Dengue Cases Forecasting Based on eXtreme Gradient Boosting Ensemble with Coyote Optimization

Matheus H. D. M. Ribeiro*† ,Ramon G. Silva*, Viviana C. Mariani‡§ and Leandro S. Coelho*§
*Industrial and Systems Engineering Graduate Program (PPGEPS), Pontifical Catholic University of Parana (PUCPR),
Curitiba, Parana, Brazil. 80215-901
†Department of Mathematics, Federal University of Technology - Parana (UTFPR),
Pato Branco, Parana, Brazil. 85503-390
‡Mechanical Engineering Graduate Program (PPGEM), Pontifical Catholic University of Parana (PUCPR),
Curitiba, Parana, Brazil. 80215-901
§Department of Electrical Engineering, Federal University of Parana (UFPR),
Curitiba, Parana, Brazil. 81530-000

Emails: mribeiro@utfpr.edu.br, gomes.ramon@pucpr.edu.br, viviana.mariani@pucpr.br, leandro.coelho@pucpr.br

***Abstract*—Dengue is considered a public health problem in tropical regions, periodically affecting an increasing number of citizens. Consequently, the development of efficient models is essentials to short and long-term forecasting, supporting health care officials to optimally disseminate available resources in the dengue-prone areas. Hybridization of two or more models is a common solution to this problem where one can take advantage of diversity among models to reduce both the bias and variances of the prediction error obtained using single models. Fortunately, the use of ensemble approaches becomes attractive. In this paper, we propose a novel ensemble learning approach combining the eXtreme Gradient Boosting (XGBoost) and Coyote Optimization Algorithm (COA) to capture the non-linearity in a dataset and perform dengue cases forecasting. The performance of the XGBoost model depends upon the appropriate choice of its hyperparameters. In this study, COA has been employed to tune the XGBoost hyperparameters. The proposed hybrid COA-XGBoost model is applied to predicting dengue time-series dataset from Parana, Brazil. Averages of precipitation, temperature, thermal amplitude, relative humidity, and previous dengue cases are considered as input variables as well as dengue cases are used as output variables. The performance of the proposed COA-XGBoost model has been compared with XGBoost when hyperparameters are obtained using other optimization techniques like Differential Evolution, Genetic Algorithm, Cuckoo Search Optimization, Grey Wolf Optimizer, and Firefly Algorithm. The results indicate that the proposed COA–XGBoost can be competitive model when compared to other classical techniques.**

***Keywords*—Ensemble learning, time series forecasting, dengue, metaheuristics.**

I. Introduction

Dengue is considered a public health problem in tropical regions, periodically affecting an increasing number of citizens. Consequently, the development of efficient models is essentials to short and long-term forecasting, supporting health care officials to optimally disseminate available resources in the dengue-prone areas. In 2020 in Brazil, there were 979,764 probable cases reported (incidence rate of 466.2 cases per 100 thousand inhabitants) [1].

Several studies have already been developed aiming to obtain efficient forecasting systems for epidemiological context, especially for dengue outbreak control. Indeed, Guo et al. [2] used an ensemble learning model to forecast weekly dengue incidence and detect outbreak occurrence defined using different cutoffs, during the periods of 2011–2016 in Guangzhou, South of China. Proposed ensemble model provided near-real time estimates of dengue incidence, and captured the fluctuations of dengue dynamics. Liebig et al. [3] proposed an epidemiological model to forecast model the probability of local dengue outbreaks in Queensland, Australia. As result, the authors revealed the airports where dengue infected travelers are most likely to arrive and geographic locations associated with high outbreak probabilities. Abualamah et al. [4] adopted seasonal autoregressive integrated moving average model (SARIMA) to forecast the morbidity and mortality of dengue fever in the Kingdom of Saudi Arabia. The authors observed that the seasonal association of dengue fever during May to September and its relation to air temperature should be communicated to all stakeholders. Mussumeci and Coelho [5] evaluated the predictive performance of Long-Short-Term-Memory (LSTM), random forest regression (RF), and least absolute shrinkage and selection operator (LASSO) model to forecast weekly dengue incidence (four-weeks-ahead) in 790 cities in Brazil using multivariate predictors. The authors argued that LSTM can achieve better performance, in terms of mean prediction errors in quantile scale, than RF and LASSO models.

Notably there is a great attention give to the development of efficient forecasting models in the context of dengue disease. Indeed, a class of models usually employed to achieve high accuracy in the forecasting field is the ensemble learning

method. An ensemble learning approach is a set of combined (weak or base) models that learn different data patterns and thus when results of each model are aggregated an efficient model can be obtained [6]. In this respect, we can highlight the use of the eXtreme gradient boosting (XGBoost) model [7], an ensemble learning model which uses an iterative process for training sequential decision trees, where the objective is to prevent overfitting and optimize the available computational resources. However, to achieve high forecasting performance, the most suitable set of hyperparameters must be defined, and metaheuristics can be used for this purpose, as observed in some fields such as train arrival delay prediction [8], predict a cumulative abnormal return of stocks following earnings release [9], and reservoir production [10].

In this paper, we present an exploratory study is performed to evaluate the viability of using a recently proposed metaheuristic named Coyote Optimization Algorithm (COA) [11] to tune the XGBoost hyperparameters. Then the performance of COA–XGBoost is evaluated in the task of forecasting dengue cases multi-month-ahead (one, two, and three-months-ahead) in Parana (PR) state, Brazil. These two approaches were selected due to having already performed well in time series forecasting [12] and the optimization approach is relatively novel and efficient to global optimization [13] and it is expected to perform optimally.

This paper introduces as the main contribution, the study of the effectiveness of COA–XGBoost compared to well know metaheuristics Differential Evolution (DE), Genetic Algorithm (GA), Firefly Algorithm (FFA), Cuckoo Search Optimization (CSO), and Grey Wolf Optimizer (GWO) to obtain the XGBoost hyperparameters by the root mean squared error (RMSE) minimization. Moreover, the mean absolute percentage error (MAPE), improvement percentage (IP), RMSE, Friedman and Nemenyi hypothesis tests are used to evaluate the compared approaches. This paper represents a contribution to the epidemiological field and time series forecasting combining ensemble learning method and a metaheuristic algorithm, as well as, presenting the importance of some predictive variables on the appearance of dengue cases.

The remainder of this paper is structured as follows: Section II-A presents the data sets adopted in this paper. Section II-B describes the methods employed in this paper. Section III presents the data modeling steps. Section IV shows the results and discussions. Finally, Section V concludes this paper and presents the proposals for future research.

II. Material and Methods

In this section, the data as well as the adopted methods used in this paper are presented.

A. Material

The dataset used in this paper refers to monthly dengue cases registered in Paraná (PR) state, Brazil. It is chosen because there is a high incidence of dengue as well as due to the subtropical and tropical climate. For adopted series, information between the years 2007 and 2017 are available on the Department of Informatics of the Unified Health System (*Departamento de Informática do Sistema Único de Saúde*, DATASUS, in Portuguese) database [14]. In this case, 70% of the data (first 93 observations) is used as a training set and 30% is adopted as a test set (last 39 observations).

In Figures 1 and 2 are presented the the cases of dengue in PR state and the partial autocorrelation function (PACF), respectively.

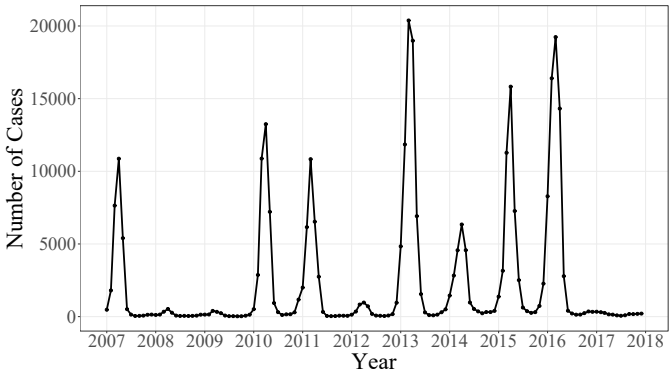

Fig. 1. Dengue cases over the time for PR state.

Exogenous variables such as precipitation, maximum and minimum temperature, humidity, thermal amplitude [15] and previous dengue cases (lagged on one up to three months) are used as inputs of the XGBoost model. Table I shows the statistical indicators for the observed data and climatic features.

TABLE I
SUMMARY OF THE STATISTICAL INDICATORS OF DATASET.

Variable	Dataset	Statistical Indicator				
		# Samples	Minimum	Median	Mean	Maximum
Dengue Cases	Whole	132	23	308	2287	20380
	Training	92	23	292	2061	20380
	Test	40	59	322	2806	19237
Precipitation (mm)	Whole	132	2.92	142.47	147.56	342.20
	Training	92	7.06	128.48	142.09	326.49
	Test	40	2.92	159.76	160.14	342.20
Maximum Temperature (°C)	Whole	132	19.62	26.59	26.10	30.80
	Training	92	20.00	26.28	25.98	30.80
	Test	40	19.62	27.06	26.37	30.58
Minimum Temperature (°C)	Whole	132	8.34	15.41	14.96	20.18
	Training	92	8.34	15.16	14.68	19.82
	Test	40	8.95	15.83	15.60	20.18
Humidity (kg/m^3)	Whole	132	60.93	79.87	78.96	90.06
	Training	92	68.16	80.18	79.23	90.06
	Test	40	60.93	78.77	78.34	86.15

According to the Augmented Dick-Fuller test the dengue cases time series is stationary (DF = -5.35, p-value < 0.05). Aiming at evaluating the presence of seasonality within the data, the Kruskal-Wallis test is performed. In this case, there is evidence of monthly seasonality (χ^2_{11} = 68.04, p-value $>$ 0.05) [16].

B. Methods

1) Extreme Gradient Boosting: The boosting approach proposed by [17] consists of finding an additive model that minimizes a loss function for a weak model (decision tree).

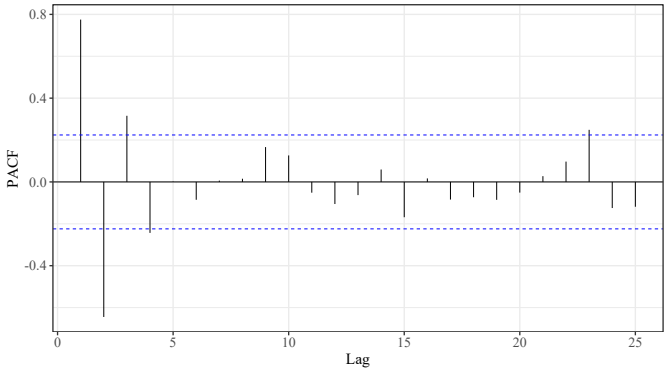

Fig. 2. PACF analysis.

Giving the residual of the previous model, a new model is fitted to minimize the loss function. The current model is added to the previous model, and the procedure continues until convergence criterion is met. The XGBoost is an extension of this approach, which objective is to prevent overfitting and optimize the available computational resources. In this approach through the use of a regularization term added to the loss function, the model's complexity is controlled [7]. The mathematical formulation can be stated as follows:

$$F_{obj}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \Omega(\boldsymbol{\theta}), \tag{1}$$

where $L(\boldsymbol{\theta}) = l\left(\widehat{y_i}, y_i\right)$ and $\Omega(\boldsymbol{\theta}) = \gamma T + \frac{1}{2}\lambda\|w\|^2$, $F_{obj}(\boldsymbol{\theta})$ is the objective function, $\boldsymbol{\theta} = [\gamma, T, w, \lambda, \hat{y}_i]$, $L(\boldsymbol{\theta})$ is the loss function between prediction $\widehat{y_i}$ and real value y_i, $\Omega(\boldsymbol{\theta})$ is the regularization term, γ is the learning rate, T is the number of leaves in the tree, λ is the regularization parameter, and w is the weights of the leaves.

2) Coyote Optimization Algorithm: Every time series has its characteristics, thus hyperparameters that allow for the model to have a good generalization capacity and therefore achieve accurate results in out-of-sample data are desirable. To achieve the promising results in dengue case data, a suitable set of hyperparameters was obtained using the COA. The COA is a population-based approach and it considers the social relations inside the packs of the *Canis latrans* species. The population of coyotes is divided into N_p packs with N_c coyotes each, where the number of coyotes is the same for all packs. This algorithm is classified as both swarm intelligence and evolutionary metaheuristic and it is inspired by the coyotes' behavior [11].

The main steps of COA are described in the sequence, and the mathematical formulation is omitted, which can be found in [11]. The optimization process starts when the global population of coyotes is defined (all candidate solutions). In the sequence, the coyotes' adaptation in the current social condition (set of decision variables, or in this case, current hyperparameters value) is evaluated using the objective function (in this paper the RMSE - described in Subsection III). In it is turning, the alpha coyote of the pack is defined and the social tendency is stated. For each coyote, based on social tendency, the social condition is updated, evaluated and its adaptation verified. Taken into account of biological events of life, the birth and death of coyotes are stated. After this, the transition between packs and coyotes' age (in years) is updated. The process ends when the best coyote (solution) is selected.

C. Hypothesis tests

In this paper, the Friedman is used to verify if at least two of the models represents different results. The statistic of the test is stated according to,

$$FD = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \sim \chi^2_{k-1}, \tag{2}$$

where is distributed according to chi-squared distribution with $k-1$ degrees of freedom, with n observations and k groups, R_j^2 is the squared rank of j-th compared approach. Under the null hypothesis, there is no difference between results from different groups.

In the sequence, according to [18], if Friedman's null hypothesis is rejected, it is necessary to apply a post-hoc test to find which groups have different results. Hence, a multiple comparison test of Nemenyi can be applied. In this approach, a threshold is obtained by

$$CD = \frac{q_{\infty,k,\alpha}}{\sqrt{2}}\sqrt{\frac{k(k+1)}{6}}, \tag{3}$$

where CD being the critical difference to assume that there is a difference or not in the measures of groups, $q_{\infty,k,\alpha}$ is the studentized range statistic, k the number of algorithms, at α significance level, and n number of samples. If the critical differences of the rank sum $|R_i - R_j|$ is greater than CD, there is a difference between results from algorithms i and j [19]. Therefore, these two approaches are applied to compare the errors from all proposed algorithms.

III. Methodology

The forecasting and optimization results presented in this paper were developed in `R` software by using `caret` [20] and `metaheuristicOpt` [21] packages, respectively.

For each optimization method, a set of decision variables is initialized, then the XGBoost (linear booster) architecture is trained, and the optimization is performed. Cross-validation for time series using sliding window is employed in training set. The models' structured to one, two, and three-months-ahead forecast are presented in Eq. (4), and computed as follows,

$$\hat{y}_{(t+h)} = \begin{cases} \hat{f}\left\{y_{(t+h-1)}, y_{(t+h-2)}, y_{(t+h-3)}, x_{(t+h-1)}\right\} \text{ if } h = 1, \\ \hat{f}\left\{\hat{y}_{(t+h-1)}, y_{(t+h-2)}, y_{(t+h-3)}, x_{(t+h-2)}\right\} \text{ if } h = 2, \\ \hat{f}\left\{\hat{y}_{(t+h-1)}, \hat{y}_{(t+h-2)}, y_{(t+h-3)}, x_{(t+h-3)}\right\} \text{ if } h = 3. \end{cases} \tag{4}$$

In fact, f is a function that maps the observed data, $\hat{y}(t+h)$ is the forecast dengue case in forecasting horizon $h = 1, 3$ at time

t $(1, \ldots, 132)$, $y(t+h-1), \hat{y}(t+h-2), y(t+h-3)$ are the previous observed and predicted dengue cases, $\boldsymbol{X}(t+h-n_x)$ is the inputs vector composed by average relative humidity, maximum and minimum temperatures, precipitation and thermal amplitude at the maximum lag of inputs ($n_x = 1, 2, 3$).

The XGBoost hyperparameters (a linear booster) and search space are shown in Table II.

TABLE II
THE XGBOOST HYPERPARAMETERS, DESCRIPTION, LOWER AND UPPER BOUNDARIES.

	Parameter	Description	Lower boundarie	Upper boundarie
T	Boosting iterations	Number of base learners	1	300
α	L_1 regularization	Equivalent to Lasso Regression on weights	0.0001	1
λ	L_2 regularization	Equivalent to Ridge Regression on weights	0.0001	1
η	Shrinkage	Learning rate applied to base learner expansion	0.001	0.3

The RMSE minimization starts when each method initializes the set of hyperparameters according to the boundaries shown in Table II. In the sequence, the predicted values are obtained and the RMSE is calculated according to (5) and the optimization algorithm is executed. Indeed, the RMSE is computed according to (5) as follows,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}, \tag{5}$$

Moreover, the MAPE and IP are adopted to evaluate the accuracy of each forecasting model and are computed according to,

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

$$\text{IP} = 100 \times \frac{M_p - M_c}{M_c}, \tag{7}$$

where n represents the number of observations of the training and test sets, y_i is the i-th observed value and $\hat{y}_i$ is the i-th predicted value obtained by XGBoost. Also, the M_c and M_p represent the performance measure of compared and proposed model, respectively. Besides, the optimization process starts and ends when the number of generations or iterations is met. To check the robustness of each optimization approach, 30 runs are conducted. Therefore, the statistical indicators such as minimum, median, arithmetic average, maximum, standard deviation (Std), and trimmead average (arithmetic average without 10% of bigger and lower values) are computed.

In Figure 3 is depicted the proposed methodology.

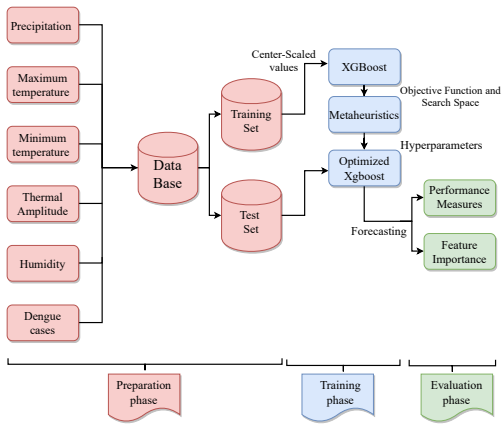

Fig. 3. Flowchart of proposed methodology.

The following metaheuristics were tested and compared in the XGBoost tuning: CS, DE, FFA, GA, and GWO. The optimizers compared with COA were of CRAN library given in https://cran.r-project.org/web/packages/metaheuristicOpt/metaheuristicOpt.pdf with standard settings, and are presented in Table III.

TABLE III
PARAMETERS OF EACH ADOPTED OPTIMIZER.

Method	Parameter	Value
CS	Population Size	25
	Abandoned Fraction	0.5
COA	Number of Packs	10
	Number of Coyotes	8
DE	Population Size	25
	Mutation Factor	0.8
	Crossover Ration	0.5
FFA	Population Size	25
	Attractiveness Firefly	1
	Light Absorption Coefficient	1
	Randomization Parameter	0.2
GA	Population Size	25
	Mutation Probability	0.5
	Crossover Probability	0.8
GWO	Population Size	25

In Table IV are showed the hyperparameters (for a linear booster) selected by each optimization algorithm.

TABLE IV
XGBOOST HYPERPARAMETERS SELECTED BY EACH METAHEURISTC.

Method	Statistical Indicator	Hyperparameter			
		# Boosting Iterations	L1 Regularization	L2 Regularization	Learning Rate
COA	Average	141.3179	0.7588	0.5268	0.1661
	Std	56.2504	0.2026	0.2845	0.0824
CS	Average	145.8778	0.6817	0.5161	0.1442
	Std	79.5053	0.2063	0.2931	0.0827
DE	Average	171.3180	0.6258	0.4306	0.1372
	Std	103.9385	0.2327	0.2861	0.0905
FFA	Average	183.9358	0.6521	0.4372	0.1554
	Std	90.8867	0.2538	0.2868	0.0956
GA	Average	138.2354	0.7623	0.4810	0.1371
	Std	86.7825	0.1463	0.2941	0.0818
GWO	Average	153.8231	0.7205	0.4691	0.1736
	Std	82.1156	0.2191	0.2541	0.0804

The results presented in Section IV are generated using the processor Intel(R) Core(TM) i5-4200U central processing unit of 1.6Hz, 8 gigabyte of random access memory in Windows 10 operating system. The R software [22] is adopted to perform the modeling.

IV. RESULTS

The effectiveness of the proposed ensemble, COA-XGBoost, is measured in terms of RMSE and MAPE. Table V presents statistical indicators such as minimum, median, arithmetic average, maximum, Std, and trimmed average (arithmetic average without 10% of bigger and lower values) for each optimization method. For each statistic, the best results are presented in bold.

To forecast dengue cases one month ahead, the COA outperforms the compared methods in terms of RMSE for minimum, arithmetic average, and trimmed average. In fact, the IP ranges between 0.09% - 1.06%, 0.05% - 2.22%, and 0.38% - 4.27% for minimum, arithmetic average, and trimmead average, respectively. When the MAPE criterion is adopted, similar results are achieved for median and trimmead average. Regarding the remaining measures, competitive results are achieved when results are compared with GA and GWO metaheuristics.

Concerning the two-month-ahead forecasting horizon, for RMSE and MAPE analysis, COA outperforms compared metaheuristics in most of the indicators, except in minimum and maximum for

TABLE V
RESULTS OF THE OPTIMIZED XGBOOST MODEL IN TERMS OF RMSE AND MAPE (30 RUNS) TO FORECAST DENGUE CASES ONE UP TO THREE-MONTHS-AHEAD.

Forecasting Horizon	Statistical Indicator	COA RMSE	COA MAPE	CS RMSE	CS MAPE	DE RMSE	DE MAPE	FFA RMSE	FFA MAPE	GA RMSE	GA MAPE	GWO RMSE	GWO MAPE
One-Month-Ahead	Minimum	**1713.02**	0.84	1731.29	0.89	1731.15	0.86	1729.77	0.90	1714.62	**0.78**	1729.11	0.90
	Median	1766.03	**0.95**	1789.70	1.05	1804.11	1.10	1794.90	1.09	1775.90	0.98	**1764.10**	1.04
	Arithmetic Average	**1781.11**	1.04	1822.97	1.07	1867.30	1.10	1838.31	**1.01**	1782.03	1.03	1798.53	1.05
	Maximum	2324.51	1.37	2367.71	1.32	2469.16	1.33	2365.05	1.33	**1952.43**	**1.31**	2027.33	1.32
	Std	106.51	0.16	125.53	0.14	169.64	0.15	158.45	0.16	**48.44**	0.15	88.34	**0.12**
	Trimmead Average	**1771.63**	**1.03**	1806.79	1.07	1850.67	1.10	1823.37	1.10	1778.35	1.03	1792.84	1.05
Two-Months-Ahead	Minimum	2137.04	1.66	2134.07	1.76	2119.51	1.77	2149.11	1.67	**1972.87**	**1.47**	2148.96	1.67
	Median	**2255.32**	**2.01**	2308.71	2.08	2359.21	2.11	2342.11	2.05	2267.86	2.03	2285.98	2.06
	Arithmetic Average	**2304.89**	**2.01**	2414.14	2.12	2452.36	2.16	2482.09	2.09	2325.46	2.05	2470.95	2.12
	Maximum	2579.12	**2.35**	3459.93	2.83	3485.67	2.89	3524.81	2.90	**2578.88**	2.36	3521.92	2.90
	Std	**122.48**	**0.18**	314.00	0.23	363.39	0.28	356.59	0.27	145.87	**0.18**	424.53	0.36
	Trimmead Average	**2301.09**	**2.01**	2386.79	2.10	2427.34	2.15	2456.74	2.08	2329.00	2.06	2444.91	2.11
Three-Months-Ahead	Minimum	4648.69	3.94	4359.29	3.12	4481.53	3.44	4329.44	**2.88**	**4098.81**	3.90	4712.15	3.92
	Median	**4716.16**	6.19	4807.40	**5.09**	4944.75	5.34	4925.31	5.27	4794.93	6.14	4912.82	5.79
	Arithmetic Average	**4837.34**	5.93	4886.85	5.45	4968.31	5.62	4965.94	**5.50**	4841.11	5.87	5001.54	5.86
	Maximum	**5128.22**	7.56	5744.22	**7.55**	5781.80	7.96	5828.92	7.55	5198.96	8.01	5787.45	7.62
	Std	**119.14**	1.32	268.26	1.45	286.06	1.37	306.57	1.36	201.01	1.42	319.13	**1.27**
	Trimmead Average	**3577.82**	3.90	3630.58	**3.68**	3687.54	3.80	3706.88	3.71	3579.39	3.88	3711.34	3.92

RMSE, and minimum for MAPE. For RMSE, the IP ranges between 0.55% - 4.40%, 0.88% - 6.72%, 16.03% - 66.29%, and 1.20% - 6.34% for median, arithmetic average, Std, and trimmead average, respectively. For MAPE, 1.10% and 4.90%, 2.11% - 6.90%, 0.70% - 19.08%, 1.22% - 48.72%, and 2.54% - 6.34% for median, arithmetic average, maximum, Std, and trimmead average, respectively. For this forecasting horizon, GA is the optimizer which the most similar performance of COA.

Looking for three-months-ahead, for RMSE, similar results of one and two-month-ahead are obtained. However, in terms of MAPE, COA is outperformed by CS, GWO, and FFA metaheuristics. In a broader perspective, considering RMSE for all forecasting horizons, the COA reaches better accuracy in 66.67% of the comparisons, while for MAPE, in 38.89% of the cases. The ranking of metaheuristics is composed of COA, GA, CS, GWO, FFA, and DE to tune the XGBoost hyperparameters and forecasting dengue cases for the PR state. The aforementioned ranking is due to the performance of each approach evaluated for all forecasting horizons and criteria.

In Figure 4 are showed the observed and predicted number of dengue cases for one up three-months-ahead. For the first two forecasting horizons, the COA–XGBoost learns data behavior, in most of the cases, which allows predictions compatible with the observed values. There is a challenge in the three-months-ahead forecast task, once the recursive method accumulates the errors of previous forecasts to the next. Moreover, in this forecasting horizon, in the ups and downs, the forecasting model has difficulty in capturing the data variability.

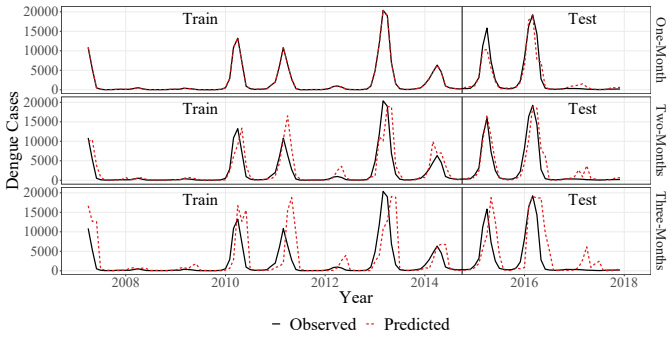

Fig. 4. Forecasts and observed dengue cases.

In Figure 5 is presented the importance of each adopted input to forecasting dengue cases. These scores are computed using the properties of the base learners used by the XGBoost method.

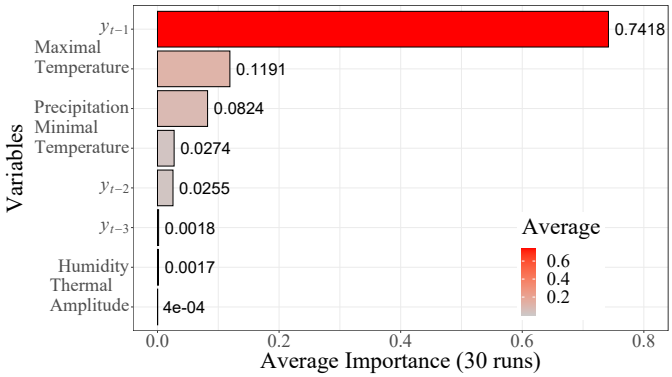

Fig. 5. Inputs importance.

The proposed ensemble COA-XGBoost showed that dengue previous cases exert greater importance in predicting future cases, followed by, minimum and maximum temperature, precipitation, humidity, and thermal amplitude, in this order.

Considering Friedmann's test of the RMSE and MAPE over 30 runs for the adopted metaheuristics, there are statistically different results to at least two metaheuristics at 5% level (χ^2_5 = 478.03 - 485.37, p-value < 0.05). In Figures 6 and 7 are showed the CD plot based on Nemenyi test.

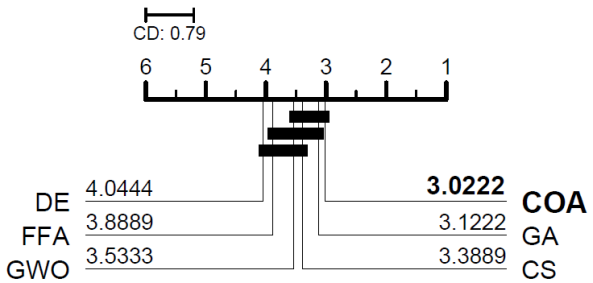

Fig. 6. Critical distance plot for RMSE.

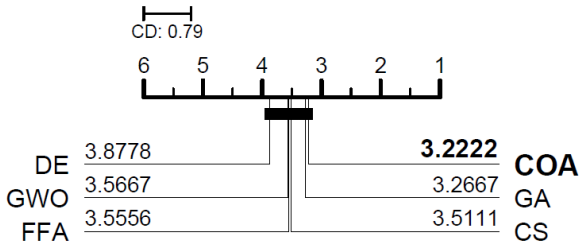

Fig. 7. Critical distance plot for MAPE.

Those metaheuristics that are not joined by a line can be regarded as different. The CD to consider the errors statistically different are 0.79 in both cases. In both cases, the COA achieved the best ranking, followed by GA, CS, FFA/GWO, and DE metaheuristics.

V. Conclusion

In this paper, a novel combination of COA and XGBoost has been proposed to forecast dengue cases one and three-month-ahead, in PR state, Brazil. The importance of average precipitation, maximum and minimum temperature, thermal amplitude, and humidity were evaluated. The COA was employed to obtain a set of XGBoost hyperparameters by minimizing the RMSE. The performance of the proposed XGBoost model using COA was compared with XGBoost coupled with DE, CS, COA, DE, FFA, GA, and GWO for hyperparameters tuning. Regarding accuracy (average of RMSE and MAPE, over 30 runs), the proposed COA-XGBoost outperforms compared approaches.

Even with good results achieved in terms of evaluated criteria, this study has the following limitations: (i) The spatio-temporal analysis was didn't perform to taking account the most affected locations, the proposed model was not able to perform well for three-months-ahead, the parameters of COA optimizer were selected by trial and error. The proposed ensemble learning approach based on XGBoost and COA got the most suitable set of hyperparameters in a shorter time when compared to other optimization approaches. Fortunately, the COA-XGBoost is an efficient tool for dengue cases forecasting helping health managers in the decision-making process. As future research is intended to perform signal decomposition, considering converge analysis of the optimizers, comparisons of other forecasting models such as random forest and artificial neural networks, to adopt optimization approaches for feature engineering and selection, and other evolutionary and swarm intelligence algorithms, such as particle swarm optimization, owls [23] and falcon [24].

Acknowledgments

The authors would like to thank National Council of Scientific and Technologic Development of Brazil – CNPq (Grants number: 307958/2019-1-PQ, 404659/2016-0-Univ, 307966/2019-4-PQ), PRONEX '*Fundação Araucária*' 042/2018, and *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brazil – CAPES (Finance Code: 001) for its financial support of this work.

References

[1] Brasil, *Monitoramento dos casos de arboviroses urbanas transmitidas pelo Aedes Aegypti (dengue, chikungunya e zika), semanas epidemiológicas 1 a 50, 2020. Ministério da Saúde*, 2021, access in 17 may. 2021, (in Portuguese). [Online]. Available: https://www.gov.br/saude/pt-br/assuntos/media/pdf/2020/dezembro/28/boletim_epidemiologico_svs_51.pdf

[2] P. Guo, Q. Zhang, Y. Chen, J. Xiao, J. He, Y. Zhang, L. Wang, T. Liu, and W. Ma, "An ensemble forecast model of dengue in Guangzhou, China using climate and social media surveillance data," *Science of The Total Environment*, vol. 647, pp. 752–762, 2019.

[3] J. Liebig, F. de Hoog, D. Paini, and R. Jurdak, "Forecasting the probability of local dengue outbreaks in Queensland, Australia," *Epidemics*, vol. 34, no. 100422, 2021.

[4] W. A. Abualamah, N. A. Akbar, H. S. Banni, and M. A. Bafail, "Forecasting the morbidity and mortality of dengue fever in ksa: A time series analysis (2006–2016)," *Journal of Taibah University Medical Sciences*, vol. 16, no. 3, pp. 448–455, 2021.

[5] E. Mussumeci and F. Codeço Coelho, "Large-scale multivariate forecasting models for dengue - lstm versus random forest regression," *Spatial and Spatio-temporal Epidemiology*, vol. 35, no. 100372, 2020.

[6] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Computing Surveys*, vol. 45, no. 1, pp. 10:1–10:40, Dec. 2012.

[7] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, New York, NY, USA, 2016, pp. 785–794.

[8] R. Shi, X. Xu, J. Li, and Y. Li, "Prediction and analysis of train arrival delay based on XGBoost and bayesian optimization," *Applied Soft Computing*, vol. 109, no. 107538, 2021.

[9] Z. J. Ye and B. W. Schuller, "Capturing dynamics of post-earnings-announcement drift using a genetic algorithm-optimized xgboost," *Expert Systems with Applications*, vol. 177, no. 114892, 2021.

[10] J. Gu, W. Liu, K. Zhang, L. Zhai, and Y. Zhang, "Reservoir production optimization based on surrogate model and differential evolution algorithm," *Journal of Petroleum Science and Engineering*, no. 108879, 2021.

[11] J. Pierezan and L. S. Coelho, "Coyote optimization algorithm: A new metaheuristic for global optimization problems," in *IEEE Congress on Evolutionary Computation (CEC)*, Rio de Janeiro, Brazil, 2018, pp. 2633–2640.

[12] M. H. D. M. Ribeiro and L. S. Coelho, "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," *Applied Soft Computing*, vol. 86, no. 105837, 2020.

[13] M. Qais, H. Hasanien, S. Alghuwainem, and A. Nouh, "Coyote optimization algorithm for parameters extraction of three-diode photovoltaic models of photovoltaic modules," *Energy*, vol. 187, p. 116001, 2019.

[14] Brasil, *Ministério da Saúde, Departamento de Informática do Sistema Único de Saúde (DATASUS)*, 2019, access in 22 jul. 2019, (in Portuguese). [Online]. Available: http://www2.datasus.gov.br/DATASUS/index.php?area=0203&id=29892234&VObj=http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinannet/cnv/menin

[15] Brazil, "Ministério da Agricultura, Pecuária e Abastecimento. [Ministry of Livestock Agriculture and Supply]," 2019, http://www.inmet.gov.br/projetos/rede/pesquisa/. Access in 10 Oct. 2019 (In Portuguese). [Online]. Available: http://www.inmet.gov.br/projetos/rede/pesquisa/

[16] P. A. Morettin and C. Toloi, *Análise de séries temporais*, 2nd ed. São Paulo, Brasil: Blucher, 2006, (in Portuguese).

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[18] J. Carrasco, S. García, M. Rueda, S. Das, and F. Herrera, "Recent trends in the use of statistical tests for comparing swarm and evolutionary computing algorithms: Practical guidelines and a critical review," *Swarm and Evolutionary Computation*, vol. 54, no. 100665, 2020.

[19] P. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Statistics Schoool, Princeton University, Princeton, USA, 1963.

[20] M. Kuhn, "Building predictive models in r using the caret package," *Journal of Statistical Software, Articles*, vol. 28, no. 5, pp. 1–26, 2008.

[21] L. Septem Riza, Iip, E. Prasetyo Nugroho, M. B. Adi Prabowo, E. Junaeti, and A. G. Abdullah, `metaheuristicOpt: Metaheuristic for Optimization`, 2019, r package. [Online]. Available: https://CRAN.R-project.org/package=metaheuristicOpt

[22] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: http://www.R-project.org/

[23] E. H. de Vasconcelos Segundo, V. C. Mariani, and L. S. Coelho, "Metaheuristic inspired on owls behavior applied to heat exchangers design," *Thermal Science Engineering Progress*, vol. 14, no. 100431, 2019.

[24] ——, "Design of heat exchangers using Falcon optimization algorithm," *Applied Thermal Engineering*, vol. 156, pp. 119–144, 2019.