Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables

Ramon Gomes da Silva[a,*], Matheus Henrique Dal Molin Ribeiro[a,b],
Viviana Cocco Mariani[c,d], Leandro dos Santos Coelho[a,d]

[a] *Industrial & Systems Engineering Graduate Program (PPGEPS), Pontifical Catholic University of Parana (PUCPR), 1155, Rua Imaculada Conceicao, Curitiba, PR, Brazil, 80215-901*
[b] *Department of Mathematics, Federal Technological University of Parana (UTFPR), Via do Conhecimento, KM 01 - Fraron, Pato Branco, PR, Brazil, 85503–390*
[c] *Mechanical Engineering Graduate Program (PPGEM), Pontifical Catholic University of Parana (PUCPR), 1155, Rua Imaculada Conceicao, Curitiba, PR, Brazil, 80215-901*
[d] *Department of Electrical Engineering, Federal University of Parana (UFPR), 100, Avenida Coronel Francisco Heraclito dos Santos, Curitiba, PR, Brazil, 81530-000*

ARTICLE INFO

ABSTRACT

The novel coronavirus disease (COVID-19) is a public health problem once according to the World Health Organization up to June 24th, 2020, more than 9.1 million people were infected, and more than 470 thousand have died worldwide. In the current scenario, the Brazil and the United States of America present a high daily incidence of new cases and deaths. Therefore, it is important to forecast the number of new cases in a time window of one week, once this can help the public health system developing strategic planning to deals with the COVID-19. The application of the forecasting artificial intelligence (AI) models has the potential of deal with dynamical behavior of time-series like of COVID-19. In this paper, Bayesian regression neural network, cubist regression, *k*-nearest neighbors, quantile random forest, and support vector regression, are used stand-alone, and coupled with the recent pre-processing variational mode decomposition (VMD) employed to decompose the time series into several intrinsic mode functions. All AI techniques are evaluated in the task of time-series forecasting with one, three, and six-days-ahead the cumulative COVID-19 cases in five Brazilian and American states, with a high number of cases up to April 28th, 2020. Previous cumulative COVID-19 cases and exogenous variables as daily temperature and precipitation were employed as inputs for all forecasting models. The models' effectiveness are evaluated based on the performance criteria. In general, the hybridization of VMD outperformed single forecasting models regarding the accuracy, specifically when the horizon is six-days-ahead, the hybrid VMD–single models achieved better accuracy in 70% of the cases. Regarding the exogenous variables, the importance ranking as predictor variables is, from the upper to the lower, past cases, temperature, and precipitation. Therefore, due to the efficiency of evaluated models to forecasting cumulative COVID-19 cases up to six-days-ahead, the adopted models can be recommended as a promising models for forecasting and be used to assist in the development of public policies to mitigate the effects of COVID-19 outbreak.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

The new coronavirus disease (COVID-19) is a virus infectious disease induced by severe acute respiratory syndrome coronavirus 2 (SARS-CoV2). According to the World Health Organization (WHO), most of the population will mild to moderate respiratory illness and recover without requiring special treatment [1].

However, several studies are being developed, and preliminary results indicated that people with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, obesity, and cancer are more likely to develop serious injuries [2–7]. Also, the COVID-19 can cause extensive and multiple lung injuries [8], thus compromising the respiratory system of patients. In this context, the demand for devices that assist in the performance of breathing-related movements have increased.

Due to the serious damage caused by COVID-19, according to WHO, up to June 24th 2020, more than 9.1 million people were already infected, as well as more than 470 thousand people worldwide have now died with the coronavirus. Indeed, considering the

* Corresponding author.
E-mail addresses: gomes.ramon@pucpr.edu.br (R.G. da Silva), mribeiro@utfpr.edu.br (M.H.D.M. Ribeiro), viviana.mariani@pucpr.br (V.C. Mariani), leandro.coelho@pucpr.br (L.d.S. Coelho).

current scenario of the health system worldwide, the overcrowding could be observed in some countries, like Italy, Spain and perhaps Brazil. In Brazilian context, believed that the average of 3388 municipalities could have a significant deficit in hospital beds. Especially, the deficit is projected to occur in Brazilian North and Northeast regions, which means exceeding health care capacity due to the COVID-19 [9].

Considering the importance of knowing the difficult epidemiological scenario for COVID-19 on a short-term horizon, to mitigate the effects of this pandemic, the development of efficient and effective forecasting models also has a positive impact on product reasonably accurate success rates forecasts the immediate future. Also, these models allow health managers to develop strategic planning and perform decision-making as assertively as possible. For this purpose, epidemiological models can be used, as it has been widely adopted in [10,11]. Alternatively, linear forecasting models [12–14], artificial intelligence (AI) approaches [15,16], as well as hybrid forecasting models [17,18] proved to be effective tools to forecast COVID-19 cases. The advantages of AI approaches for time series forecasting lie in the flexibility of dealing with different kinds of response variables, as well as to the ability of these approaches to learning data dynamical behavior, complexity and accommodate nonlinearities, such as the observed in epidemiological data [19]. Besides, hybrid methodologies allow us to combine several techniques such as pre-processing methods and single forecasting models.

By the coupling of some methods, it is possible to use the specialty of each one to deal with different characteristics and therefore building an effective model. In context of the preprocessing techniques, especially signal decomposition methods, the variational mode decomposition (VMD) [20] is an effective approach to decompose a dimensional signal into an ensemble of band-limited modes with specific bandwidth in a spectral domain applied in several fields [21–23], once can deal with nonlinearities, and nonstationarity inherent to time series. Considering the intrinsic mode function (IMF) obtained through VMD, it is hard to choose AI models to train and forecasting the VMD components. Therefore, based on this understanding, some models are coupled with VMD and are described in the following.

Due to the necessity of understanding the COVID-19 outbreak, and the associated factors, or exogenous variables, some studies are being conducted considering the social environment, climatic variables, pollution, and population density [24–28]. In this direction, in a general aspect, Sobral et al. [29] investigated the effects of climatic variables in COVID-19 spread for 166 countries. The authors argued that increasing the temperature reduced the COVID-19 cases, and precipitation also has a positive correlation with SARS-CoV2 cases. In the sequence, for Brazil, Auler et al. [30] evaluated how meteorological conditions such as temperature, humidity, and rainfall can affect the spread of COVID-19 in five Brazilian cities. The authors concluded that higher mean temperatures and average relative humidity might support the COVID-19 transmission. Considering the United States of America (USA) weather aspects, especially for the New York state, Bashir et al. [31] inferred that average and minimum temperature and air quality are significantly associated with the COVID-19 pandemic. All previously mentioned studies tried related the climatic variables with COVID-19 but in those papers were not incorporated in time series models to forecasting COVID-19 cases. However, we think that incorporating the exogenous climatic variables in forecasting models can help to understand the data dynamic, and perhaps more efficient forecasting models could be obtained [32].

In this respect, for forecasting of cumulative cases of COVID-19, the objective of this paper is to explore and compare the predictive capacity of Bayesian regression neural network (BRNN), cubist regression (CUBIST), k-nearest neighbors (KNN), quantile random forest (QRF), and support vector regression (SVR) when are used stand-alone, and a hybrid framework composed by VMD coupled with previously mentioned models. In this study were used as datasets the number about the cumulative cases of COVID-19 from five Brazilian states (Amazonas - AM, Ceara - CE, Pernambuco - PE, Rio de Janeiro - RJ, and Sao Paulo - SP), the first state from north region, the second and third states from northeast region, and the other two states from southeast region. Also were considered five American states (California - CA, Illinois - IL, Massachusetts - MA, New Jersey - NJ, and New York - NY). The choice of these states was made through the largest number of new cases of COVID-19 up to April 28th 2020.

In the task of forecasting horizons of the time series one, three, and six-days-ahead of cumulative COVID-19 cases are adopted to evaluates the forecasting efficiency of the different models. Additionally, previous COVID-19 cases, and exogenous variables such as daily temperature (maximum and minimum), and precipitation are employed as inputs for each evaluated model. The output-of-sample forecasting accuracy of each model is compared by performance metrics such as the improvement percentage (IP) index, symmetric mean absolute percentage error (sMAPE), and relative root mean squared error (RRMSE). Also, the importance of each input variable is presented for each country.

Forecasting models are impacted by the small dataset effect and the prediction of cases of COVID-19 a challenging task. The choice of the forecasting and pre-processing approaches is due to the fact that even that non-linear and AI models need large datasets to properly learn the data pattern, the use of exogenous variables (climate variables) and past values of the response variable overcomes this drawback.

VMD decomposes a time series into its intrinsic mode functions adaptively and non-recursively obtaining a set of sub-series with different features from low-frequency to high-frequency. The adoption of VMD with modes in conjunction with nonlinear prediction models of machine learning is a powerful framework to approach small datasets in forecasting task. In addition, BRNN and SVR approaches are capable of handling small samples, which makes them attractive for this study.

The contributions of this paper can be summarized as follows:

- The first contribution is related to the proposal of two frameworks, non-decomposed and decomposed models, applied in the task of forecasting the new cumulative cases of COVID-19 in five Brazilian and American states. It is expected that these evaluated models can be used as most accurate approaches to perform decision-making to structure the health system to avoid overcrowding in hospitals, and preventing new deaths.
- The second contribution, we can highlight the use of a distinct set of AI models based on machine learning approaches regarding learning structure, even as the recent effective preprocessing VMD to forecasting the Brazilian and American COVID-19 new cumulative cases. The forecasting models BRNN, CUBIST, KNN, QRF, SVR, and pre-processing VMD method were chosen once that have reached success into several fields of regression and time series forecasting [33–36];
- Also, this paper evaluates AI models in a multi-day-ahead forecasting strategy coupled with climatic exogenous inputs. The range of the forecasting time horizon allows us to verify the effectiveness of the predicting models in different scenarios, associated with inputs such as previous COVID-19 cumulative cases, temperature, and precipitation, allowing that the models achieve high forecasting accuracy. Finally, their results can help in planning actions to improve the health system to contain the COVID-19 deaths.

The remainder of this paper is organized as follows: Section 2.1 a brief description of the dataset adopted in this

paper is presented. The forecasting models applied in this study are described in Section 2.2. Section 3 details the procedures applied in the research methodology. Results obtained and related discussion about models forecasting performance are mentioned in Section 4. Finally, Section 5 concludes this study with considerations and some directions for future research proposals.

2. Material and methods

This section presents a description of the material analyzed (Section 2.1), as well as the models description applied in this paper (Section 2.2).

2.1. Dataset description

The collected dataset refers to the COVID-19 cumulative cases that occurred in five states of the Brazil and the USA until April 28th, 2020. For the Brazilian context, the dataset was collected from an API (Application Program Interface) [37] that retrieves the daily information about COVID-19 cases from all 27 Brazilian State Health Offices, assembles and makes them publicly available. And for USA context, the dataset was collected from "COVID-19 Data Repository" on Github provided by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [38]. The cumulative confirmed cases and deaths of each state, and the period from the first and last reports, are illustrated in Table 1.

The climatic exogenous variables were retrieved from the "*Instituto Nacional de Meteorologia*" (INMET) [39] for dataset from Brazil, while the USA climate dataset were taken into a count from the daily global historical climatology network that was retrieved from the National Centers for Environmental Information (NCEI) from the National Oceanic and Atmospheric Administration [40], by using `rnoaa` package [41]. For each state, considering the daily available information, minimum and maximum temperature ($°C$), and precipitation (mm) were select as climatic exogenous inputs to each forecasting model applied in this study. The measurement period of each state is variable, this is due because the record of the first case of the disease may differ from state to state. The summary of the climatic variables used is described in Table 2.

The heat-map of the cumulative confirmed cases from the Brazil and the USA in each of the five states analyzed are presented in Fig. 1. In that figure can be seen that the states with the highest number of COVID-19 cumulative cases are SP and NY, respectively, in Brazil and the USA, the states with the highest demographic index in both countries.

2.2. Methodologies

This section presents a summary of each model employed in the data analysis.

- BRNN is a kind of feedforward artificial neural network, a two-layer neural network, composed by one input and one hidden layer, which uses the Bayesian methods, such as empirical Bayes, for parameter estimation, to avoid overfitting [42]. In the BRNN formulation, the variances are regularization parameters, in which the trade-off between goodness-of-fit and smoothing can be controlled. Also, in this approach the method of [43] is used to assign initial weights of neural network and the Gauss-Newton training algorithm to perform the optimization. For the datasets evaluated in this paper, the BRNN becomes attractive once it can deal with small samples, as well as it has a lower computational cost.

- CUBIST is a rule-based algorithm used to build forecasting models (in the time series field) based on the analysis of input data [44]. It estimates the target values by establishing regression models with one or more rules (committee/ensemble of rules) based on the input set. These rules are employed based on a combination of conditions with a linear function (in general linear regression). When the rule satisfies all conditions defined in the learning process, this approach can execute multiple rules once and find different linear functions suitable to forecast COVID-19 cases. However, if the standard deviation reduction value is smaller or equal to the expected error for sub-tree, some leaves are pruned to avoid overfitting [15].

- KNN is an instance-based learner model designed to solve classification and regression problems [45]. In fact, in the time series context, the KNN searches k nearest past similar values in the input set (past COVID-19 values, and climatic variables), in which these k values are namely nearest neighbors. In this context, to find the nearest values, a similarity measure is adopted. The k-nearest neighbors are those that similarity measure between past cases and new cases is the smallest. Considering that the set of k-nearest neighbors are defined, the forecasting of new COVID-19 cases is obtained through of average of past similar values. In contrast to the simplicity of this supervised learning, the computational cost may be a disadvantage [32].

- QRF approach is an extension of the random forests (RF) ensemble learning model [46]. It provides information about the full conditional distribution of the response variable, not only about the conditional mean. In this approach, the use of conditional quantile is to enhance the RF performance, which makes this a consistent approach [47]. The main assumption about QRF lies in that weighted observations can be used for estimating the conditional mean [48]. Additionally, while the RF approach keeps in the results information as regards the average cases of COVID-19 of the leaves, the QRF keeps all COVID-19 cases contained in the leaves.

- SVR is a type support vector machine that consists in determining support vectors close to a hyperplane, which maximizes the

Table 1
Summary of COVID-19 cases by country and state.

Country	State	no. of observed days	First reported	Last reported	Cumulative cases	Cumulative deaths
Brazil	AM	47	13/03/2020	28/04/2020	4337	351
	CE	44	16/03/2020	28/04/2020	6985	403
	PE	48	12/03/2020	28/04/2020	5724	508
	RJ	55	05/03/2020	28/04/2020	8504	738
	SP	64	25/02/2020	28/04/2020	24,041	2049
USA	CA	94	26/01/2020	28/04/2020	46,164	1864
	IL	96	24/01/2020	28/04/2020	48,102	2125
	MA	87	01/02/2020	27/04/2020	56,462	3003
	NJ	55	05/03/2020	28/04/2020	113,856	6442
	NY	58	02/03/2020	28/04/2020	295,106	22,912

Table 2
Descriptive measures for climatic variables by country and state.

Country	State	Variable	Minimum	Median	Mean	Maximum
Brazil	AM	Minimum temperature (°C)	24.76	26.20	26.36	28.28
		Maximum temperature (°C)	25.29	27.05	27.24	29.55
		Precipitation (mm)	0.00	0.11	0.33	2.40
	CE	Minimum temperature (°C)	25.14	26.62	26.60	27.90
		Maximum temperature (°C)	25.91	27.73	27.68	28.99
		Precipitation (mm)	0.00	0.12	0.25	1.31
	PE	Minimum temperature (°C)	23.36	25.18	25.05	26.74
		Maximum temperature (°C)	24.27	26.30	26.10	27.96
		Precipitation (mm)	0.00	0.14	0.23	1.33
	RJ	Minimum temperature (°C)	19.07	21.23	21.57	25.33
		Maximum temperature (°C)	19.69	22.16	22.56	26.49
		Precipitation (mm)	0.00	0.03	0.13	1.32
	SP	Minimum temperature (°C)	17.60	19.99	20.09	23.40
		Maximum temperature (°C)	18.76	21.11	21.37	25.03
		Precipitation (mm)	0.00	0.00	0.12	1.19
USA	CA	Minimum temperature (°C)	-1.76	4.90	4.91	11.92
		Maximum temperature (°C)	10.60	18.33	18.54	28.59
		Precipitation (mm)	0.01	4.66	20.98	162.67
	IL	Minimum temperature (°C)	-19.42	-0.42	-0.75	14.63
		Maximum temperature (°C)	-6.39	8.40	8.99	26.06
		Precipitation (mm)	0.00	4.29	23.30	196.47
	MA	Minimum temperature (°C)	-14.75	-0.86	-1.76	5.39
		Maximum temperature (°C)	-2.77	8.32	8.16	18.70
		Precipitation (mm)	0.00	6.84	34.66	320.86
	NJ	Minimum temperature (°C)	-11.80	1.60	0.96	7.27
		Maximum temperature (°C)	0.54	10.78	11.22	22.02
		Precipitation (mm)	0.00	5.44	31.53	274.04
	NY	Minimum temperature (°C)	-20.48	-2.61	-3.75	4.60
		Maximum temperature (°C)	-7.98	5.62	6.23	17.88
		Precipitation (mm)	0.00	9.94	27.61	167.94

(a) Brazil   (b) USA
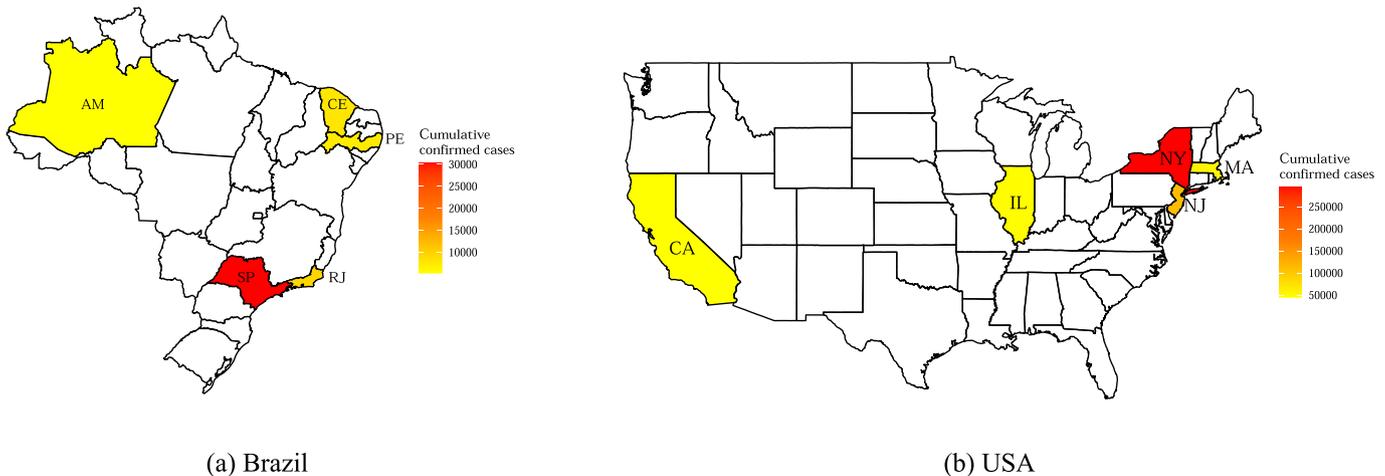
Fig. 1. Heatmap of the cumulative conïrmed cases to five states from Brazil and USA.

margin between two-point classes obtained from the difference between the target value and a threshold. To deal with nonlinear problems SVR takes into account kernel functions, which calculates the similarity between two observations through the inner product. In this paper, the linear kernel is adopted. The main advantage of the use of SVR lie in its capacity to capture the predictor non-linearity and then use it to improve the forecasting cases. Also, it is advantageous to employ to forecast COVID-19 cumulative cases, once the samples are small [15,49].

- VMD is a pre-processing technique in the field of decomposition approaches, which decomposes a time series into a finite and predefined k number of IMF or mode functions. In a general way, VMD reproduces the decomposed signal with different sparsity properties [20]. There are three main concepts related to VMD, which are Wiener filtering, Hilbert transform and analytic signal, and frequency mixing and heterodyne demod-

ulation. Sparsity prior of each mode is chosen as bandwidth in the spectral domain and can be accessed by the following scheme for each model: (i) compute associated analytic signal utilizing the Hilbert transform to obtain a unilateral frequency spectrum; (ii) shift frequency spectrum of mode to baseband by mixing the exponential tune to the respective estimated center frequency; and (iii) the bandwidth estimated through the Gaussian smoothness of the demodulated signal [21].

3. Proposed forecasting framework

This section describes the main steps in the data analysis adopted by BRNN, CUBIST, KNN, QRF, SVR, and VMD based models.

Step 1: First, the dataset output variables are decomposed into five IMFs by performing VMD. The lag equal to 2 was chosen by
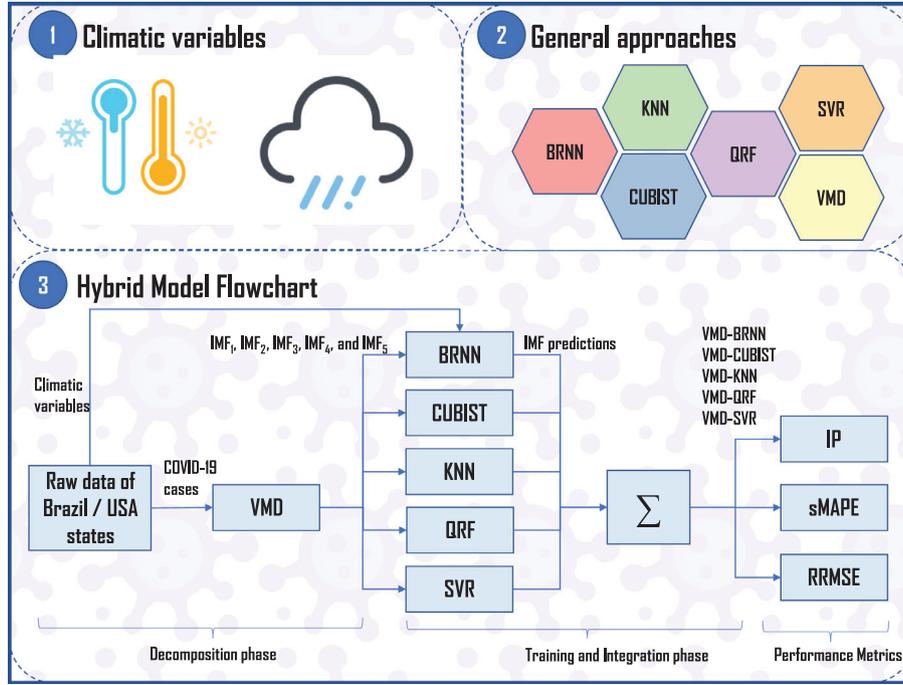
Fig. 2. Proposed forecasting framework.

grid-search, applied on the IMFs creating four inputs from the lags, and applied on the exogenous inputs as well. Further, the new data is split into training and test sets. The test set consists of the last six observations and the training set defined by the remaining samples. In the training state, leave one-out-cross-validation with time slice was adopted, such as developed by [32].

Step 2: Each IMF is trained with each model described in Section 2.2 using time-slice validation approach. Next, the IMF predictions were reconstructed by a simple summation-grouping model, in other words, the IMF is trained by the same model and is summed. Then, five predictions outputs were generated named VMD–BRNN, VMD–CUBIST, VMD–KNN, VMD–QRF, and VMD–SVR.

Step 3: A recursive strategy is employed to develop multi-days-ahead COVID-19 cases forecasting [15]. Regarding this, one model is fitted for one-day-ahead forecasting, then the recursive strategy uses this forecasting result as an input for the same model to forecast the next step, continuing until the desirable forecasting horizon. In this study, the aim is to obtain the cases up to h next days, especially up to 1 (ODA, one-day-ahead), 3 (TDA, three-days-ahead), and 6-days-ahead (SDA, six-days-ahead), respectively. The following forecasting structures are considered,

$$\hat{y}_{(t+h)} = \begin{cases} \hat{f}\{y_{(t+h-1)}, y_{(t+h-2)}, \mathbf{X}_{(t+h-1)}\} & \text{if } h = 1, \\ \hat{f}\{\hat{y}_{(t+h-1)}, \hat{y}_{(t+h-2)}, \mathbf{X}_{(t+h-3)}\} & \text{if } h = 3, \\ \hat{f}\{\hat{y}_{(t+h-1)}, \hat{y}_{(t+h-2)}, \mathbf{X}_{(t+h-6)}\} & \text{if } h = 6, \end{cases} \quad (1)$$

where $\hat{f}$ is a function that maps the cumulative COVID-19 cases, $\hat{y}(t + h)$ is the forecast of cumulative cases in horizon $h =$ 1, 3 and 6, $y(t + h - 1)$, $y(t + h - 2)$ are the previous observed, $\hat{y}(t + h - 1)$, $\hat{y}(t + h - 2)$ are the predicted cumulative cases, $\mathbf{X}(t + h - n_x)$ is the exogenous inputs vector at the maximum lag of inputs ($n_x = 1$ if $h = 1$, $n_x = 3$ if $h = 3$, and $n_x = 6$ if $h = 6$).The analyses are developed using R software [50]. All hyperparameters employed in this study are presented in Tables B.1 and B.2 in Appendix B.

Step 4: To evaluate the effectiveness of adopted models, from obtained forecasts out-of-sample (test set), performance IP, sMAPE,

and RRMSE criteria are computed as

$$IP = 100 \times \frac{M_c - M_b}{M_c}, \quad (2)$$

$$sMAPE = \frac{2}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}, \quad (3)$$

$$RRMSE = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}{\frac{1}{n} \sum_{i=1}^{n} y_i}, \quad (4)$$

where n is the number of observation, y_i and $\hat{y}_i$ are the i-th observed and predicted values, respectively. Also, the M_c and M_b represent the performance measure of compared and best models, respectively.

Fig. 2 presents the proposed forecasting framework.

4. Results

This section describes the results of the developed experiments in forecasting out-of-sample (test set). First, Section 4.1 compares the results of evaluated models over ten datasets and three forecasting horizons adopted. In Tables A.1 and A.2 in Appendix A, the best results regarding accuracy are presented in bold. Additionally, Figs. 3 and 4 illustrate the relation between observed and predicted values achieved by models with the best set of performance measures depicted in Tables A.1 and A.2, as well as box-plots for out-of-sample errors, are illustrated in Fig. 5. Also, Fig. 6 illustrates the variable importance of each input (both lags and exogenous inputs) used in the models' predictions.

4.1. Performance measures for compared models

In this section, the main results achieved by the best model regarding sMAPE and RRMSE criteria are presented for short-term forecasting multi-days-ahead of cumulative cases of COVID-19 from five Brazilian and five American states.
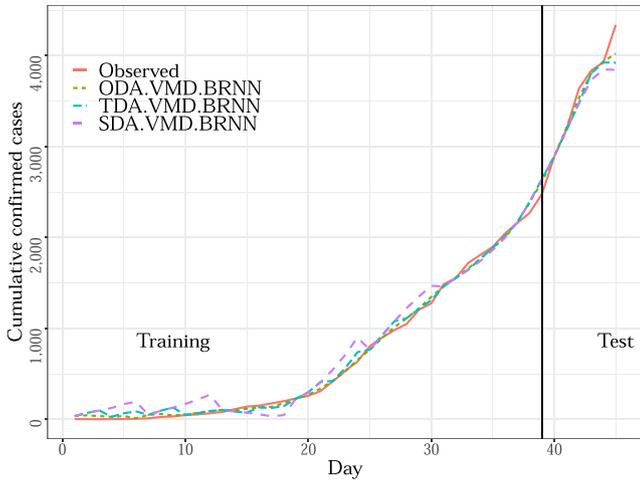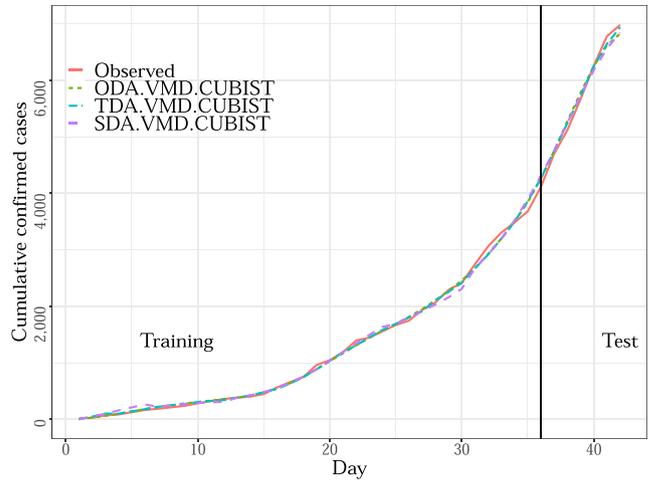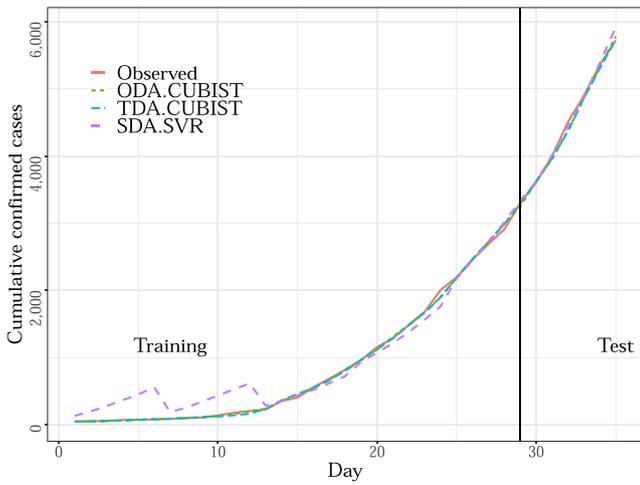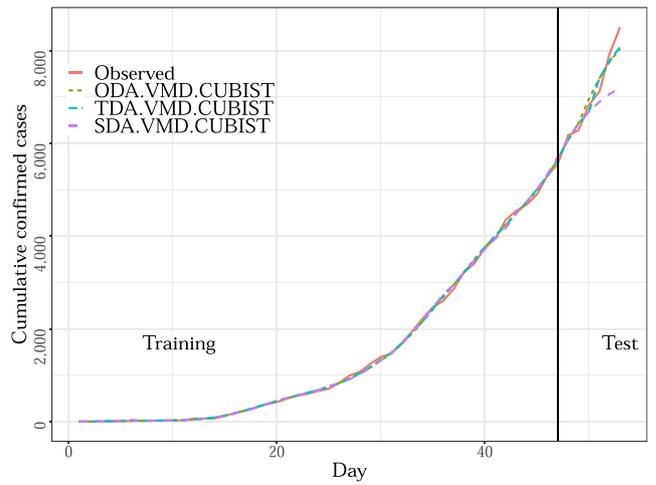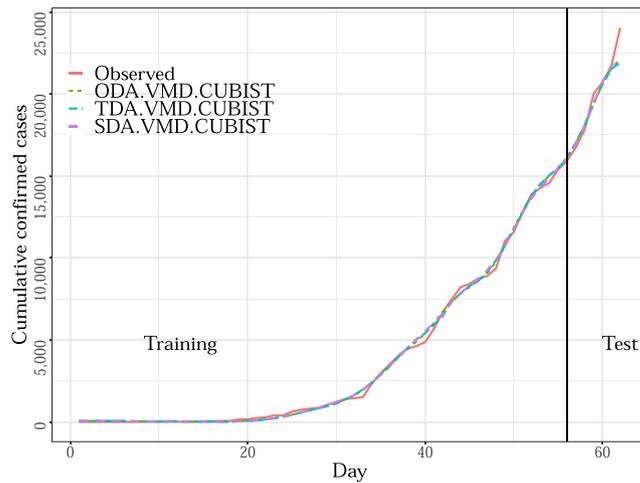
Fig. 3. Prediction versus observed COVID-19 cases for Brazilian States.
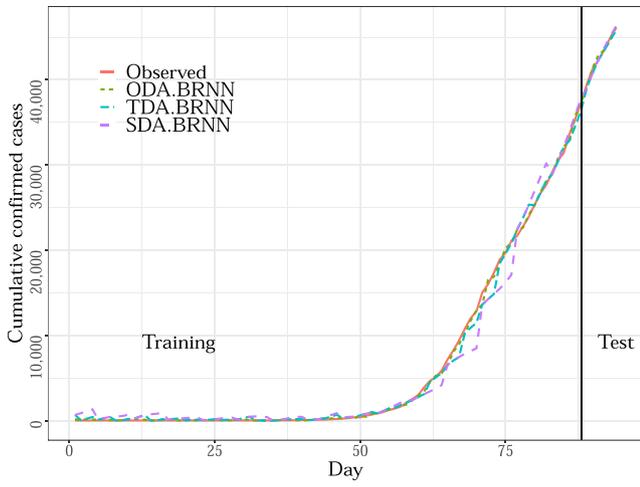
(a) CA
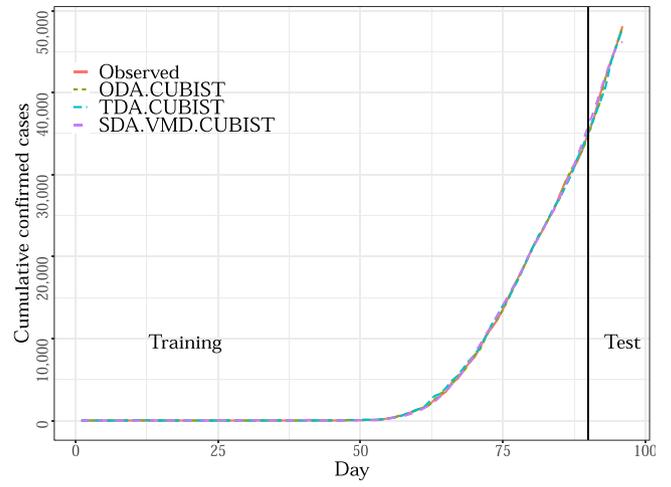
(b) IL
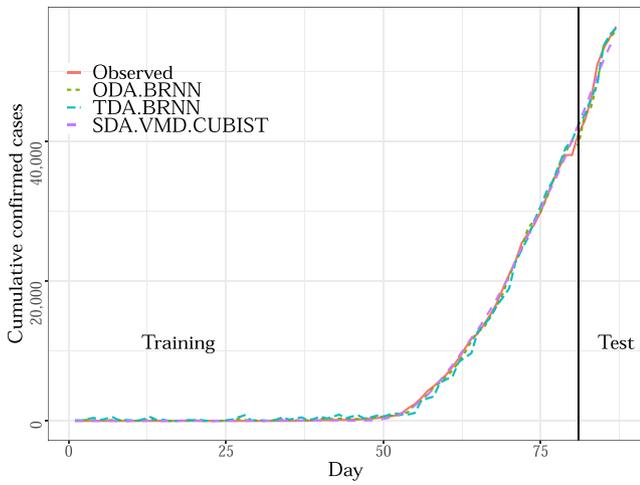
(c) MA
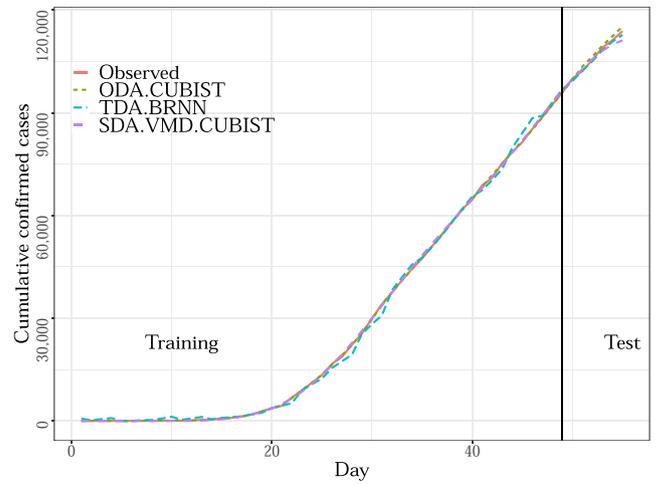
(d) NJ
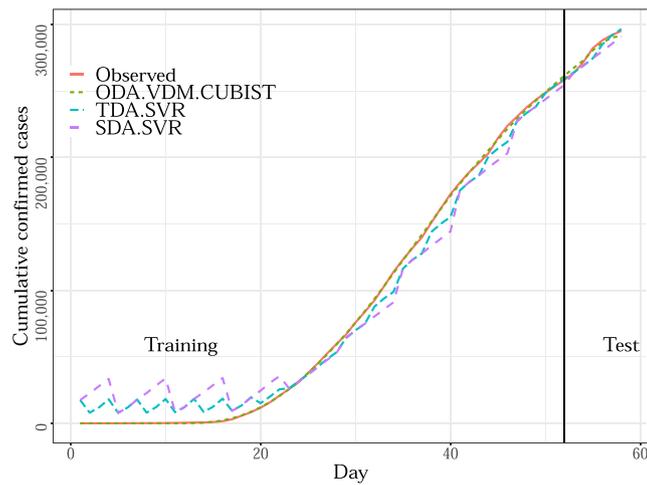
(e) NY

Fig. 4. Prediction versus observed COVID-19 cases for American States.
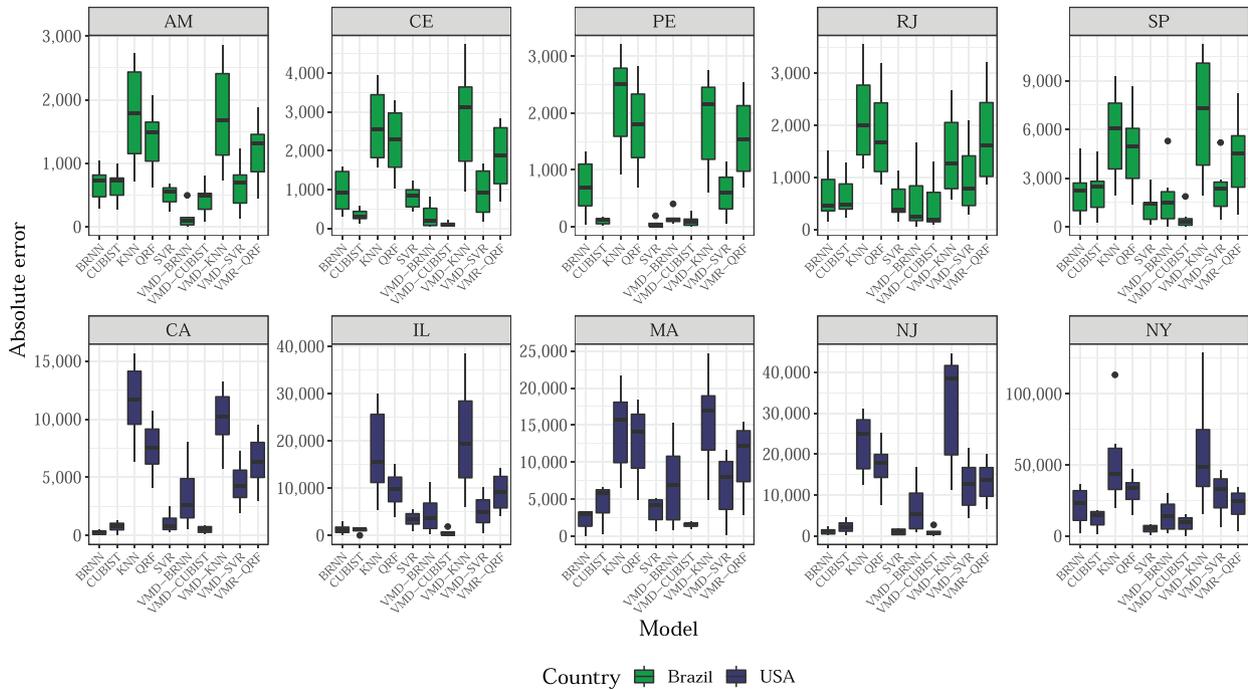
Fig. 5. Box-plot for absolute error according to model and state for COVID-19 forecasting for SDA.
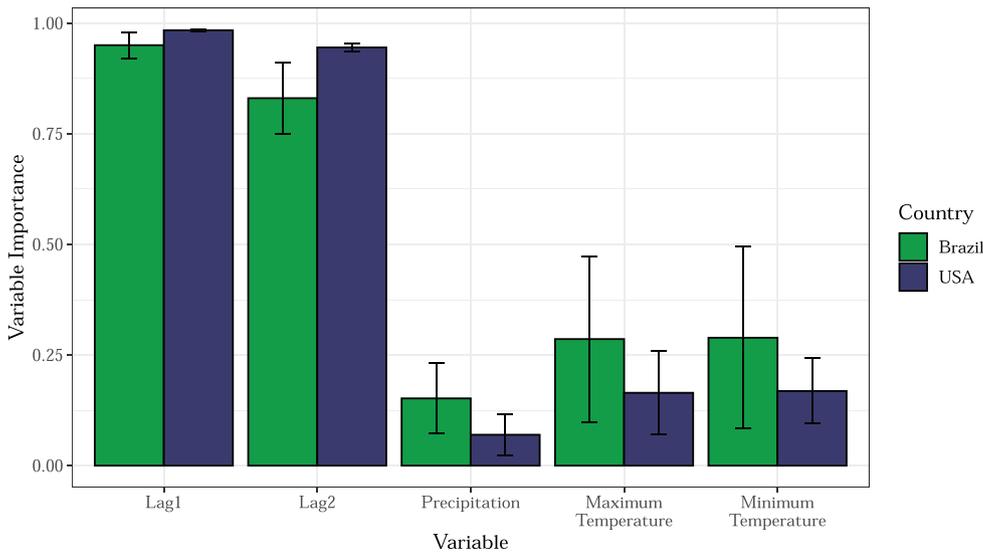
Fig. 6. Variable importance for Brazil and USA.

Firstly, considering the results for the Brazil context, the main results are highlighted as follows.

- AM: In this state, VMD–BRNN could be considered to forecasting COVID-19 cases, once the model outperformed all the single and VMD models in both performance criteria in all forecasting horizons. The improvement in the sMAPE achieved by VMD–BRNN ranges between 39.47% - 96-06%, 55.97% - 94.88%, and 67.41% - 94.25%, for ODA, TDA, and SDA horizon respectively. Regarding RRMSE analysis, the improvement ranges between 9.86% - 94.81%, 33.44% - 93.29%, and 56.66% - 93.89%, respectively.
- CE, RJ, and SP: For these states, in all forecasting horizons, the VMD–CUBIST approach achieved better accuracy than other models, for both sMAPE and RRMSE criteria in the multi-days-ahead forecasting task of the confirmed number of COVID-19.

In fact, the improvement in sMAPE is ranged in 8.67% - 96.57%, 12.15% - 97.78%, and 59.37% - 97.09%, respectively, in ODA, TDA, and SDA forecasting horizons. Moreover, the improvement in RRMSE is ranged in 12.41% - 97.32%, 2.61% - 98.29%, and 49.99% - 97.95%, respectively.

- PE: In this state, CUBIST and SVR present better performance to forecasting COVID-19 cases. For ODA and TDA, CUBIST outperforms models, while for SDA the SVR achieves better accuracy regarding sMAPE and RRMSE than others. The improvement in the sMAPE for ODA and TDA achieved by CUBIST ranges between 6.81% - 97.93%, and 24.94% - 98.23%, respectively. For SDA, SVR outperforms other models, and this criterion is reduced in the range of 49.36% - 98.27%. Moreover, the same behavior is observed when the improvement in the RRMSE criterion is obtained.

Remark: In this experiment, regarding the Brazilian states, 150 scenarios (5 datasets, 3 forecasting horizons, and 10 models) were evaluated for the task of forecasting cumulative COVID-19 cases. In an overview, the best models for each state, obtained sMAPE ranged between 1.14% - 3.05%, 1.06% - 2.79%, and 1.05% - 3.03% for ODA, TDA, and SDA forecasting, respectively. In the Brazilian context, the ranking of the model in all scenarios is VMD–CUBIST, VMD–BRNN, SVR, CUBIST, VMD–SVR, BRNN, VMD–QRF, QRF, VMD–KNN, and KNN. From a broader perspective, the efficiency of the VMD models is due to the capability of the approach to deal with non-linearity and non-stationarity of the data. Moreover, the efficiency of the CUBIST is due mainly to its ensemble learning of rules, in which the approach takes advantage of each rule based on the input set. On the other hand, the difficulty of the KNN model to forecasting cumulative COVID-19 cases could be attributed to the fact that this approach requires more observations to effectively learn the data pattern, once the forecasting is obtained by an average of past similar values.

In the next, considering the results for the USA context, the main results are highlighted as follows.

- CA: In CA state, BRNN outperformed other models, in all forecasting horizons, for both sMAPE and RRMSE criteria. In this aspect, the improvement in sMAPE ranges between 29.98% - 97.86%, 4.64% - 97.71%, and 48.56% - 97.99%, for ODA, TDA, and SDA, respectively. Regarding RRMSE, the improvement ranges in 24.00% - 97.67%, 6.57% - 97.78%, and 48.62% - 98.11%, respectively.

- IL, MA, and NJ: For both performance criteria, CUBIST outperformed other models in ODA, for IL and NJ states, and TDA, for IL. BRNN presented better accuracy than other models, for MA state in ODA and TDA. Moreover, VMD–CUBIST outperformed other models in SDA for these three states. In fact, the improvement in sMAPE is ranged in 6.63% - 98.76%, 31.89% - 98.09%, and 3.76% - 97.98%, respectively, in ODA, TDA, and SDA forecasting horizons. Moreover, regarding the RRMSE, the improvement ranges between 7.54% - 98.48%, 0.83% - 98.25%, and 3.25% - 98.11%, respectively.

- NY: For NY state, in both performance criteria, VMD–CUBIST presented better accuracy than other model in ODA forecasting, while SVR outperformed the other models in TDA and SDA forecasting. Regarding sMAPE, the improvement ranges 17.86% - 95.44%, 16.12% - 95.69%, and 42.39% - 92.71%, for ODA, SDA, and TDA, respectively. For RRMSE, the improvement ranges 25.78% - 96.09%, 7.78% - 95.43%, and 43.76% - 93.45%, respectively.

Remark: In this experiment, regarding the American states, 150 scenarios (5 datasets, 3 forecasting horizons, and 10 models) were evaluated for the task of forecasting cumulative COVID-19 cases. In an overview, the best models for each state, obtained sMAPE ranged between 0.54% - 1.90%, 0.55% - 1.59%, and 0.62% - 3.08% for ODA, TDA, and SDA forecasting, respectively. In the American context, the ranking of the models in all scenarios is VMD–CUBIST, BRNN, CUBIST, SVR, VMD–BRNN, VMD–SVR, VMD–QRF, QRF, KNN, and VMD–KNN. The same behavior presented in Brazilian cases is presented in the American, which the VMD–CUBIST in overall had better average performance compared to the other models.

According to the information depicted in Figs. 3 and 4 it is possible to identify that the behavior of the data is learned by the evaluated models, which can forecasting compatible cases with the observed values. In most states, the good performance presented in the training stage persists in the test phase. In Figs. 3a, 3c, 4a, and 4e the models presented some difficulties to capture the behavior of the data in the training stage, however in test phase the models could perform accurately presenting low errors.

Furthermore, Fig. 5 presents the box-plots of test set forecasting errors in the SDA horizon for each model and each state.

Due to the recursive strategy adopted, the SDA horizon was chosen to the analysis, once the errors tend to grow as the forecast horizon increases. The box diagram depicts the variation of absolute errors for each model, which reflects the stability of each model. In this context, the dots out of boxes are considered outliers errors.

Analyzing the box-plot, models with lower variation in the errors are indicated by the boxes with a smaller size. Fig. 5 corroborates the results presented in Tables A.1 and A.2. Models with lower errors achieve better stability, which means that the most appropriate model for each state can maintain a learning pattern, obtaining homogeneous forecasting errors.

The variable importance is an overall quantification of the relationship between the predictor variables (inputs) and the predicted value. Finally, Fig. 6 is presented the variable importance of each input used to fit and train the models. As expected, the lag inputs present high importance due to their high correlation to the output. However, it is important to notice that climate data indeed presented some influence in predicting COVID-19 cumulative cases, especially in the Brazilian context, that the variance of the Temperature data reaches up to 50% of importance. In other words, the climatic exogenous inputs are in some level relevant to the prediction of cumulative cases of COVID-19 in both Brazil's and USA's context for the five evaluated states.

5. Conclusion and future research

In this paper, machine learning approaches named BRNN, CUBIST, KNN, QRF, and SVR, as well as VMD approach, were employed in the task of forecasting one, three, and six-days-ahead the COVID-19 cumulative confirmed cases in five Brazilian states and five American states with a high daily incidence. The COVID-19 cumulative confirmed cases for AM, CE, PE, RJ, and SP states, as well as CA, IL, MA, NJ, and NY were used. The IP, sMAPE and RRMSE criteria were adopted to evaluate the performance of the compared approaches. The stability of out-of-sample errors was evaluated through box-plots. Further, the variable importance of the lag and climatic exogenous inputs were analyzed.

In respect of obtained results, it is possible to infer that CUBIST coupled with the VMD model are suitable tools to forecast COVID-19 cases for most of the adopted states, once that these approaches were able to learn the non-linearities inherent to the evaluated epidemiological time series. Also, BRNN and SVR models deserve attention for the development of this task as well. Therefore, the ranking of models in all scenarios for Brazilian states is VMD–CUBIST, VMD–BRNN, SVR, CUBIST, VMD–SVR, BRNN, VMD–QRF, QRF, VMD–KNN, and KNN, and for USA states is VMD–CUBIST, BRNN, CUBIST, SVR, VMD–BRNN, VMD–SVR, VMD–QRF, QRF, KNN, and VMD–KNN. Also, looking for COVID-19 forecasts six-days-ahead, hybrid models are more suitable tools than non-decomposed models. Further, it was observed that climatic variables, such as temperature and precipitation indeed influence increasing the accuracy when predicting COVID-19 cases, wherein some cases climate inputs reached up to 50% of importance in the forecasting model.

For future works, it is intended to adopt (i) deep learning approaches, (ii) different decomposition approaches, (iii) multi-objective optimization to tune hyperparameters of forecasting models, and (iv) more climatic data and demographic features.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Ramon Gomes da Silva: Conceptualization, Methodology, Formal analysis, Validation, Writing - original draft, Writing - review & editing. **Matheus Henrique Dal Molin Ribeiro:** Conceptualization, Methodology, Formal analysis, Validation, Writing - original draft, Writing - review & editing. **Viviana Cocco Mariani:** Conceptualization, Writing - review & editing. **Leandro dos Santos Coelho:** Conceptualization, Writing - review & editing.

Acknowledgments

Appendix A. Performance Measures

Tables A.1 and A.2 present the performance measures for each model in each state and forecasting horizon.

Appendix B. Hyperparameters

Tables B.1 and B.2 present the hyperparameters obtained by grid-search for the models employed in this paper.

Table A.1
Performance measures for each evaluated model for Brazilian states.

Country	State	Forecasting Horizon	Criterion	Model									
				BRNN	CUBIST	KNN	QRF	SVR	VMD–BRNN	VMD–CUBIST	VMD–KNN	VMD–QRF	VMD–SVR
Brazil	AM	ODA	sMAPE	11.59%	5.34%	50.92%	45.99%	5.40%	**2.00%**	3.31%	38.17%	20.32%	4.30%
			RRMSE	12.96%	5.88%	73.92%	65.17%	5.59%	**3.83%**	4.25%	50.10%	23.55%	5.11%
		TDA	sMAPE	14.54%	8.86%	52.37%	45.99%	7.65%	**2.68%**	6.09%	51.73%	24.93%	9.58%
			RRMSE	17.00%	11.05%	77.14%	65.17%	9.20%	**5.18%**	7.77%	75.87%	30.25%	11.99%
		SDA	sMAPE	19.62%	19.15%	63.80%	45.99%	14.48%	**3.67%**	11.26%	63.53%	38.35%	18.68%
			RRMSE	24.11%	23.31%	102.85%	65.17%	16.62%	**6.29%**	14.50%	102.53%	52.68%	24.83%
	CE	ODA	sMAPE	7.97%	2.70%	53.33%	45.42%	5.49%	2.55%	**1.83%**	40.75%	17.82%	3.26%
			RRMSE	9.48%	3.89%	78.05%	64.86%	5.52%	3.52%	**2.09%**	56.19%	20.56%	3.85%
		TDA	sMAPE	13.17%	4.24%	55.23%	45.42%	8.46%	3.46%	**1.23%**	52.98%	22.64%	8.03%
			RRMSE	16.46%	4.26%	82.59%	64.86%	8.85%	5.35%	**1.41%**	79.73%	26.70%	9.57%
		SDA	sMAPE	16.69%	5.91%	56.75%	45.42%	14.36%	5.08%	**1.82%**	62.54%	35.71%	16.20%
			RRMSE	21.78%	6.62%	86.22%	64.86%	16.85%	7.71%	**2.08%**	101.64%	49.40%	22.04%
	PE	ODA	sMAPE	10.43%	**1.14%**	54.97%	45.27%	1.22%	2.54%	2.09%	37.56%	26.69%	1.96%
			RRMSE	13.54%	**1.36%**	83.80%	66.03%	1.86%	3.32%	2.39%	51.61%	32.54%	2.65%
		TDA	sMAPE	12.21%	**1.06%**	60.19%	45.27%	1.42%	3.30%	1.78%	44.05%	30.83%	6.11%
			RRMSE	16.57%	**1.39%**	94.20%	66.03%	2.33%	4.37%	2.11%	62.59%	40.21%	7.94%
		SDA	sMAPE	15.32%	2.08%	61.09%	45.27%	**1.05%**	3.28%	2.16%	48.27%	39.07%	12.76%
			RRMSE	21.34%	2.53%	96.22%	66.03%	**1.81%**	4.15%	3.08%	71.19%	55.22%	17.24%
	RJ	ODA	sMAPE	4.87%	3.65%	34.70%	28.70%	3.54%	3.61%	**3.05%**	16.88%	15.87%	3.80%
			RRMSE	5.82%	4.37%	46.38%	38.06%	4.62%	5.32%	**3.60%**	21.62%	19.64%	5.24%
		TDA	sMAPE	6.68%	5.62%	34.95%	28.70%	5.34%	5.58%	**2.73%**	18.69%	20.51%	7.67%
			RRMSE	9.23%	7.00%	46.81%	38.06%	6.72%	9.02%	**3.30%**	23.93%	25.51%	9.88%
		SDA	sMAPE	9.38%	8.95%	34.95%	28.70%	7.57%	7.67%	**3.02%**	21.92%	28.10%	14.24%
			RRMSE	12.73%	11.40%	46.81%	38.06%	9.83%	12.35%	**3.95%**	28.99%	37.48%	19.27%
	SP	ODA	sMAPE	4.70%	3.58%	29.85%	26.12%	2.81%	3.73%	**2.57%**	32.58%	20.39%	3.91%
			RRMSE	6.81%	4.82%	39.71%	34.83%	4.22%	5.85%	**4.22%**	43.82%	27.35%	5.06%
		TDA	sMAPE	6.44%	4.56%	29.85%	26.12%	3.17%	6.14%	**2.79%**	37.21%	21.95%	6.65%
			RRMSE	9.62%	6.45%	39.71%	34.83%	4.74%	10.04%	**4.62%**	52.39%	29.47%	8.73%
		SDA	sMAPE	10.67%	11.29%	32.05%	26.12%	5.95%	8.71%	**2.42%**	40.67%	23.02%	11.82%
			RRMSE	14.65%	14.92%	43.16%	34.83%	8.19%	13.68%	**4.10%**	58.44%	31.18%	15.78%

Table A.2
Performance measures for each evaluated model for American states.

Country	State	Forecasting Horizon	Criterion	Model									
				BRNN	CUBIST	KNN	QRF	SVR	VMD–BRNN	VMD–CUBIST	VMD–KNN	VMD–QRF	VMD–SVR
USA	CA	ODA	sMAPE	**0.56%**	0.90%	26.37%	19.05%	1.80%	2.19%	0.80%	24.23%	11.44%	3.60%
			RRMSE	**0.73%**	0.96%	31.31%	22.12%	1.96%	2.63%	1.07%	28.42%	12.87%	3.76%
		TDA	sMAPE	**0.66%**	1.10%	28.87%	19.05%	2.40%	4.66%	0.69%	25.63%	13.07%	6.11%
			RRMSE	**0.78%**	1.22%	35.11%	22.12%	2.97%	6.08%	0.84%	30.44%	14.62%	6.74%
		SDA	sMAPE	**0.62%**	1.87%	30.85%	19.05%	2.41%	8.20%	1.21%	26.23%	15.89%	10.75%
			RRMSE	**0.72%**	2.18%	38.04%	22.12%	3.01%	10.86%	1.40%	31.31%	18.44%	12.38%
	IL	ODA	sMAPE	1.07%	**0.54%**	34.26%	25.04%	3.05%	3.16%	1.83%	43.71%	17.74%	3.80%
			RRMSE	1.24%	**0.92%**	44.21%	31.31%	3.13%	4.43%	2.63%	60.63%	21.50%	4.49%
		TDA	sMAPE	1.84%	**1.05%**	36.63%	25.04%	4.73%	6.33%	1.55%	49.22%	19.41%	6.98%
			RRMSE	2.39%	**1.51%**	47.86%	31.31%	5.18%	8.95%	2.07%	70.55%	23.08%	8.53%
		SDA	sMAPE	2.89%	2.58%	52.64%	25.04%	8.10%	11.03%	**1.42%**	67.54%	23.61%	12.43%
			RRMSE	3.78%	3.00%	78.46%	31.31%	9.53%	15.66%	**2.04%**	107.77%	29.59%	16.27%
	MA	ODA	sMAPE	**1.90%**	2.45%	30.79%	28.07%	3.01%	4.51%	2.65%	26.67%	17.49%	3.91%
			RRMSE	**2.42%**	3.46%	39.30%	35.70%	3.59%	5.39%	2.69%	33.79%	21.69%	4.77%
		TDA	sMAPE	**1.59%**	3.41%	31.52%	28.07%	3.35%	8.25%	2.34%	31.47%	19.83%	7.12%
			RRMSE	**2.39%**	5.34%	40.65%	35.70%	4.58%	10.86%	2.41%	40.87%	24.21%	8.82%
		SDA	sMAPE	4.38%	8.98%	32.54%	28.07%	6.92%	14.65%	**3.08%**	35.52%	22.66%	13.85%
			RRMSE	5.33%	10.98%	42.30%	35.70%	8.18%	20.17%	**3.16%**	47.50%	28.56%	18.13%
	NJ	ODA	sMAPE	0.97%	**0.88%**	21.41%	17.87%	0.94%	1.63%	0.99%	25.40%	9.88%	3.78%
			RRMSE	1.01%	**0.93%**	24.93%	20.38%	1.03%	2.18%	1.32%	30.87%	10.99%	3.92%
		TDA	sMAPE	**0.55%**	1.25%	22.03%	18.20%	1.09%	3.52%	0.83%	28.62%	11.17%	6.75%
			RRMSE	**0.62%**	1.48%	25.81%	20.82%	1.17%	4.74%	1.21%	35.63%	12.36%	7.47%
		SDA	sMAPE	1.02%	1.99%	23.76%	18.20%	0.94%	6.54%	**0.91%**	35.12%	13.20%	12.35%
			RRMSE	1.27%	2.40%	28.32%	20.82%	1.14%	8.95%	**1.23%**	45.47%	15.09%	14.70%
	NY	ODA	sMAPE	3.26%	1.26%	16.81%	12.09%	1.02%	1.46%	**0.84%**	18.44%	5.51%	3.27%
			RRMSE	3.63%	1.43%	19.88%	13.64%	1.23%	1.64%	**0.92%**	23.44%	6.30%	3.38%
		TDA	sMAPE	5.40%	2.20%	18.23%	12.09%	**0.92%**	2.76%	1.10%	21.34%	6.39%	5.85%
			RRMSE	6.43%	2.69%	22.36%	13.64%	**1.27%**	3.31%	1.38%	27.91%	7.28%	6.62%
		SDA	sMAPE	7.64%	4.54%	21.02%	12.09%	**1.75%**	5.23%	3.05%	24.06%	7.96%	10.97%
			RRMSE	9.37%	5.30%	26.69%	13.64%	**2.05%**	6.65%	3.65%	31.38%	9.32%	12.93%

Table B.1
Hyperparameters selected by grid-search for each evaluated model for Brazilian states.

Country	State	Component	BRNN	CUBIST		KNN	QRF	SVR
			no. of Neurons	no. of Committees	no. of Instances	no. of Neighbors	no. of Randomly Selected Predictors	Cost
Brazil	AM	IMF$_1$	4	1	0	9	5	1
		IMF$_2$	5	20	5	5	5	1
		IMF$_3$	3	20	0	5	5	1
		IMF$_4$	5	10	0	7	5	1
		IMF$_5$	4	1	5	5	5	1
		Non-decomposed	3	1	5	5	4	1
	CE	IMF$_1$	3	1	5	13	2	1
		IMF$_2$	5	20	5	5	5	1
		IMF$_3$	5	10	9	13	5	1
		IMF$_4$	5	10	9	5	5	1
		IMF$_5$	5	10	0	5	5	1
		Non-decomposed	1	1	9	5	4	1
	PE	IMF$_1$	2	10	0	13	3	1
		IMF$_2$	5	20	5	5	3	1
		IMF$_3$	1	20	5	13	3	1
		IMF$_4$	5	1	9	5	3	1
		IMF$_5$	5	10	9	11	5	1
		Non-decomposed	5	10	0	5	4	1
	RJ	IMF$_1$	3	10	0	5	5	1
		IMF$_2$	1	1	0	5	4	1
		IMF$_3$	4	20	5	5	5	1
		IMF$_4$	5	1	5	5	5	1
		IMF$_5$	5	1	5	5	5	1
		Non-decomposed	1	20	5	5	4	1
	SP	IMF$_1$	2	10	0	11	4	1
		IMF$_2$	5	1	9	5	4	1
		IMF$_3$	5	10	5	13	4	1
		IMF$_4$	5	20	0	5	5	1
		IMF$_5$	1	10	5	5	5	1
		Non-decomposed	1	10	0	5	5	1

Table B.2
Hyperparameters selected by grid-search for each evaluated model for American states.

Country	State	Component	BRNN	CUBIST		KNN	QRF	SVR
			no. of Neurons	no. of Committees	no. of Instances	no. of Neighbors	no. of Randomly Selected Predictors	Cost
USA	CA	IMF_1	1	1	9	5	4	1
		IMF_2	1	1	0	5	4	1
		IMF_3	4	1	9	11	3	1
		IMF_4	5	20	0	9	5	1
		IMF_5	5	1	5	5	5	1
		Non-decomposed	1	20	5	5	4	1
	IL	IMF_1	5	20	5	5	5	1
		IMF_2	1	20	5	5	4	1
		IMF_3	5	20	5	5	3	1
		IMF_4	5	20	9	5	5	1
		IMF_5	5	10	0	5	5	1
		Non-decomposed	1	20	5	5	4	1
	MA	IMF_1	3	1	5	5	5	1
		IMF_2	1	20	5	5	4	1
		IMF_3	4	20	5	5	4	1
		IMF_4	4	20	5	13	5	1
		IMF_5	5	1	0	5	5	1
		Non-decomposed	1	20	5	5	5	1
	NJ	IMF_1	1	10	0	5	5	1
		IMF_2	5	20	5	5	4	1
		IMF_3	5	10	9	5	4	1
		IMF_4	4	10	9	13	5	1
		IMF_5	5	1	0	5	5	1
		Non-decomposed	1	20	5	5	5	1
	NY	IMF_1	1	1	0	13	5	1
		IMF_2	5	20	5	5	5	1
		IMF_3	5	1	0	5	5	1
		IMF_4	3	10	5	9	5	1
		IMF_5	5	1	5	5	4	1
		Non-decomposed	5	20	0	5	4	1

References

[1] World Health Organization (WHO). Coronavirus (COVID-19). 2020 (accessed in 24 June, 2020); https://www.who.int/health-topics/coronavirus#tab=tab_1.

[2] Bansal M. Cardiovascular disease and COVID-19. Diabet Metab Syndrome 2020;14(3):247–50. doi:10.1016/j.dsx.2020.03.013.

[3] Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R. Severe acute respiratory syndrome coronavirus 2 (SARS-cov-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges. Int J Antimicrob Agents 2020;55(3):105924. doi:10.1016/j.ijantimicag.2020.105924.

[4] Hussain A, Bhowmik B, do Vale Moreira NC. COVID-19 And diabetes: knowledge in progress. Diabetes Res Clin Pract 2020;162:108142. doi:10.1016/j.diabres.2020.108142.

[5] Moujaess E, Kourie HR, Ghosn M. Cancer patients and research during COVID-19 pandemic: a systematic review of current evidence. Crit Rev Oncol Hematol 2020;150:102972. doi:10.1016/j.critrevonc.2020.102972.

[6] Abbas AM, Fathy SK, Fawzy AT, Salem AS, Shawky MS. The mutual effects of COVID-19 and obesity. Obesity Medicine 2020;19:100250. doi:10.1016/j.obmed.2020.100250.

[7] Su H, Yang M, Wan C, Yi L-X, Tang F, Zhu H-Y, et al. Renal histopathological analysis of 26 postmortem findings of patients with COVID-19 in China. Kidney Int 2020. doi:10.1016/j.kint.2020.04.003.

[8] Guan W-j, Ni Z-y, Hu Y, Liang W-h, Ou C-q, He J-x, et al. Clinical characteristics of coronavirus disease 2019 in China. N top N Engl J Med 2020;382(18):1708–20. doi:10.1056/NEJMoa2002032.

[9] Requia WJ, Kondo EK, Adams MD, Gold DR, Struchiner CJ. Risk of the Brazilian health care system over 5572 municipalities to exceed health care capacity due to the 2019 novel coronavirus (COVID-19). Sci Total Environ 2020;730:139144. doi:10.1016/j.scitotenv.2020.139144.

[10] Ndarou F, Area I, Nieto JJ, Torres DF. Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. Chaos Soliton Fract 2020;135:109846. doi:10.1016/j.chaos.2020.109846.

[11] Barmparis G, Tsironis G. Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach. Chaos Soliton Fract 2020;135:109842. doi:10.1016/j.chaos.2020.109842.

[12] Zhang X, Ma R, Wang L. Predicting turning point, duration and attack rate of COVID-19 outbreaks in major western countries. Chaos Soliton Fract 2020;135:109829. doi:10.1016/j.chaos.2020.109829.

[13] Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ 2020;729:138817. doi:10.1016/j.scitotenv.2020.138817.

[14] Ahmar AS, del Val EB. SutteARIMA: short-term forecasting method, a case COVID-19 and stock market in Spain. Sci Total Environ 2020;729:138883. doi:10.1016/j.scitotenv.2020.138883.

[15] Ribeiro MHDM, da Silva RG, Mariani VC, Coelho LdS. Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. Chaos Soliton Fract 2020;135:109853. doi:10.1016/j.chaos.2020.109853.

[16] Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Soliton Fract 2020;135:109864. doi:10.1016/j.chaos.2020.109864.

[17] Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. Chaos Soliton Fract 2020;135:109850. doi:10.1016/j.chaos.2020.109850.

[18] Singh S, Parmar KS, Kumar J, Makkhan SJS. Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19. Chaos Soliton Fract 2020;135:109866. doi:10.1016/j.chaos.2020.109866.

[19] Ribeiro MHDM, da Silva RG, Fraccanabbia N, Mariani VC, Coelho LdS. Forecasting epidemiological time series based on decomposition and optimization approaches. In: 14th Brazilian computational intelligence meeting (CBIC); 2019. p. 1–8. Belém, Brazil

[20] Dragomiretskiy K, Zosso D. Variational mode decomposition. IEEE Trans Signal Process 2014;62(3):531–44.

[21] Moreno SR, da Silva RG, Mariani VC, Coelho LdS. Multi-step wind speed forecasting based on hybrid multi-stage decomposition model and long short-term memory neural network. Energy Convers Manage 2020;213:112869. doi:10.1016/j.enconman.2020.112869.

[22] Wu Q, Lin H. Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network. Sustain Cities Soc 2019;50:101657. doi:10.1016/j.scs.2019.101657.

[23] Li J, Zhu S, Wu Q. Monthly crude oil spot price forecasting using variational mode decomposition. Energy Econ 2019;83:240–53. doi:10.1016/j.eneco.2019.07.009.

[24] Prata DN, Rodrigues W, Bermejo PH. Temperature significantly changes COVID-19 transmission in (sub)tropical cities of Brazil. Sci Total Environ 2020;729:138862. doi:10.1016/j.scitotenv.2020.138862.

[25] Coccia M. Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. Sci Total Environ 2020;729:138474. doi:10.1016/j.scitotenv.2020.138474.

[26] Shi P, Dong Y, Yan H, Zhao C, Li X, Liu W, et al. Impact of temperature on the dynamics of the COVID-19 outbreak in China. Sci Total Environ 2020;728:138890. doi:10.1016/j.scitotenv.2020.138890.

[27] Wu Y, Jing W, Liu J, Ma Q, Yuan J, Wang Y, et al. Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. Sci Total Environ 2020;729:139051. doi:10.1016/j.scitotenv.2020.139051.

[28] Ahmadi M, Sharifi A, Dorosti S, Jafarzadeh Ghoushchi S, Ghanbari N. Investigation of effective climatology parameters on COVID-19 outbreak in Iran. Sci Total Environ 2020;729:138705. doi:10.1016/j.scitotenv.2020.138705.

[29] Sobral MFF, Duarte GB, da Penha Sobral AIG, Marinho MLM, de Souza Melo A. Association between climate variables and global transmission of SARS-cov-2. Sci Total Environ 2020;729:138997. doi:10.1016/j.scitotenv.2020.138997.

[30] Auler A, Cssaro F, da Silva V, Pires L. Evidence that high temperatures and intermediate relative humidity might favor the spread of COVID-19 in tropical climate: acase study for the most affected Brazilian cities. Sci Total Environ 2020;729:139090. doi:10.1016/j.scitotenv.2020.139090.

[31] Bashir MF, Ma B, Bilal, Komal B, Bashir MA, Tan D, et al. Correlation between climate indicators and COVID-19 pandemic in New York, USA. Sci Total Environ 2020;728:138835. doi:10.1016/j.scitotenv.2020.138835.

[32] Ribeiro MHDM, Coelho LdS. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Appl Soft Comput 2020;86(105837). doi:10.1016/j.asoc.2019.105837.

[33] Ribeiro VHA, Reynoso-Meza G. Multi-objective support vector machines ensemble generation for water quality monitoring. In: IEEE Congress on Evolutionary Computation (CEC). Rio de Janeiro, Brazil: IEEE; 2018. p. 1–6.

[34] Fernndez-Delgado M, Sirsat M, Cernadas E, Alawadi S, Barro S, Febrero-Bande M. An extensive experimental survey of regression methods. Neural Networks 2019;111:11–34. doi:10.1016/j.neunet.2018.12.010.

[35] Zuo G, Luo J, Wang N, Lian Y, He X. Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting. J Hydrol (Amst) 2020;585:124776. doi:10.1016/j.jhydrol.2020.124776.

[36] Zhu Q, Zhang F, Liu S, Wu Y, Wang L. A hybrid VMD biGRU model for rubber futures time series forecasting. Appl Soft Comput 2019;84:105739. doi:10.1016/j.asoc.2019.105739.

[37] Justen A.. COVID-19: Coronavirus newsletters and cases by municipality per day. 2020. (accessed in 28 April, 2020).; https://brasil.io/api/dataset/covid19/caso/data/?place_type=state.

[38] Center for Systems Science and Engineering (CSSE). Novel coronavirus (COVID-19) cases, provided by JHU CSSE. 2020. (accessed in 28 April, 2020); https://github.com/CSSEGISandData/COVID-19.

[39] Brazil. Instituto Nacional de Meteorologia (INMET), Ministrio da Agricultura, Pecuria e Abastecimento. 2020. (accessed in 28 April, 2020), (in Portuguese); http://www.inmet.gov.br/portal/index.php?r=estacoes/estacoesAutomaticas.

[40] U.S.. National Oceanic and Atmospheric Administration (NOAA): National Centers for Environmental Information. 2020. (accessed in 28 April, 2020); https://www.ncdc.noaa.gov/.

[41] Chamberlain S.. rnoaa: 'NOAA' weather data from R. 2020. R package version 0.9.6; https://CRAN.R-project.org/package=rnoaa.

[42] MacKay DJC. Bayesian interpolation. Neural Comput 1992;4(3):415–47. doi:10.1162/neco.1992.4.3.415.

[43] Nguyen D, Widrow B. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In: IJCNN International Joint Conference on Neural Networks, 3; 1990. p. 21–6. San Diego, USA

[44] Quinlan JR. Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence. Hobart, Tasmania: World Scientific; 1992. p. 343–8.

[45] Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. Mach Learn 1991;6(1):37–66. doi:10.1007/BF00153759.

[46] Breiman L. Random forests. Mach Learn 2001;45(1):5–32. doi:10.1023/A:1010933404324.

[47] Meinshausen N. Quantile regression forests. J Mach Learn Res 2006;7:983–99.

[48] Vaysse K, Lagacherie P. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 2017;291:55–64. doi:10.1016/j.geoderma.2016.12.017.

[49] Drucker H, Burges CJC, Kaufman L, Smola AJ, Vapnik V. Support vector regression machines. In: Mozer MC, Jordan MI, Petsche T, editors. Advances in neural information processing systems 9. MIT Press; 1997. p. 155–61.

[50] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria; 2018.