar{u}_t$ is the mean value in the interval $[1, t]$.

$$ID = \frac{\sum_{t=1}^{n}(\hat{u}_t - u_t)^2}{\sum_{t=1}^{n}(|\hat{u}_t - \bar{u_t}| + |u_t - \bar{u}_t|)^2}. \tag{11}$$

$$Theil = \frac{\sum_{t=2}^{N}(u_t - \hat{u}_t)^2}{\sum_{t=2}^{N}(u_t - u_{t-1})^2}. \tag{12}$$

$$WPOCID = 1 - \frac{\sum_{t=2}^{N} D_t}{N-1}, \tag{13}$$

$$D_t = \begin{cases} 1, & \text{if} \quad (u_t - u_{t-1})(\hat{u}_t - \hat{u}_{t-1}) \geq 0 \\ 0, & \text{if} \quad (u_t - u_{t-1})(\hat{u}_t - \hat{u}_{t-1}) < 0. \end{cases} \tag{14}$$

In turn, $\mathrm{WR}^2 = 1 - R^2$, as well as Reg_Intercept and Reg_Slope are related to the linear model adjusted to the pairs u_t e $\hat{u}_t$, via minimal squared estimation. In this way, one can consider the general equation $u_t = Reg\_Intercept + Reg\_Slope \times \hat{u}_t$ [67]. In turn, R^2, the determination coefficient, reflects the performance of the model in capturing the variability of the time series [67]. R^2 is defined as

$$R^2 = \frac{\sum_{t=1}^{N}(u_t - \hat{u}_t)^2}{\sum_{t=1}^{N}(u_t - \bar{u})^2}, \tag{15}$$

in which $\bar{u}$ is the average of the observed series.

For general analyses of these metrics, it is considered the aggregate performance metric:

$$APM = \frac{\sum_{i=1}^{m}(n.Metric_i)}{m}, \tag{16}$$

in which $n.Metric_i$ is the $metric_i$, from the aforementioned ones, normalised according to its values regarding the models taken into account (i.e. MNCB, NCB, SEIR, SIR, SIS, SIR.sin, ARIMA, and ETS) and m is the number of metrics ($m = 9$ in the present paper, reflecting RMSE, MAPE, ARV, ID, Theil'U, WPOCID, $|Reg\_Intercept|$, $||Reg\_Slope|-1|$, and WR2). APM is based on the reasoning that the near to zero the value of APM is, the better the model is. Thus, for APM, $|Reg\_Intercept|$ and $||Reg\_Slope| - 1|$ are adopted instead of $Reg\_Intercept$ and Reg_Slope. Therefore, APM is the simple average of the normalised version of the previous mentioned metrics, according to

$$n.Metric_i = \frac{Metric_i - min_i}{max_i - min_i}, \tag{17}$$

in which min_i and max_i are, respectively, the observed minimal and maximal values of $Metric_i$ among the adjusted models under study.

4.4. Performance results

Tables 3 and 4 summarise the performance of the near-optimal predictors when fitting and forecasting COVID-19 incidence in US, in this order. The second and third columns of the tables highlight the model with the worst and best figures, respectively. One can see that MNCB has always beaten the remaining models during training, but in terms of MAPE and WPOCID, in which ARIMA and SIR.sin have been the best, respectively. During test phase, MNCB has only been overcome in terms of Reg_Slope and Reg_WR2, in which SEIR and SIR.sin have obtained the best results. Thus, under an aggregate point of view (the two last lines of the tables), MNCB has been attractive. Anyway, the expressive values of

RMSE reflect the challenge of fitting and predicting this series. The RMSE is presented in the same measure unit of the time series (i.e. daily pandemic incidence). Thus, during training, the RMSE of MNCB and SIR.sin was approximately 15,832 and 28,923 cases per day, respectively. On the other hand, the MNCB model has been able to capture (R^2 = 1-WR2=) 95.1% of the variability of the US series during training phase. In turn, MNCB has presented a MAPE of 0.353 in the test phase.

Metric	Worst	Best	MNCB	NCB	SIR	SEIR	SIS	SIR.sin	ARIMA	ETS
RMSE	SIR.sin	MNCB	**15,832.070**	27,023.010	26,237.894	26,687.214	26,319.020	28,923.013	16,378.936	16,378.999
MAPE	SEIR	ARIMA	0.592	28.454	144.712	322.618	166.832	0.756	**0.184**	0.190
ARV	SIR.sin	MNCB	**0.052**	0.165	0.149	0.151	0.148	0.229	0.056	0.056
ID	SIR.sin	MNCB	**0.013**	0.040	0.037	0.038	0.037	0.050	0.014	0.014
Theil	SIR.sin	MNCB	**0.702**	2.045	1.927	1.952	1.939	2.342	0.751	0.751
WPOCID	ARIMA, ETS	SIR.sin	0.462	0.459	0.459	0.459	0.459	**0.445**	0.569	0.569
Reg_Intercept	SIR.sin	MNCB	**-107.778**	5,692.467	1,674.086	2,676.145	1,032.923	-8,871.184	1,554.125	1,553.912
Reg_Slope	SIR.sin	MNCB	**1.002**	0.960	0.975	0.960	0.974	1.132	0.996	0.996
Reg_WR2	SIR.sin	MNCB	**0.049**	0.139	0.134	0.137	0.134	0.151	0.052	0.052
n.RMSE	SIR.sin	MNCB	0.000	0.855	0.795	0.829	0.801	1.000	0.042	0.042
n.MAPE	SEIR	ARIMA	0.001	0.088	0.448	1.000	0.517	0.002	0.000	0.000
n.ARV	SIR.sin	MNCB	0.000	0.638	0.549	0.559	0.542	1.000	0.025	0.025
n.ID	SIR.sin	MNCB	0.000	0.723	0.640	0.665	0.640	1.000	0.027	0.027
n.Theil	SIR.sin	MNCB	0.000	0.819	0.747	0.762	0.754	1.000	0.030	0.030
n.WPOCID	ARIMA, ETS	SIR.sin	0.133	0.111	0.111	0.111	0.111	0.000	1.000	1.000
n.Reg_Intercept	SIR.sin	MNCB	0.000	0.637	0.179	0.293	0.106	1.000	0.165	0.165
n.Reg_Slope	SIR.sin	MNCB	0.000	0.291	0.179	0.293	0.185	1.000	0.021	0.021
n.Reg_WR2	SIR.sin	MNCB	0.000	0.878	0.826	0.862	0.833	1.000	0.030	0.030
n.Mean	SIR.sin	MNCB	0.015	0.560	0.497	0.597	0.499	0.778	0.149	0.149
n.Sd	SIR.sin	MNCB	0.044	0.314	0.282	0.304	0.294	0.441	0.323	0.323

Table 3: Performance of the forecasting models (MNCB, NCB, SIR, SEIR, SIS, SIR.sin, ARIMA, ETS) when predicting US time data (Training phase). The best values for each metric are highlighted in bold.

Metric	Worst	Best	MNCB	NCB	SIR	SEIR	SIS	SIR.sin	ARIMA	ETS
RMSE	SIS	MNCB	**32,320.046**	245,461.479	251,736.233	231,123.149	283,187.135	150,253.595	102,052.099	102,007.888
MAPE	SIS	MNCB	**0.353**	3.558	3.649	3.351	4.100	2.180	1.481	1.481
ARV	ETS	MNCB	**0.664**	1.370	1.360	1.403	1.308	1.738	2.414	2.415
ID	SIS	MNCB	**0.193**	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Theil	SIS	MNCB	**5.475**	315.658	331.970	279.828	420.085	118.366	54.592	54.545
WPOCID	ARIMA, ETS	MNCB	**0.476**	0.524	0.524	0.524	0.524	0.524	1.000	1.000
Reg_Intercept	SIR.sin	MNCB	**29,534.892**	332,978.945	341,990.174	397,202.198	288,882.852	704,656.213	86,912.969	86,912.969
Reg_Slope	SIR.sin	SEIR	0.591	-0.771	-0.783	**-1.011**	-0.570	-2.710	NA	NA
Reg_WR2	ARIMA, ETS	SIR.sin	0.226	0.308	0.290	0.275	0.323	**0.195**	1.000	1.000
n.RMSE	SIS	MNCB	0.000	0.850	0.875	0.792	1.000	0.470	0.278	0.278
n.MAPE	SIS	MNCB	0.000	0.855	0.880	0.800	1.000	0.488	0.301	0.301
n.ARV	ETS	MNCB	0.000	0.403	0.397	0.422	0.367	0.613	0.999	1.000
n.ID	SIS	MNCB	0.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999
n.Theil	SIS	MNCB	0.000	0.748	0.787	0.662	1.000	0.272	0.118	0.118
n.WPOCID	ARIMA, ETS	MNCB	0.000	0.091	0.091	0.091	0.091	0.091	1.000	1.000
n.Reg_Intercept	SIR.sin	MNCB	0.000	0.449	0.463	0.545	0.384	1.000	0.085	0.085
n.Reg_Slope	SIR.sin	SEIR	0.234	0.129	0.122	0.000	0.247	1.000	NA	NA
n.Reg_WR2	ARIMA, ETS	SIR.sin	0.039	0.140	0.117	0.099	0.159	0.000	1.000	1.000
n.Mean	ETS	MNCB	0.030	0.518	0.526	0.490	0.583	0.548	0.598	0.598
n.Sd	ETS	MNCB	0.078	0.354	0.367	0.360	0.406	0.389	0.436	0.436

Table 4: Performance of the forecasting models (MNCB, NCB, SIR, SEIR, SIS, SIR.sin, ARIMA, ETS) when predicting US time data (Test phase). The best values for each metric are highlighted in bold.

Tables 5 and 6 allow one to compare the performance of the models in the light of the thirteen countries taken into account, in terms of APM. In general, considering the test sets, MNCB has been the best model whilst it has never been the worst alternative. On the other hand, in the training set, MNCB has been beaten by ARIMA, only.

Country	Worst	Best	MNCB	NCB	SIR	SEIR	SIS	SIR.sin	ARIMA	ETS
Argentina	SIR	ARIMA	0.215	0.52	0.891	0.437	0.685	0.868	0.071	0.44
Brazil	SIS	ARIMA	0.203	0.711	0.602	0.323	0.885	0.394	0.081	0.456
China	SIR.sin	MNCB	0.142	0.353	0.326	0.221	0.534	0.731	0.477	0.469
France	SEIR	MNCB	0.091	0.615	0.45	0.799	0.537	0.651	0.172	0.167
Germany	SEIR	ARIMA	0.333	0.684	0.631	0.758	0.59	0.635	0.036	0.559
India	SIR.sin	MNCB	0.117	0.61	0.653	0.617	0.678	0.804	0.125	0.227
Iran	SEIR	ETS	0.078	0.713	0.657	0.742	0.658	0.733	0.093	0.076
Italy	SEIR	ARIMA	0.111	0.434	0.313	0.921	0.359	0.416	5e-03	0.072
Japan	SEIR	ARIMA	0.197	0.725	0.638	0.751	0.658	0.413	0.032	0.147
Korea, South	SEIR	ETS	0.064	0.311	0.309	0.929	0.301	0.558	0.013	0.011
Spain	SIS	ARIMA	0.269	0.642	0.662	0.704	0.752	0.732	0.043	0.599
United Kingdom	NCB	ETS	0.242	0.861	0.677	0.667	0.708	0.704	0.099	0.096
US	SIR.sin	MNCB	0.015	0.56	0.497	0.597	0.499	0.778	0.149	0.149
APM (rank)	SEIR	ARIMA	0.16 (2)	0.595 (5)	0.562 (4)	0.651 (8)	0.603 (6)	0.647 (7)	0.107 (1)	0.267 (3)

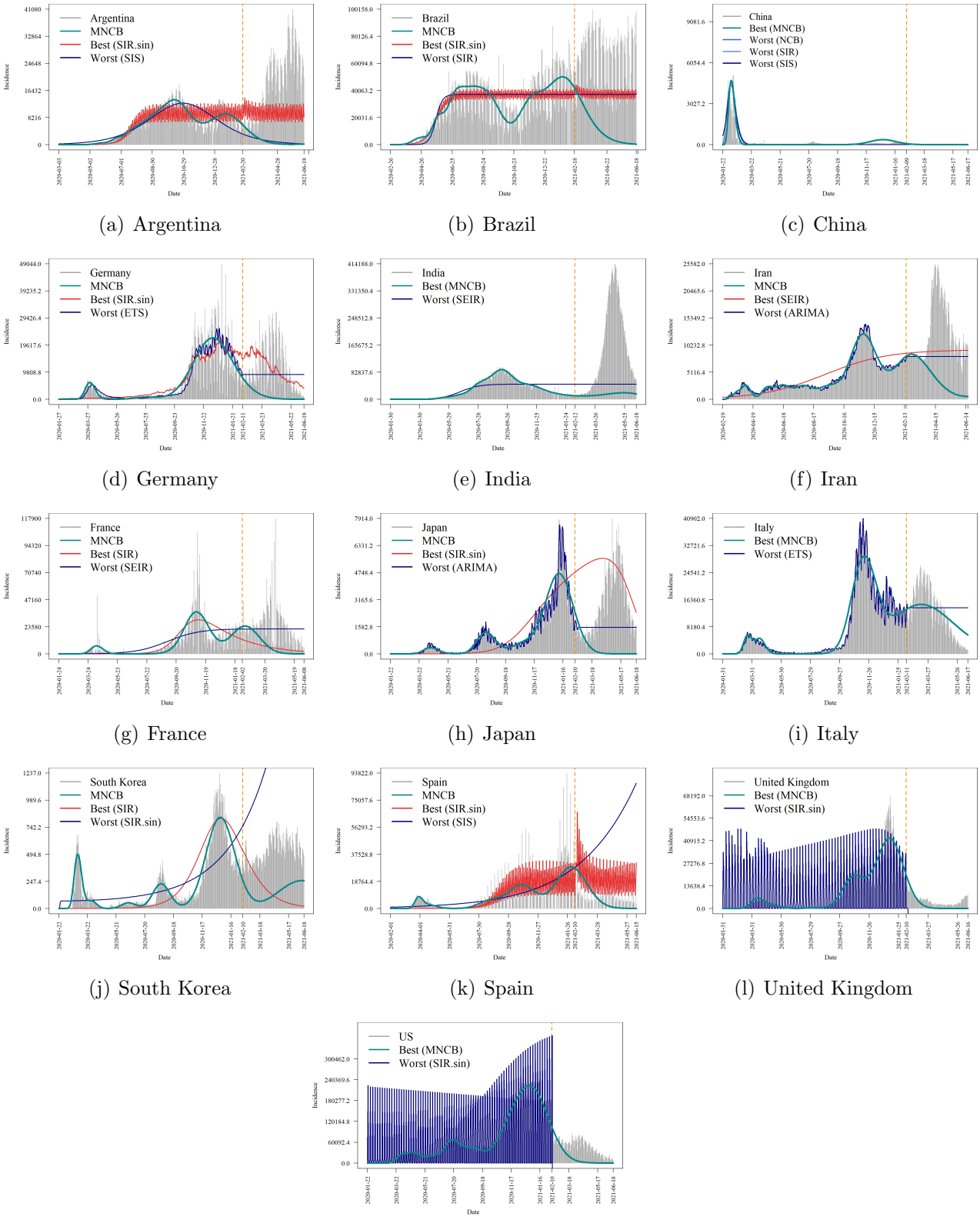
Table 5: Aggregate mean normalised performance (APM) of the forecasting models (MNCB, NCB, SIR, SEIR, SIS, SIR.Sin, ARIMA, ETS) when predicting Argentina, Brazil, China, France, Germany, India, Iran, Italy, Japan, Korea, South, Spain, United Kingdom, US time series (Training phase). The rank of each model is in parentheses, in the last line.

Country	Worst	Best	MNCB	NCB	SIR	SEIR	SIS	SIR.sin	ARIMA	ETS
Argentina	ETS	ARIMA	0.426	0.307	0.461	0.542	0.52	0.422	0.118	0.558
Brazil	SIR	SIS	0.471	0.352	0.657	0.362	0.151	0.424	0.187	0.6
China	ARIMA	SEIR	0.381	0.336	0.336	0.314	0.336	0.582	0.649	0.648
France	ARIMA	SIR.sin	0.306	0.408	0.353	0.326	0.472	0.235	0.682	0.681
Germany	SIR.sin	ARIMA	0.46	0.47	0.495	0.441	0.457	0.592	0.128	0.487
India	ARIMA	MNCB	0.184	0.505	0.506	0.395	0.514	0.394	0.708	0.403
Iran	ARIMA	SEIR	0.215	0.136	0.241	0.075	0.241	0.556	0.648	0.512
Italy	ETS	ARIMA	0.199	0.259	0.225	0.274	0.257	0.485	0.187	0.488
Japan	NCB	SEIR	0.336	0.456	0.429	0.15	0.436	0.378	0.433	0.383
Korea, South	NCB	SEIR	0.329	0.639	0.397	0.313	0.487	0.392	0.517	0.517
Spain	ETS	SIR.sin	0.452	0.509	0.407	0.511	0.448	0.354	0.48	0.621
United Kingdom	SIS	MNCB	0.088	0.599	0.549	0.472	0.655	0.638	0.604	0.604
US	ETS	MNCB	3e-02	0.518	0.526	0.49	0.583	0.548	0.598	0.598
APM (rank)	ETS	MNCB	0.298 (1)	0.423 (3)	0.429 (5)	0.359 (2)	0.427 (4)	0.462 (7)	0.457 (6)	0.546 (8)

Table 6: Aggregate mean normalised performance (APM) of the forecasting models (MNCB, NCB, SIR, SEIR, SIS, SIR.sin, ARIMA, ETS) when predicting Argentina, Brazil, China, France, Germany, India, Iran, Italy, Japan, Korea, South, Spain, United Kingdom, US time series (Test phase). The rank of each model is in parentheses, in the last line.

Figure 3 exhibits the available COVID-19 incidence time series and the respective models forecasts. The vertical orange dashed line separates the training and test data sets. The machine learning has been based on the training set. Then, the predictors were challenged to infer the test series. One can see the difficult of the predictors in fitting the pandemic incidence trajectory though some adherence can be verified. It is argued that any change in the national and local intervention policies might affect the pandemic trajectory, leading to the fluctuations of the target series around the expected values inferred from the models, mainly in the training set. Anyway, the target has been underestimated in the last part of the test set in the the cases of Brazil, France, and India (Figures 3(b), 3(g), and 3(e)). In turn, in the cases of Japan and Spain (Figures 3(h) and 3(k)), the pattern of the test series has been usually overestimated. In fact, regardless of the class of the formalism, the

performance in predicting the transitions between Phases (i), (ii), and (iii) has usually been challenging. As highlighted by [13], ARIMA and ETS might be more useful to perform short-term (e.g. one-step-ahead) than middle-long-term forecasts, thus tending to present small oscillations through the latter. In turn, SIR.sin has not successfully achieved the best cycle variations, though being the most time demanding approach. Finally, MNCB, ARIMA, and ETS have usually achieved attractive results in the training phases. In turn, considering the test set, MNCB have provided good results in a reasonable number of cases. For instance, it was the model that better predicted a transition from Phase (iii) to Phase (i) in India, Iran, France, and Italy. Further, its performance in predicting the transition from Phases (ii) to (iii) in UK and USA seems remarkable.



(a) Argentina (b) Brazil (c) China

(d) Germany (e) India (f) Iran

(g) France (h) Japan (i) Italy

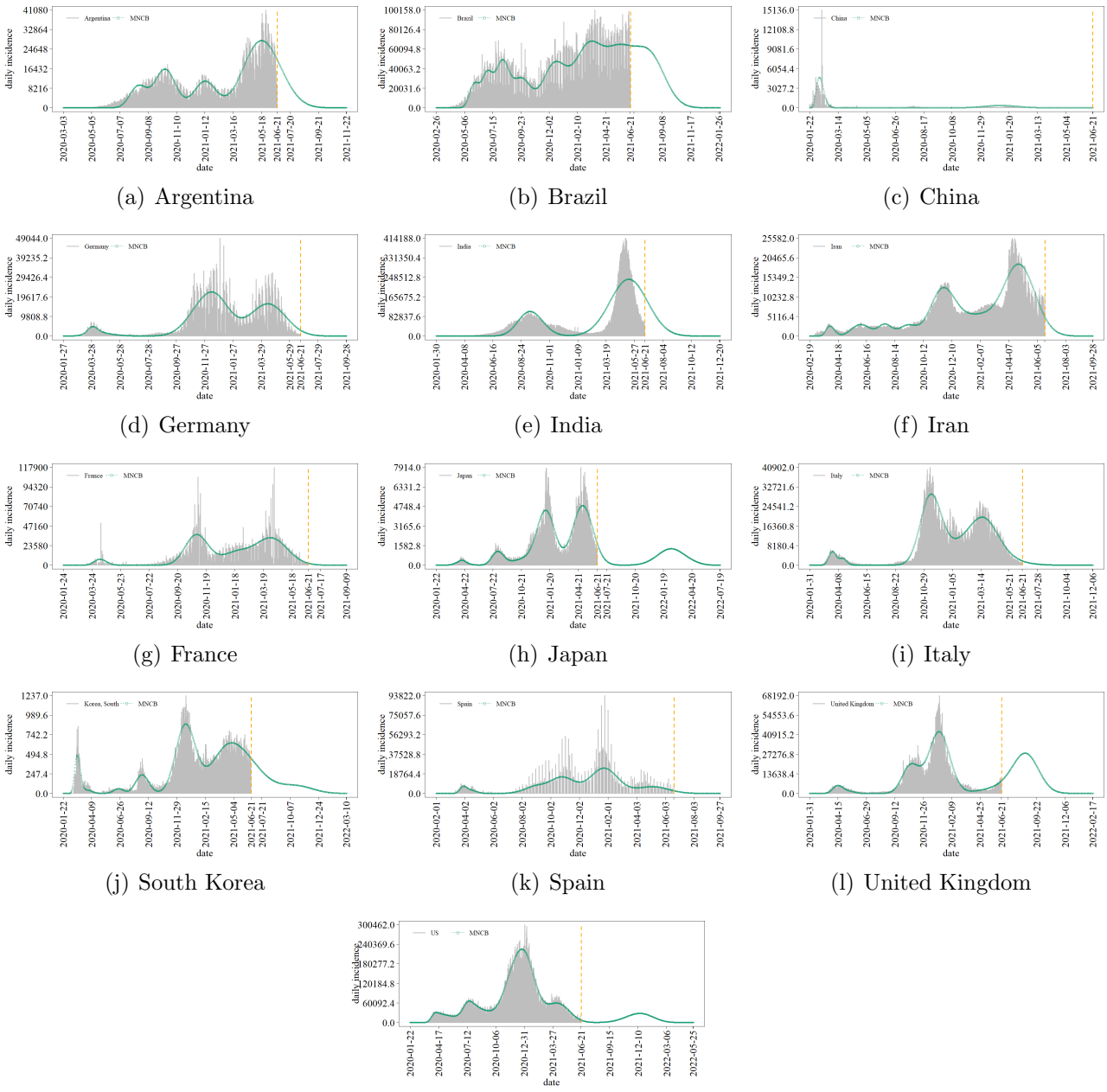
(j) South Korea (k) Spain (l) United Kingdom

(m) United States of America

Figure 3: Prediction of the national daily incidence of COVID-19 since the first register, according to Johns Hopkins University data set [55]. The vertical purple dashed line separates training and test series.

4.5. Comparing national pandemic incidences

Disregarding from the limitations of [13] in dealing with multiple-peaks pandemics time series, the proposed MNCB approach can be useful for better summarising and comparing the difficulty imposed by these diseases among territories. Figure 5 presents the shape of the MNCB models in the face of the national incidence of COVID-19 until 2021-03-28. Now, the MNCB models have been adjusted under a time bound of (GSA.max.time =) 60 seconds, taking the respective previous model of the country as initial solution for GSA algorithm. The way the MNCB models fit the available time series is remarkable. It opens possibilities for better comparative analyses and evaluation among intervention policies of the countries. Most generally, the approach has predicted transitions from Phase (i) to (ii) and (iii) (i.e. Argentina, Germany, India, France, Japan, Italy, South Korea, United Kingdom, and United States of America). On the other hand, the transition from Phase (ii) to (iii) has also been verified (i.e. Brazil, Iran). Finally, new complete epidemics cycles have also been predicted (i.e. South Korea, Spain). This last pattern prediction is particularly attractive once it reflects the fact that in the attempt of better fitting the observed time series (with emphasis to the most recent incidences, according to Eq (6)), new pandemic cycles might be incorporated in the model.



(a) Argentina (b) Brazil (c) China

(d) Germany (e) India (f) Iran

(g) France (h) Japan (i) Italy

(j) South Korea (k) Spain (l) United Kingdom

(m) United States of America

Figure 4: NCB Prediction of the national daily incidence of COVID-19 since the first register, according to Johns Hopkins University data set [55]. The vertical dashed orange line marks the end of the available time series.

Table 7 involves specific figures of the MNCB models. Besides n and the respective training cumulative incidence (Cum_n), the table also exhibits the inferred near-optimal MNCB architecture (n_{NCBs}), the cumulative pandemic incidence through TH days ($\hat{TIP}_{obs}$), and remarkable instants, i.e. the starting ($date_0$), main peak ($date_m$), and finish ($date_{end}$)

dates. For instance, until 2021-03-28, the most contaminated country was US, with ($Cum_n =$) 30,262,377 infected, followed by Brazil. In turn, considering the available data, it is also expected that US stays in the worst position at the end of the pandemic, at ($date_{end} =$) 2022-07-26, involving a total of ($\hat{TIP}_{obs} =$) 38,886,706 cases. In turn, Brazil seems to be in the main peak of the disease spread. Thus, it is reasonable to infer that the corresponding fitted model be more sensible to the ongoing trajectory of the pandemic. From the models, special attention must also be taken with respect to Spain and Korea, South, for which the forecasts to the pandemic finish are 2022-10-27 and 2022-10-14, respectively. The most expensive MNCB models were dedicated to Argentine, US, and Brazil, demanding mixtures of ($n_{NCBs} =$) 38, 19, and 14 NCB PDFS, in this order.

Attribute	Argentina	Brazil	China	France	Germany	India	Iran	Italy	Japan	Korea, South	Spain	United Kingdom	US
n	391	397	432	430	427	424	404	423	432	432	422	423	432
Cum_n	2,308,597	12,534,688	101,676	4,680,709	2,784,652	12,039,644	1,855,674	3,532,205	468,346	102,141	3,340,077	4,347,013	30,262,377
$date_0$	2020-03-03	2020-02-26	2020-01-22	2020-01-24	2020-01-27	2020-01-30	2020-02-19	2020-01-31	2020-01-22	2020-01-22	2020-02-01	2020-01-31	2020-01-22
$\hat{TIP}_{obs}$	3,568,969	15,374,035	103,363	7,315,650	5,406,239	18,227,106	2,417,892	4,672,420	857,364	207,243	4,259,325	7,205,678	38,886,706
n_{NCBs}	38	14	4	4	4	7	9	4	5	12	10	4	19
$date_m$	2020-10-16	2021-03-27	2020-02-10	2021-04-18	2021-05-13	2020-09-13	2020-11-22	2020-11-14	2021-01-05	2020-12-24	2021-01-23	2021-01-06	2020-12-26
$date_{end}$	2022-02-28	2021-08-21	2021-07-16	2021-10-25	2021-11-06	2021-11-26	2021-12-16	2021-09-02	2021-12-04	2022-10-14	2022-10-27	2022-03-19	2022-07-26
Training time	60.25	60	60.2	60.3	60.02	60.03	61.45	60	60.03	60.06	60.01	60.02	60.06

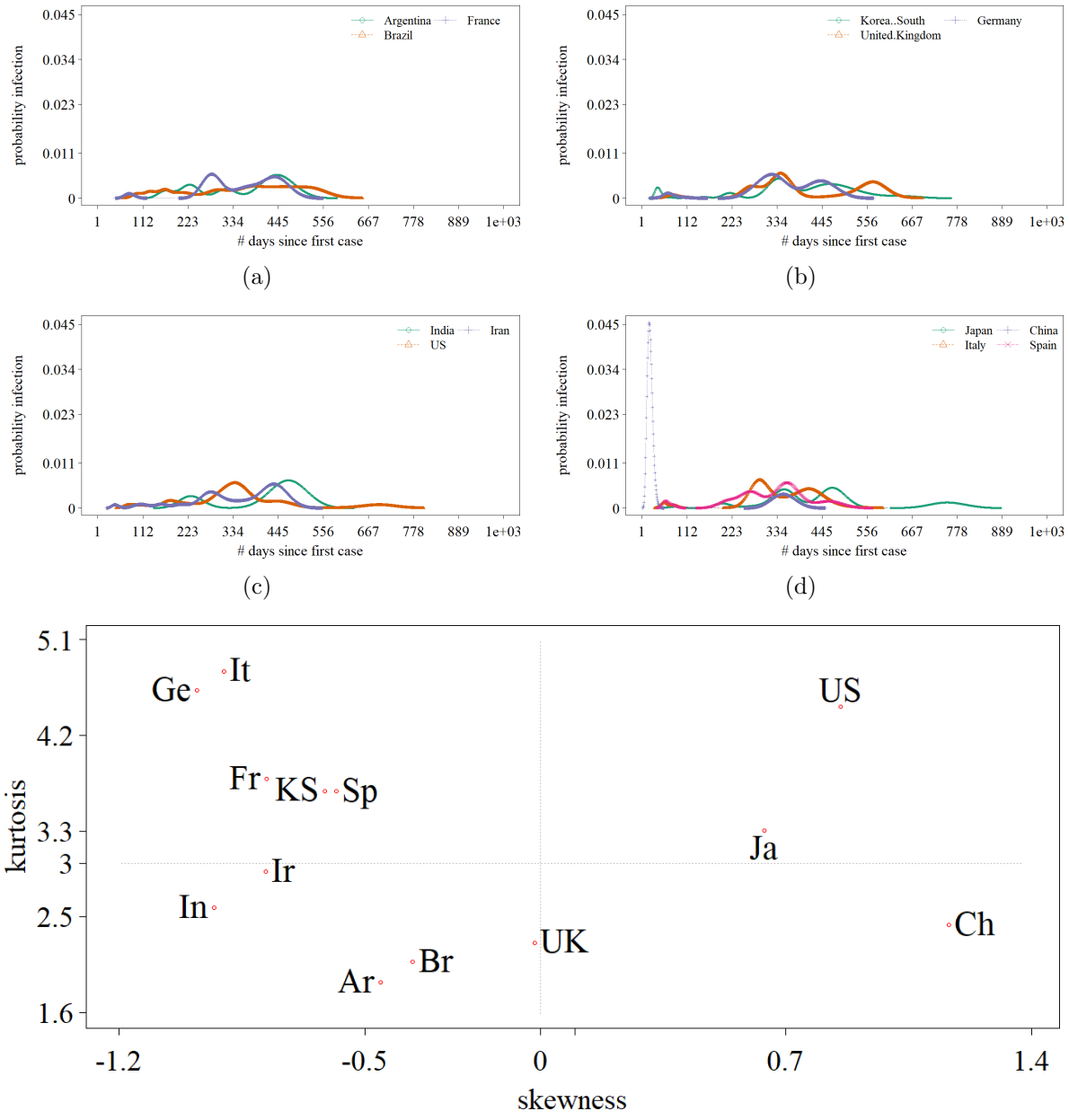
Table 7: Estimates of the GenSA-based near-optimal MNCB models for the national COVID-19 daily incidence time data until 2021-03-28 with respect to Argentina, Brazil, China, France, Germany, India, Iran, Italy, Japan, Korea, South - KS, Spain, United Kingdom - UK, and United States of America - US.

Figure 5 brings the shapes of the COVID-19 national incidence trajectories to the same picture, considering the MNCB fitted models. One can therefore compare the velocity of the occurrence of new cases (i.e. probability infection) before and after the global and local plateau phases since the first contamination case. For the sake of illustration, from Figures 5(a) - 5(d), one can conclude that China has presented the most vertiginous increasing-plateau-decreasing pandemic cycle, followed by spaced and less expressive cycles. As a matter of fact, the MNCB approach has also modelled departed cycles in Spain, United Kingdom, and Italy. However, differently from China, the first pandemic cycle in these countries has not been the main one. Given the available data, it is predicted that all countries overcome the pandemic until 800 days since the first case, but Korea, South and Spain.

Figure 5(e) allows one to compare the thirteen countries in the same terms, via skewness and kurtosis estimates of the fitted NCB probability infection distributions. The greater the value of the skew of the probability distribution, in absolute terms, the greater the difference between the left and right sides of the pandemic model, centred at a given location point (e.g. the mean) of the time until contamination since the first case. Extending [13], it is claimed here that a skew less or equal to zero might be preferred. In the case of negative skew, the decision makers would have more time to plan and review intervention policies during the first part of the pandemic trajectory, and experiment a faster decay in the number of new cases during the last part. Further, in the case of a zero-skew, the health managers wold have a better forecasting capability once the behaviour of the first part of the pandemic series would reflect the one of the second part in some way.

Regarding the kurtosis, it reflects the velocity with which the MNCB model achieves

100% of its probability mass. In other terms, the kurtosis reflects how fast the territory would face and beat the disease. Thus, from Figure 5(e), one can infer that China has presented the fastest and more traumatic pandemic cycle, though under a high level of subsequent control against the disease. In fact, China seems to be a case apart from the remaining countries under study. In turn, Germany and Italy seem to be facing the better trajectory in terms of preparation to the challenging scenarios of the pandemic that would be coming. The skewness-kurtosis plane also suggest similarity of the COVID-19 relative incidence time series in countries like Japan and Iran.



(e) Skewness vs kurtosis

Figure 5: Comparison of shapes of the MNCB probability distributions, regarding the prediction of the national daily incidence of COVID-19 since the first register, according to Johns Hopkins University data set [55]. In Figures 5(a) - 5(d), the shape of the national MNCB-based daily probability infection is exhibited. In Figure 5(e), one has the first two letters as mnemonics for Argentina, Brazil, China, France, Germany, India, Iran, Italy, Japan, and Spain. In turn, it was considered KS for Korea, South, and UK for United Kingdom.

4.6. Computational Cost Analysis

Table 8 summarises the time consumption by the part of the models during training and test exercises, per country and on average. Differently from [13], one can see that the ETS modelling is the cheaper framework, followed by ARIMA and then NCB. In turn, MNCB, SIR, SEIR, and SIS have required the maximum allowed time of the GSA optimiser. SIR.sin, on the other hand, has demanded more than the maximal available time, evidencing that each MSE-based fitness execution of this formalism is more expensive than the one of the remaining alternatives.

model	phase	Ar	Br	Ch	Fr	Ge	In	Ir	It	Ja	KS	Sp	UK	US	Average
MNCB	training	600.030	600.090	600.160	600.170	600.150	600.090	600.160	600.080	600.190	600.100	600.090	600.090	600.050	600.112
	test	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.004
NCB	training	446.530	600.610	505.550	385.720	500.580	435.390	539.740	383.590	600.490	538.370	601.800	600.340	600.150	518.374
	test	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.000
SIR	training	602.22	600.82	601.06	602.84	603.00	601.12	602.18	601.06	602.35	601.36	600.39	600.28	601.05	601.518
	test	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.00	0.01	0.00	0.00	0.012
SEIR	training	628.02	606.20	600.01	600.13	613.27	607.84	608.01	608.44	610.22	607.53	619.46	604.58	603.61	609.025
	test	0.01	0.00	0.02	0.02	0.00	0.02	0.00	0.00	0.02	0.02	0.01	0.01	0.01	0.011
SIS	training	600.99	602.09	600.65	606.32	607.83	603.09	602.75	600.95	601.86	602.54	601.44	600.89	602.21	602.585
	test	0.01	0.01	0.00	0.02	0.00	0.02	0.03	0.02	0.01	0.02	0.02	0.00	0.01	0.013
SIR.sin	training	1500.81	600.96	2926.44	1270.48	602.59	1025.70	741.70	600.15	603.46	789.12	600.53	600.58	1165.78	1002.177
	test	2.38	1.84	8.06	1.98	1.33	1.72	1.88	1.99	0.87	3.10	0.25	0.48	1.68	2.120
ARIMA	training	0.66	0.26	0.06	0.07	1.05	0.22	0.12	1.92	0.81	0.03	0.44	0.03	0.01	0.437
	test	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.000
ETS	training	0.09	0.17	0.17	0.11	0.24	0.14	0.19	0.18	0.16	0.11	0.12	0.14	0.08	0.146
	test	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.000
Average		270.123	215.211	374.431	247.694	209.279	233.949	221.187	199.880	215.734	224.443	216.034	214.809	255.328	238.316

Table 8: Time consumption (in seconds) for training and testing near-optimal MNCB, NCB, SIR, SEIR, SIS, SIR.sin, ARIMA, and ETS models for each COVID-19 daily incidence time data taken into account (Argentina - Ar, Brazil - Br, China - Ch, France - Fr, Germany - Ge, India - In, Iran - Ir, Italy - It, Japan - Ja, Korea, South - KS, Spain - Sp, United Kingdom - UK, US)

5. Conclusion

Pandemics have been a public health issue for organisations and governments around the world. For instance, COVID-19 has played decisive role for profound culture, economic, and social changes. Thus, predicting and comparing incidence trajectories among territories are paramount. The present paper has provided a relatively simple method for performing these two exercises. It is expected that the proposed NCB approach complements analogous studies at the regional and national level and might be useful in the assessment of plans and emerging disease outbreaks.

Though limited to one-peak shape fitting, the NCB approach has performed better than near-optimal versions of established epidemic models (e.g. SIR, SEIR, SIS) as well as time series formalisms (e.g. ARIMA and ETS) for both fitting previous incidence times series and forecasting future values. The results of the methods with respect to a number of performance metrics underlie this argument. In turn, the NCB probability distribution shape has showed useful for summarising and comparing the pandemic incidence trajectory among countries, via kurtosis and skewness estimates. From that, caution with respect to Iran, Argentina, and China, for the sake of illustration, has been suggested.

The method has showed to be cheap, demanding less than 7 seconds, of an intermediate notebook, to model and forecast the COVID-19 daily incidence in each one of thirteen

countries, for a time horizon of 600 days. Thus, considering a database platform that promotes a daily update of the incidence time series (e.g. the Johns Hopkins University [55]), the proposed models can be easily updated. The `Pand-Pred` user interface, freely provided at `www.mesor.com.br`, makes use of this reasoning.

A conceptual limitation of the NCB-based framework is the supposition that the incidence of the disease in a given day follows a binomial distribution. Thus, it is considered that one positive diagnostic is independent from another one in that day, something that might disregard from the reality. In addition, the need to set a maximum time horizon for the end of the pandemic cycle may lead to an underestimation of the spread of the disease in the territory. On the other hand, the impossibility of shaping several peaks is a disadvantage of the NCB models. Countries like Iran and the US seem to demand a more flexible approach, in the case of COVID-19. Thus, ongoing research are dedicated to develop modelling alternatives, such as mixtures of PDFs, adapted artificial neural networks, support vector regression, and copulas formalisms.

CRediT authorship contribution statement

Paulo Renato Alves Firmino: Investigation, Methodology, Supervision, Writing - original draft, and Editing. **Jair Paulino de Sales**: Review, Writing - original draft, and editing. **Jucier Gonçalves Junior**: Review, Writing, and editing. **Taciana Araújo da Silva**: Review, Writing, and editing.

Acknowledgements

This work was partially supported by Brazilian national council for scientific and technological development - CNPq.

References

[1] H. Morita, H. Kato, Y. Hayashi, International comparison of behavior changes with social distancing policies in response to covid-19, Available at SSRN 3594035 (2020).

[2] A. Olivera-La Rosa, E. G. Chuquichambi, G. P. Ingram, Keep your (social) distance: Pathogen concerns and social perception in the time of covid-19, Personality and Individual Differences 166 (2020) 110200.

[3] W. H. Organization, Who coronavirus (covid-19) dashboard, 2021. URL: `https://covid19.who.int/`.

[4] R. F. Garry, Mutations arising in sars-cov-2 spike on sustained human-to-human transmission and human-to-animal passage, image 908 (2021) 292.

[5] K. M. Bindayna, S. Crinion, Variant analysis of sars-cov-2 genomes in the middle east, Microbial Pathogenesis 153 (2021) 104741.

[6] J. Patel, F. Nielsen, A. Badiani, S. Assi, V. Unadkat, B. Patel, R. Ravindrane, H. Wardle, Poverty, inequality and covid-19: the forgotten vulnerable, Public health 183 (2020) 110.

[7] R. S. Goldwasser, M. S. d. C. Lobo, E. F. d. Arruda, S. A. Angelo, J. R. L. Silva, A. A. d. Salles, C. M. David, Dificuldades de acesso e estimativas de leitos públicos para unidades de terapia intensiva no estado do rio de janeiro, Revista de Saúde Pública 50 (2016) 19.

[8] O. Bargain, A. Ulugbek, Poverty and covid-19 in developing countries, Bordeaux University (2020).

[9] H. H. Nguyen, C. W. Chan, Multiple neural networks for a long term time series forecast, Neural Computing & Applications 13 (2004) 90–98.

[10] S. Jiang, Z.-G. Yu, V. V. Anh, Y. Zhou, Long-and short-term time series forecasting of air quality by a multi-scale framework, Environmental Pollution 271 (2021) 116381.
[11] S. Moein, N. Nickaeen, A. Roointan, N. Borhani, Z. Heidary, S. H. Javanmard, J. Ghaisari, Y. Gheisari, Inefficiency of sir models in forecasting covid-19 epidemic: a case study of isfahan, Scientific Reports 11 (2021) 1–9.
[12] I. Holmdahl, C. Buckee, Wrong but useful—what covid-19 epidemiologic models can and cannot tell us, New England Journal of Medicine 383 (2020) 303–305.
[13] P. R. A. Firmino, J. P. de Sales, J. G. Júnior, T. A. da Silva, A non-central beta model to forecast and evaluate pandemics time series, Chaos, Solitons & Fractals 140 (2020) 110211.
[14] G. Chowell, H. Nishiura, L. M. Bettencourt, Comparative estimation of the reproduction number for pandemic influenza from daily case notification data, Journal of the Royal Society Interface 4 (2007) 155–166.
[15] M. G. Lucero, M. T. Inobaya, L. T. Nillos, A. G. Tan, V. L. F. Arguelles, C. J. C. Dureza, E. S. Mercado, A. N. Bautista, V. L. Tallo, A. V. Barrientos, et al., National influenza surveillance in the Philippines from 2006 to 2012: seasonality and circulating strains, BMC infectious diseases 16 (2016) 762.
[16] T. M. Sharp, K. R. Ryff, L. Alvarado, W.-J. Shieh, S. R. Zaki, H. S. Margolis, B. Rivera-Garcia, Surveillance for chikungunya and dengue during the first year of chikungunya virus circulation in Puerto Rico, The Journal of infectious diseases 214 (2016) S475–S481.
[17] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg, et al., Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand (2020).
[18] J. P. Ioannidis, S. Cripps, M. A. Tanner, Forecasting for covid-19 has failed, International journal of forecasting (2020).
[19] P. McCullagh, et al., What is a statistical model?, Annals of statistics 30 (2002) 1225–1310.
[20] A. Agosto, A. Campmas, P. Giudici, A. Renda, A statistical model to monitor covid-19 contagion growth, Available at SSRN 3585930 (2020).
[21] S. Tewari, C. J. Geyer, N. Mohan, A statistical model for wind power forecast error and its application to the estimation of penalties in liberalized markets, IEEE Transactions on Power Systems 26 (2011) 2031–2039.
[22] R. M. Corder, G. A. Paula, A. Pincelli, M. U. Ferreira, Statistical modeling of surveillance data to identify correlates of urban malaria risk: A population-based study in the amazon basin, PloS one 14 (2019) e0220980.
[23] N. P. Jewell, J. A. Lewnard, B. L. Jewell, Caution warranted: using the institute for health metrics and evaluation model for predicting the course of the covid-19 pandemic, 2020.
[24] D. M. Thomas, R. Sturdivant, N. V. Dhurandhar, S. Debroy, N. Clark, A primer on covid-19 mathematical models, Obesity 28 (2020) 1375–1377.
[25] Z. Liao, P. Lan, Z. Liao, Y. Zhang, S. Liu, Tw-sir: time-window based sir for covid-19 forecasts, Scientific reports 10 (2020) 1–15.
[26] Y.-C. Chen, P.-E. Lu, C.-S. Chang, T.-H. Liu, A time-dependent sir model for covid-19 with undetectable infected persons, IEEE Transactions on Network Science and Engineering 7 (2020) 3279–3294.
[27] H. W. Hethcote, Three basic epidemiological models, in: Applied mathematical ecology, Springer, 1989, pp. 119–144.
[28] S. Kushwaha, S. Bahl, A. K. Bagha, K. S. Parmar, M. Javaid, A. Haleem, R. P. Singh, Significant applications of machine learning for covid-19 pandemic, Journal of Industrial Integration and Management 5 (2020).
[29] A. L. Booth, E. Abels, P. McCaffrey, Development of a prognostic model for mortality in covid-19 infection using machine learning, Modern Pathology 34 (2021) 522–531.
[30] M. Mele, C. Magazzino, Pollution, economic growth, and covid-19 deaths in india: a machine learning evidence, Environmental Science and Pollution Research 28 (2021) 2669–2677.
[31] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann,

C. McCague, L. Beer, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans, Nature Machine Intelligence 3 (2021) 199–217.
[32] Z. Car, S. Baressi Šegota, N. Anelić, I. Lorencin, V. Mrzljak, Modeling the spread of covid-19 infection using a multilayer perceptron, Computational and mathematical methods in medicine 2020 (2020).
[33] P. H. Borghi, O. Zakordonets, J. P. Teixeira, A covid-19 time series forecasting model based on mlp ann, Procedia Computer Science 181 (2021) 940–947.
[34] Z. E. Rasjid, R. Setiawan, A. Effendi, A comparison: Prediction of death and infected covid-19 cases in indonesia using time series smoothing and lstm neural network, Procedia Computer Science 179 (2021) 982–988.
[35] K. ArunKumar, D. V. Kalaga, C. M. S. Kumar, M. Kawaji, T. M. Brenza, Forecasting of covid-19 using deep layer recurrent neural networks (rnns) with gated recurrent units (grus) and long short-term memory (lstm) cells, Chaos, Solitons & Fractals 146 (2021) 110861.
[36] P. Melin, D. Sánchez, J. C. Monica, O. Castillo, Optimization using the firefly algorithm of ensemble neural networks with type-2 fuzzy integration for covid-19 time series prediction, Soft Computing (2021) 1–38.
[37] A. Kumar, K. Kaur, A hybrid som-fuzzy time series (somfts) technique for future forecasting of covid-19 cases and mcdm based evaluation of covid-19 forecasting models, in: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), IEEE, 2021, pp. 612–617.
[38] N. S. Punn, S. K. Sonbhadra, S. Agarwal, Covid-19 epidemic analysis using machine learning and deep learning algorithms, MedRxiv (2020).
[39] M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani, L. dos Santos Coelho, Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil, Chaos, Solitons & Fractals 135 (2020) 109853.
[40] S. P. Neuman, Maximum likelihood bayesian averaging of uncertain model predictions, Stochastic Environmental Research and Risk Assessment 17 (2003) 291–305.
[41] S. Panigrahi, H. S. Behera, A hybrid ets–ann model for time series forecasting, Engineering Applications of Artificial Intelligence 66 (2017) 49–59.
[42] P. S. D. M. Neto, P. R. A. Firmino, H. Siqueira, Y. D. S. Tadano, T. A. Alves, J. F. L. De Oliveira, M. H. D. N. Marinho, F. Madeiro, Neural-based ensembles for particulate matter forecasting, IEEE Access 9 (2021) 14470–14490.
[43] C. M. Liapis, A. Karanikola, S. Kotsiantis, An ensemble forecasting method using univariate time series covid-19 data, in: 24th Pan-Hellenic Conference on Informatics, 2020, pp. 50–52.
[44] S. Shastri, K. Singh, S. Kumar, P. Kour, V. Mansotra, Deep-lstm ensemble framework to forecast covid-19: an insight to the global pandemic, International Journal of Information Technology (2021) 1–11.
[45] X. Lu, W. Zhou, X. Ding, X. Shi, B. Luan, M. Li, Ensemble learning regression for estimating unconfined compressive strength of cemented paste backfill, IEEE Access 7 (2019) 72125–72133.
[46] A. Zameer, J. Arshad, A. Khan, M. A. Z. Raja, Intelligent and robust prediction of short term wind power using genetic programming based ensemble of neural networks, Energy conversion and management 134 (2017) 361–372.
[47] H. R. Kadkhodaei, A. M. E. Moghadam, M. Dehghan, Hboost: A heterogeneous ensemble classifier based on the boosting method and entropy measurement, Expert Systems with Applications 157 (2020) 113482.
[48] X. Zhao, Q. Jiao, H. Li, Y. Wu, H. Wang, S. Huang, G. Wang, Ecfs-dea: an ensemble classifier-based feature selection for differential expression analysis on expression profiles, BMC bioinformatics 21 (2020) 43.
[49] J. F. de Oliveira, E. G. Silva, P. S. de Mattos Neto, A hybrid system based on dynamic selection for time series forecasting, IEEE Transactions on Neural Networks and Learning Systems (2021).
[50] H. E. Randolph, L. B. Barreiro, Herd immunity: understanding covid-19, Immunity 52 (2020) 737–741.
[51] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong,

et al., Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia, New England Journal of Medicine (2020).
[52] S. Sanche, Y. T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, R. Ke, High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2, Emerging infectious diseases 26 (2020) 1470–1477.
[53] B. Spellberg, T. B. Nielsen, A. Casadevall, Antibodies, immunity, and covid-19, JAMA internal medicine (2020).
[54] M. Lipsitch, Who is immune to the coronavirus, The New York Times 14 (2020).
[55] Center for Systems Science and Engineering, Covid-19 data set from the Johns Hopkins University, Center for Systems Science and Engineering, 2020. URL: `https://github.com/CSSEGISandData/COVID-19`.
[56] P. S. de Mattos Neto, G. D. Cavalcanti, P. R. Firmino, E. G. Silva, S. R. V. Nova Filho, A temporal-window framework for modeling and forecasting time series, Knowledge-Based Systems (2020) 105476.
[57] G. Casella, R. L. Berger, Statistical inference, 2nd ed., Duxbury Pacific Grove, CA, 2002.
[58] C. Tsallis, D. A. Stariolo, Generalized simulated annealing, Physica A 233 (1996) 395–406.
[59] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020. URL: `https://www.R-project.org/`.
[60] M. J. Keeling, P. Rohani, Modeling infectious diseases in humans and animals, Princeton University Press, 2008.
[61] J. D. Cryer, K.-S. Chan, Time series analysis: with applications in R, Springer Science & Business Media, 2008.
[62] R. Hyndman, A. B. Koehler, J. K. Ord, R. D. Snyder, Forecasting with exponential smoothing: the state space approach, Springer Science & Business Media, 2008.
[63] Yang Xiang, S. Gubian, B. Suomela, J. Hoeng, Generalized simulated annealing for efficient global optimization: the GenSA package for R., The R Journal Volume 5/1, June 2013 (2013). URL: `https://journal.r-project.org/archive/2013/RJ-2013-002/index.html`.
[64] O. Santos Baquero, F. Silveira Marques, EpiDynamics: Dynamic Models in Epidemiology, 2020. URL: `https://CRAN.R-project.org/package=EpiDynamics`, r package version 0.3.1.
[65] R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O'Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, F. Yasmeen, forecast: Forecasting functions for time series and linear models, 2020. URL: `http://pkg.robjhyndman.com/forecast`, r package version 8.12.
[66] P. R. A. Firmino, P. S. de Mattos Neto, T. A. Ferreira, Correcting and combining time series forecasters, Neural Networks 50 (2014) 1–11.
[67] P. S. de Mattos Neto, G. D. Cavalcanti, P. R. Firmino, E. G. Silva, S. R. V. Nova Filho, A temporal-window framework for modelling and forecasting time series, Knowledge-Based Systems 193 (2020) 105476.