Autoregressive Short-Term COVID-19 Cases Forecasting with Gaussian Processes Regression

Elloá B. Guedes, André Igor Nóbrega da Silva, Paulo Ribeiro Lins, and Edmar C. Gurjão, *Senior Member, IEEE*

***Abstract*— This paper addresses the short-term COVID-19 forecasting problem using an autoregressive approach with Gaussian Process Regression based on additive time series decomposition and a novel kernel selection method. Experimental results on a case study with three different scales on observational data from Brazil COVID-19 pandemic exhibited $R^2 \geq 0.97$ and strong tolerance to training with few data, a contrasting advantage over many data-hungry Machine Learning methods. Moreover, when comparing the proposed approach with predictions from an Automated Machine Learning optimization pipeline, no statistical differences were observed at the advantage of a much smaller space state search and savings on computational resources. The results obtained might support decision making to implement social distancing interventions, to improve health's supply chain logistics and to plan and implement vaccination campaigns, etc. aiming at reducing transmission and decelerating the pandemic progression in multiple scenarios.**

***Index Terms*— COVID-19, Forecasting, Gaussian Regression Processes, Scale**

I. Introduction

Big data is a contemporary phenomenon in which the data production rate is greater than the processing and storing capacity [1]. Therefore, appropriate data processing and management could expose new knowledge and facilitate in responding to emerging opportunities and challenges on time [2]. That is the case for the unprecedented public health emergency in the COVID-19 pandemic that, according to the World Health Organization (WHO), led to 261,978,819 confirmed cases and 5,205,121 deaths worldwide as of November 2021 [3].

Epidemiological time series forecasting plays an essential role in public health systems [23] and several machine learning models that can forecast the COVID-19 outbreak globally have been released [38]. Such models are crucial to fight against the ongoing pandemic helping authorities and managers to (*i*) create, adopt, revise and sustain social distancing policies (schools reopening, temporarily closing stores, and reducing hours, for example); (*ii*) rationalize COVID-19 tests, especially when there is outbreak potential; (*iii*) improve hospitals' logistics by anticipating the demand for beds, medicine, equipment, and others.; (*iv*) plan and implement vaccination campaigns, especially focusing on a vulnerable population, among others.

Brazil has been heavily affected by the coronavirus disease, even becoming, during a critical period, the global epicenter of the pandemic. So far, there has been 614,278 and 22,080,906 total deaths and cases, respectively [7]. Pandemic control in Brazil is a multi-factor problem due to urban density, the timing of the implementation and guarantee of social distancing policies, and limited testing capacity [6]. Important studies reported major consequences of the pandemic in Brazilian society, with direct effects on families' work and income, implications on the physical and mental health of individuals [9] and even a reduction in life expectancy [39]. Despite such consequences, Brazil is carrying out its historically largest vaccination campaign, currently occupying the 4th place in the global COVID-19 vaccination ranking, having 77.2 % of its inhabitants with at least a single shot, 62.9 % completely vaccinated, and 6.6 % with the boost vaccine shot [8].

Accurate information is the basis for the development and implementation of actionable disease control measures during public health emergencies [15]. However, the official panel for COVID-19 cases and deaths has some limitations in the updating process due to data gathering dynamics and complexity, so numbers are under permanent review. Furthermore, aggregating endogenous and exogenous variables to help create a COVID-19 forecast model in Brazil may not be feasible because, for instance, there are no formal data on ICU and sub intensive care beds [16]; the Brazilian epidemiological database of the severe acute respiratory syndrome reportedly has a large number of errors and inconsistencies due to manually inserted data [17]; crowdsourced data may not be representative because, according to the National Household Sample Survey, 25% of population has no internet access [18], etc.

Although the challenges, forecasting the COVID-19 epidemiological scenario on a short-term horizon in Brazil remains an important task to help mitigate the effects of this pandemic situation [20]. The availability of information can be crucial, for example, to strategic decision making in the Brazilian Unified Healthcare System used by 70% of the population [19], and also to policymakers to implement social distancing interventions to reduce transmission of the virus and decelerate the pandemic progression;Aiming at addressing such a forecasting task, significant contributions for Brazil have been found in literature [6], [21], [23]. Those works consider compartmental epidemiological

Elloá B. Guedes is with Universidade do Estado do Amazonas, Brazil (e-mail: ebgcosta@uea.edu.br).

Paulo R. Lins is with the Federal Institute of Technology, Brazil (e-mail: paulo.lins@ifpb.edu.br).

André Igor Nóbrega and Edmar C. Gurjão are with Electrical Engineering Department of Federal University of Campina Grande, Brazil (e-mails: andre.igor@ee.ufcg.edu.br and ecg@dee.ufcg.edu.br).

models (SIR, SEIR, etc.) and machine learning techniques for forecasting country and state levels.

Decreasing to a city scale, the findings in literature for Brazil turn out to be scarce, probably as a result of the dichotomy between small datasets and data-hungry forecasting models. Nevertheless, forecasting COVID-19 cases at a city level remains a critical problem for the country because it would endorse a divide-and-conquer approach to fight the pandemic, bypassing the complexity of the country's continental proportions. Among other advantages, it would help city authorities to improve health resources management, to continuously evaluate the sustainability and effectiveness of interventions to reduce transmission, and to minimize as much as possible the unfathomable cost in human lives.

Regarding this context, this paper aims to introduce an autoregressive approach to forecasting COVID-19 daily cases in Brazil based on additive time series decomposition and a novel kernel selection method for Gaussian Process Regression (GPR). In this respect, we considered a case study on three different scales (country, state, and city) using data from the country official COVID-19 panel to validate our approach in contrast with predictions from an Automated Machine Learning (AutoML) optimized pipeline. Results from a time series split cross-validation delivered high-quality predictions in the scales under evaluation with no statistical difference from the optimized counterpart, at a small cost on searching the solution space state and computational resources.

The contributions of this paper can be summarized as follows:

- Proposition of an efficient strategy for forecasting one-week-ahead new daily cases of COVID-19 based on additive time series decomposition and a novel kernel selection method for Gaussian Process Regression;
- Application and evaluation of the proposed strategy on a case study that considered three different forecasting scales (country, state, city) whose results showed the robustness of the method even with different amounts of data;
- Validation of the proposed method using a comparison with a state-of-art AutoML pipeline optimization where no evidence of statistically different performances was observed.

To introduce such contributions, the rest of the paper is organized as follows: Section II bring in interseccional studies in literature; Section III contains the core of our contribution, where experimental data is depicted in Section III-A, a brief background on GPR is presented in Section III-B, the novel kernel selection method is described in Section III-C, and performance evaluation and validation strategies are discussed in Section III-D; results obtained are delineated in Section IV; and, lastly, conclusions and perspectives are drawn in Section V.

II. Related Work

Contributions on forecasting COVID-19 cases in Brazil have been found in the literature. Some approaches were based on compartmental epidemiological models (SIR, SEIR, SUEIHCDR, etc.) [21], [40], [41], but their accuracy and suitability for predicting this disease is a matter of debate [22], [42]. Upon analyzing forecasting with Machine Learning techniques, there was heterogeneity regarding selected endogenous and exogenous independent variables, techniques and models evaluated and also on several values for the horizon in short-term prediction [6], [20], [23]–[25]. When decreasing the scale to a city level, in particular, the findings in literature turn out to be scarce, probably due to the dichotomy between small datasets and data-hungry forecasting models.

When considering GPR for COVID-19 forecasting, Velásquez & Lara used 82 days of observation and continuous learning with data from the USA to experimentally conclude that these methods can be meaningfully used to gather a quantitative picture of the epidemic spreading in such location [44]. The work of Ahmad *et al.* applied GPR models for classification and prediction of confirmed COVID-19 cases in the Middle East and Asia regions and experimentally verified that Matérn kernel with smoothness hyperparameter of 5/2 were highly accurate [45].

In Brazil, a GPR with a linear kernel was used as a meta-learner for a stacking ensemble with 4 distinct types of base-learners [23]. The resulting stacking ensemble was the most suitable tool for forecasting COVID-19 cases amongst the strategies evaluated in several state-level experimental scenarios. However, it is important to notice that GPR was not directly used as a regressor for COVID-19 cases in the work mentioned above. In recent work, Alali *et al.* considered the use of GPR followed by Bayesian optimization to forecast the recovered and confirmed COVID-19 cases in two highly impacted countries, India and Brazil [46]. For the latter case, in particular, the authors used a holdout cross-validation approach with a single test partition under the Supervised Learning Paradigm and 10 different kernel functions for GPR. As a result, authors obtained high-performance metrics for optimized GPR models compared to other 15 Machine Learning models.

Although our proposed strategy is based on Supervised Machine Learning methods and observational COVID-19 data from Brazil, it fundamentally differs from the existing literature in the following aspects: (*i*) GPR is directly used for the forecasting task without exogenous variables nor posterior optimizations; (*ii*) a cross-validation approach especially suited for time series problems is used to evaluate the method under the pandemic dynamics, where more data becomes available as time passes by; (*iii*) a composite kernel function is proposed from a kernel selection method that considers closure properties under addition instead of considering the representational limitations of single kernel functions.

III. Materials and Methods

In this paper the problem of forecasting COVID-19 daily cases was modeled as an autoregressive time series task with a one-week ahead horizon. The independent variables denoted by X_t where t is a discrete time index, is the value

of seven past lagged observations of the cumulative daily cases, such that $X_t \leq X_{t+1}$. Our goal is finding a model that best fits a function $f : \mathbb{N}^7 \mapsto \mathbb{N}$ such that:

$$y_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-7}). \tag{1}$$

By defining a target variable y_t as such, it is possible to amortize daily variation and also simplify f as being monotonic. Taking such considerations the next subsections introduce the experimental data, a background on GPR and its most relevant kernel functions, our kernel selection method, and performance evaluation, and the adopted validation strategies.

A. Experimental Data

The data used in this paper comes from the Brazilian Health Ministry who daily publishes the COVID-19 registered number of cases. Presented results are based on the data up to November 17$^{\text{th}}$, 2021 [7]. From this dataset, we considered the time series of cumulative confirmed cases in three different scales: the first one for the whole country, the second for a state-level considering the Paraiba state, and the third for the city level for Campina Grande in Paraiba.

Inspecting the dependent variable in the first dataset (country level), there were 21,977,661 confirmed cases starting from February 25$^{\text{th}}$, 2020. At the state level, there were 457,831 cases confirmed so far. At a city level, there were 47,447 confirmed cases. Graphics in Fig. 1 show the evolution of cumulative daily cases and also the logarithm growth rates in each of the locations above.

Data pre-processing was required to represent the autoregressive forecasting task under consideration correctly: time stamps were replaced by consecutive integer indexes, and the experimental samples for an independent and dependent variable were obtained. As a result, for Brazil and Paraíba, there are 632 samples, and for Campina Grande, 601 samples.

B. Gaussian Processes Regression

Gaussian Processes Regression (GPR) is a non-parametric Bayesian approach towards regression problems utilized in exploration and exploitation scenarios. It can capture relationships in data by utilizing a theoretically infinite number of parameters and letting the data determine the level of complexity via Bayesian inference [26], [29]. Theoretical and practical developments over the last decade turned GPR into a severe competitor to supervised learning applications [43].

Schulz, Speekenbrink & Krause introduce GPR in an accessible tutorial [26]. According to them, the output y of a (unknown) function f at input $\mathbf{x}$ can be written as

$$y = f(\mathbf{x}) + \epsilon, \tag{2}$$

with $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ representing the inherent observation randomness. Assuming $f(\mathbf{x})$ is a random variable with unknown distribution, uncertainty regarding f can be reduced by observing the output at different input points. This way, $f(\mathbf{x})$ is assumed to follow a Gaussian process

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{3}$$

where $\mathcal{GP}$ is a distribution over functions defined by a mean $\mu(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$. The mean function $\mu(\mathbf{x})$ is the expected function value at input $\mathbf{x}$, so that $\mu(x) = \mathbb{E}[f(\mathbf{x})]$, i.e., the average of all functions in the distribution evaluated at input $\mathbf{x}$. The covariance function, hereafter named kernel of the Gaussian process and denoted by $k(\mathbf{x}, \mathbf{x}')$, models the dependence between the function values at different input points $\mathbf{x}$ and $\mathbf{x}'$, and it is given by

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}')\right]. \tag{4}$$

The prior mean function is often set to $\mu(\mathbf{x}) = 0$ to avoid expensive posterior computations and only make inference via the covariance function [27], [29].

There are two standard kernels categories: (1) stationary kernels, whose value only depends on the difference $\mathbf{x} - \mathbf{x}'$ and (2) non-stationary kernels, whose modeled functions depend on the value of the input coordinates themselves, meaning that the corresponding GPR model will produce different predictions if the data were moved while the kernel parameters were kept fixed [28]. The choice of an appropriate kernel is based on assumptions such as smoothness and likely patterns to be expected in the data [26], and also new kernels can be composed of existing ones given the closure under sum, product and exponentiation operations [43].

A detailed example of how GPR works is shown in Fig. 2 where a function is splitted into train and test partitions with 70 % and 30 % samples, respectively. Samples from a process with prior distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ are presented in the figure. After a GPR training process, sample functions were fitted to train data, and also the exploitation result for the test set was obtained.

According to Duvenaud [28], GPR is well-suited for regression problems because: it is less prone to overfitting when compared to neural networks, for example; it has high expressiveness of modeling assumptions and given a kernel function and some observations, the predictive posterior distribution can be exactly computed in closed form. However, the flexibility of GPR contrasts with the need to choose a kernel, which is equivalent to learning a valuable representation of the input.

Since GPR is a non-parametric method, in many practical applications, it may not be easy to specify with confidence all aspects of the kernel function [43]. The available functions do not just differ in their parametrization but in their fundamental structure [30]. Thus, the kernel function choice is addressed as a model selection problem comprehending both discrete choices and the setting of continuous (hyper-) parameters of such functions [43].

A strategy to overcome such difficulty is to adopt a cross-validation procedure by splitting the data into two disjoint sets: train and test sets where the performance on

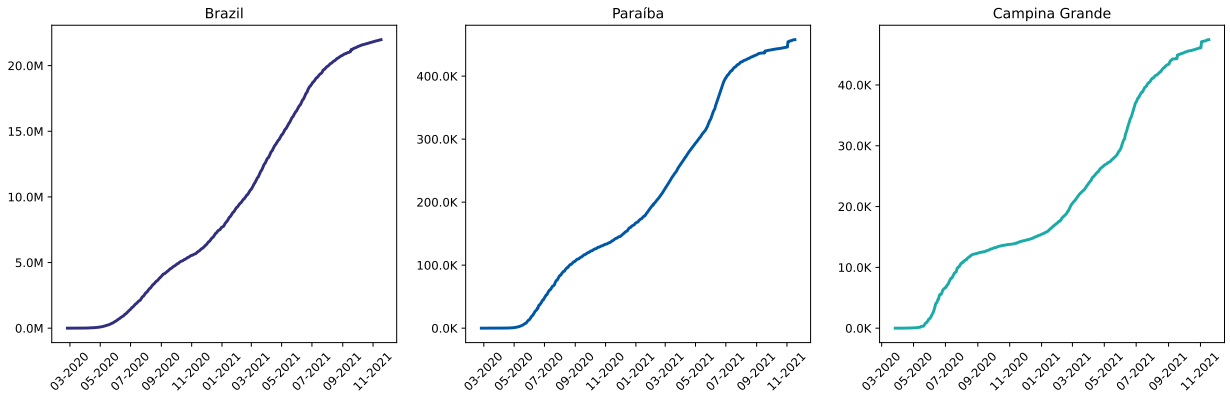


Fig. 1. Cumulative COVID-19 cases.

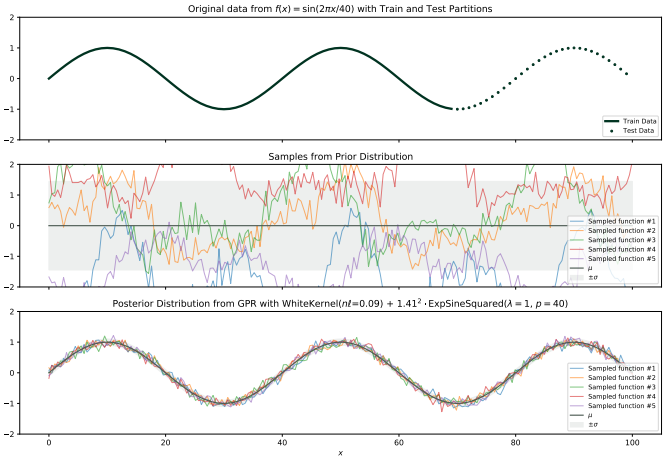


Fig. 2. Seasonal decomposition under the additive approach to target variable in each scenario. Seasonality component is restricted to the first **100** days for better visualization.

the test set is used as a proxy for the generalization error, and model selection is carried out using this measure [43]. In this work, we use a *time series-split cross-validation*: training data is divided into $k = 5$ fixed time intervals; in the first out of k rounds, the first split is used for training, and the second split is used for testing, in the second round, the first and second splits are used for training, and the third split is used for the test, and so on such that in each round test indices must be higher than before; shuffling in the time series split cross-validation approach is not adopted because it can cause leakage from future data to the model during training [31].

With such a cross-validation approach, we can evaluate GPR regression tolerance to training with few data and if GPR performance improves in a scenario where more data becomes progressively available, as it is desired when predicting a disease outbreak. However, while the cross-validation procedure strengthens the confidence in the generalization capabilities of a good kernel function, the model selection problem remains open. In order to deal with such a fundamental component of GPR, we introduce a kernel selection method for the COVID-19 forecasting problem.

C. Kernel Selection Method for Time Series Additive Decomposition

Upon examining the independent variable y as a time series, we performed an additive decomposition using the classical method as described by Hyndman & Athanasopoulos [5]. This is because we considered a seasonal period of a week (seven days) after analyzing surveillance data, and also because, at an individual level, the highest viral load was observed at the time of symptom onset and, therefore, infectiousness peaked on or before that window where the serial interval was estimated to have a mean of 5.8 days [32]. After such definition, the first step was to smooth the data using a centered moving average of order equal to the periodicity of the data $m = 7$ (7-MA) as follows:

$$d_t = \frac{1}{m} \sum_{i=-\beta}^{+\beta} y_{t+i} \tag{5}$$

where $\beta = \frac{m-1}{2}$. The resulting series d_t is the deseasonalized data. Indeed, in the case of the weekly periodicity, the value on Thursday, for instance, is the average of values from Monday to Sunday. In the next step, we obtained the *trend component* T_t, also known as the de-trended series, as follows:

$$T_t = y_t - d_t. \tag{6}$$

The *seasonal component* S_t was obtained by averaging the value for each day of the week on the de-trended series. The seasonal component for Monday, for instance, is equal to the average of all Mondays values contained in the T_t series. Lastly, the *remainder component* e was calculated by subtracting the estimated seasonal and trend-cycle components. By following such procedure, the target variable could be decomposed as shown in Eq. (7), so that each sum component represented an underlying pattern category.

$$y_t = T_t + S_t + e_t. \tag{7}$$

By applying such procedure in experimental data, as shown in Fig. 3, recognizable visual patterns for the target variable emerged in all scales considered: T seems to have linear growth, S is a periodic function, and e behaves like noise.

Considering the kernel selection problem, instead of dealing with the unfeasibility of a brute-force search over all possible kernel functions or considering kernel combinations with the shortcomings of a grid search that does not cover all state-space solutions. The seasonal additive decomposition approach of the target variable and the property of closure of kernels under the sum operation makes it straightforward to address the kernel selection problem in the COVID-19 forecasting context: a kernel function that fitted best each component was selected, then a new kernel was proposed from the sum of the kernels of the components, resulting in the kernel k given as follows:

$$\begin{aligned} k &= \text{DotProduct}(\sigma_0 = 1)^e \\ &+ \text{RationalQuadratic}(\alpha = 1, \ell) + \text{WhiteKernel}(n\ell), \end{aligned} \tag{8}$$

where $e \in [1, 10)$, $\ell \in \{1, 5, 10, 15, 50, 100, 250, 500, 1000\}$ is the length scale, and the noise level is in the range $-0.1 \leq n\ell \leq 2.0$ with discrete increments of 0.1. The other hyperparameters were kept fixed with values $\sigma_0 = \alpha = 1$. This search space resulted in 1701 different kernels configurations to be evaluated for each scenario.

D. Performance Evaluation and Validation

The R^2 score, also known as the coefficient of determination, was the performance metric chosen because it denotes the proportion of the variance in the dependent variable that is predictable from the independent variable, as follows:

$$R^2 = 1 - \frac{\sum_{i=i}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{9}$$

where y is the variable to be predicted, $\hat{y}$ is a prediction for y, and $\overline{y}$ is the mean of the target variable. The values for this metric lie in the range $(-\infty, +1]$, where $R^2 = +1$ indicates a perfect regressor [35]. A recent study states that this metric can be more (intuitively) informative than MAE, MAPE, MSE, and RMSE in regression analysis evaluation, as the former can be expressed as a percentage, whereas the latter measures have arbitrary ranges [36].

The proposed kernel selection method for COVID-19 autoregressive forecasting will be validated compared to an Automated Machine Learning (AutoML) approach. The field of AutoML aims at delivering a model that best performs in a particular application in a data-driven, objective, and automated way [33]. Literature has shown that AutoML approaches are already mature enough to rival and sometimes even outperform human machine learning experts [37], we chose such approach for validation in an analogy with the black box oracle machine concept from Computability and Complexity Theories in which such abstract computer is capable to solve complex problems in a single operation.

On this wise, we used TPOT, a Python tool that optimizes multiple machine learning algorithms (random forests, linear models, SVMs, etc.) in a pipeline with multiple preprocessing steps (missing value imputation, scaling, PCA, feature selection, etc.). All model hyperparameters and preprocessing steps, as well as multiple ways to ensemble or stack the algorithms within the pipeline [34]. We used the default settings with an initial population of 100 models and 100 generations which altogether with the cross-validation method will result in 50,000 pipeline configurations before finishing.

IV. Results and Discussion

The experimental procedures were carried out on a server with an Intel Core i9 3.7 GHz processor with 20 cores, 64 GB of primary memory, 3 GeForce RTX 3060 graphic cards with 12 GB each and 2 TB of secondary memory. The first step was to implement and execute the scripts to obtain the results from the proposed GPR approach. After that, we performed the AutoML optimization for the Brazil scenario, which resulting pipeline was then re-trained to the other scales.

The TPOT resulting pipeline is illustrated in Fig. 4, and first step is to perform a feature selection on input data taking into account the ANOVA F-value. After that, the input features and their respective p-values are given to a Gradient Boosting Regressor, then to a Linear Support Vector Classifier, and finally to a linear model trained with L_1 prior as regularizer fit with least angle regression (Lasso Lars). It can be noticed that there is explicit feature extraction and that an alternative representation using a classification problem is considered amidst the pipeline. Altogether with the model ensemble through stacking, it can be noticed that there are successive transformations to data in such a way that the relation between input and resulting output cannot be straightforwardly explained.

The resulting R^2 for the splits in the cross-validation approach considered is depicted in Table I for each scenario and considered method. It can be noticed that all $R^2 \geq 0.97$ indicate that both approaches delivered highly accurate predictions no matter the scale. Upon comparing the relative R^2 values difference per split between the proposed GPR approach with the AutoML strategy, we obtained the ΔR^2 values per split per scale. It can be observed that the largest percentage difference $\max(\Delta R^2) \approx 2.17\%$, which means that both methods had very similar performance.

In order to contrast and compare the performances, we carried out a Wilcoxon signed-rank test with a 95% confidence interval ($\alpha = 0.05$) on the experimental R^2 values shown in Table I from the forecasting scenarios with the GPR kernel selection proposed method and the AutoML approach. The null hypothesis (H_0) is that the paired R^2 values come from the same distribution and the alternative hypothesis (H_A) is the two-sided negation of the null hypothesis. As a result, we got a p-value of 0.000061, and since it is smaller than α, H_0 cannot be rejected. As a consequence, we exhibit statistical evidence of the strengths of the GPR kernel selection method

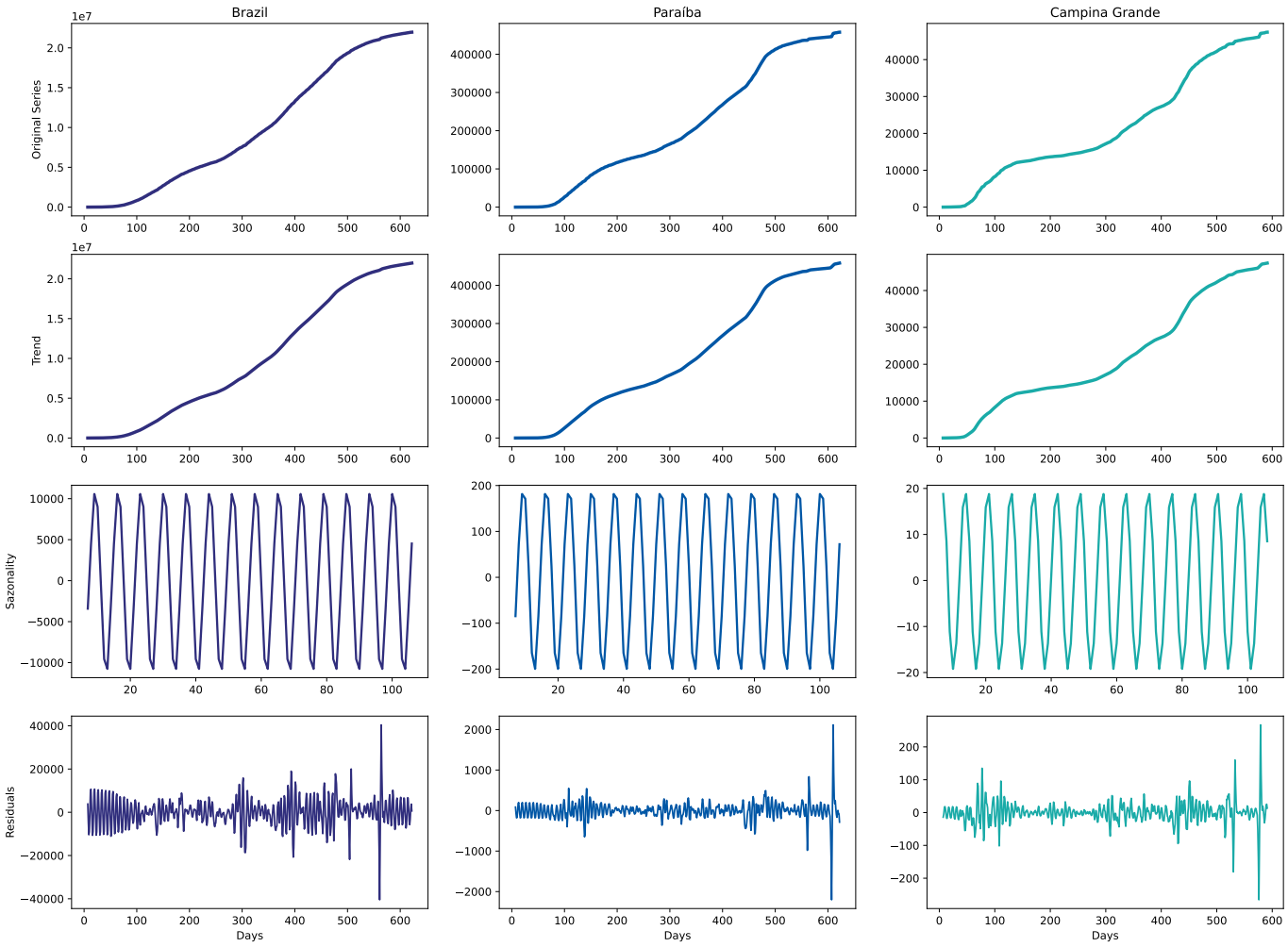
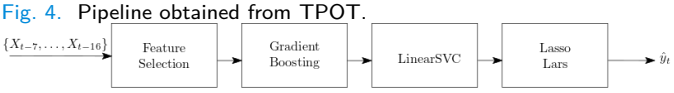


Fig. 3. Seasonal decomposition under the additive approach to target variable in each scenario. Seasonality component is restricted to the first **100** days for better visualization.

Fig. 4. Pipeline obtained from TPOT.



proposed for the problem of forecasting new COVID-19 cases on an autoregressive scenario on different scales.

Besides the R^2, we also analyzed the computational performance: the proposed GPR kernel selection method training time for Brazil was of 344.05 ± 8.32 s in 10 runs versus the AutoML approach that took 2450.28 s in a single run. It can be noticed that the proposed method delivered competitive results significantly faster, saving time and computational resources. It might be due to the size of the parameters search space which is 1701 kernel combinations versus 50,000 pipelines. More experiments are needed, however, in order to quantify better the computational performance contrasts between the approaches.

A visual argument is also presented to sustain further the positive evidence on the proposed GPR kernel selection method. By using the models above on a holdout cross-validation approach where 90% of data is used for training, and 10% is used for testing, we obtained the standardized residual plot shown in Fig. 5. The residuals were obtained from the difference between predicted and observed values then normalized. As can be noticed, for all scenarios under consideration, both GPR and AutoML models were predominantly close to a perfect fit, where residuals are zero. However, in the other cases, both strategies simultaneously yielded large residuals indicating challenging forecasting scenarios.

Fig. 5. Standardized residuals for predictions to **10%** latest data.

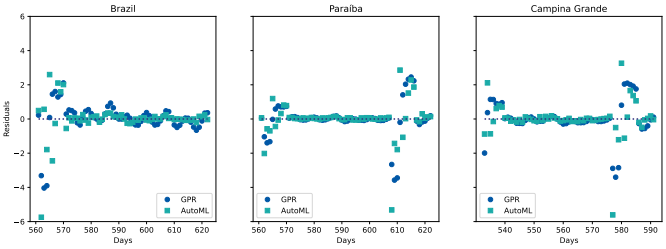


TABLE I
BEST KERNEL COMBINATIONS FOR GPR UNDER THE TIME SERIES SPLIT CROSS-VALIDATION AND COMPARISON WITH AUTOML APPROACH.

Scenario $e = 2,\ l = 5, nl = 0$			
	Brazil	AutoML	ΔR^2
Split 1	0.9879	0.9999	1.2204
Split 2	0.9958	0.9999	0.4192
Split 3	0.9987	0.9999	0.1243
Split 4	0.9997	0.9999	0.0237
Split 5	0.9995	0.9998	0.0325
$\mu \pm \sigma$	0.9963 ± 0.0045	0.9999 ± 0.0001	0.3640 ± 0.4516
Scenario $e = 2,\ l = 1, nl = 0$			
	Paraíba	AutoML	ΔR^2
Split 1	0.9977	0.9995	0.1782
Split 2	0.9996	0.9999	0.0365
Split 3	0.9906	0.9999	0.9426
Split 4	0.9917	0.9999	0.8268
Split 5	0.9942	0.9980	0.3756
$\mu \pm \sigma$	0.9948 ± 0.0034	0.9995 ± 0.0008	0.4719 ± 0.3557
Scenario $e = 2,\ l = 10, nl = 1.6$			
	Campina Grande	AutoML	ΔR^2
Split 1	0.9783	0.9998	2.1711
Split 2	0.9972	0.9999	0.2723
Split 3	0.9999	0.9999	0.0029
Split 4	0.9999	0.9999	0.0045
Split 5	0.9983	0.9989	0.0631
$\mu \pm \sigma$	0.9947 ± 0.0083	0.9997 ± 0.0004	0.5028 ± 0.8400

V. CONCLUSIONS

In this work, we introduced a data-driven autoregressive one-week-ahead forecasting approach for COVID-19 cases based on GPR. For that purpose, we designed a novel kernel selection method for GPR based on the additive decomposition of time series data that significantly reduced the hyperparameter space state search. Experimental evaluation and validation on a case study on three different scales (country, state, and city) with observational data from Brazil lead to performance metrics statistically not different from an AutoML optimization pipeline. Furthermore, the proposed method exhibited strong tolerance to training with few data and subsequent improvement as more data became available. Moreover, it is robust to many practical scenarios because it demands only a univariate COVID-19 daily cases time series and no other exogenous independent variables that may be difficult to track or correlate in a pandemic scenario.

In future work, we aim at (*i*) expanding the experimental evaluation of the method to other localities both from Brazil and other countries; (*ii*) improving the model validation by using a blocked time-series strategy; and (*iii*) design an automated procedure for extracting hyperparameter values from observational data.

ACKNOWLEDGMENT

Elloá B. Guedes acknowledges the financial support provided by FAPEAM and CNPq under the grant PPP 04/2017.

REFERENCES

[1] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods", *Journal of Business Research*, vol. 70, p.p. 263-286, issn 0148-2963, 2017.

[2] J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. "Big data challenge: a data management perspective", *Front. Comput. Sci.*, vol. 7, p.p. 157 -164, 2013.

[3] WHO, "Coronavirus disease (COVID-19) Pandemic", "Coronavirus disease (COVID-19) Pandemic", Available at https://www.who.int/emergencies/diseases/novel-coronavirus-2019, Available at https://www.who.int/emergencies/diseases/novel-coronavirus-2019 Accessed January 11, 2022.

[4] A. Nielsen, *Practical Time Series Analysis – Prediction with Statistics & Machine Learning'*, O'Reilly, 2020, Canada.

[5] R.J. Hyndman, and G. Athanasopoulos, *Forecasting: principles and practice*, 2018, OTexts, Edition 2, Melbourne, Australia. Available at https://otexts.com/fpp2/. Accessed January 11, 2022.

[6] M. M. de Oliveira, and T. L. Fuller, and P. Brasil, and C. R. Gabaglia, and K. Nielsen-Saines, "Controlling the COVID-19 pandemic in Brazil: a challenge of continental proportions", *Nature Medicine*, vol. 26, pp. 1505 - 1506, 2020.

[7] Health Ministry – Coronavirus panel, Available at https://covid.saude.gov.br/, Brasil, 2020. Accessed July 7th, 2021.

[8] Health Ministry – Brazil. Available at https://www.gov.br/saude/pt-br/vacinacao/, 2020, Accessed November 29th, 2021.

[9] W. da S. de Almeida, et al. "Changes in Brazilian socioeconomic and health conditions during the COVID-19 pandemic", *Revista Brasileira de Epidemiologia*, vol. 23, 2020.

[10] F. Petropoulos, and Spyros Makridakis, "Forecasting the novel coronavirus COVID-19", *PLoS ONE*, vol.15, number 3, 2020.

[11] K. Leung, and J.T. Wu, and G. M. Leung, "Real-time tracking and prediction of COVID-19 infection using digital proxies of population mobility and mixing", *Nature Communications*, vol. 12, number 1501, 2021.

[12] K. Leung and J.T. Wu and G. M. Leung, "Short-term real-time prediction of total number of reported COVID-19 cases and deaths in South Africa: a data driven approach", *BMC Med Res Methodol*, vol. 21, number 15, 2021.

[13] C. Menni, et al. "Real-time tracking of self-reported symptoms to predict potential COVID-19", *Nature Medicine*, vol. 26, pp. 1037-1040, 2020. https://doi.org/10.1038/s41591-020-0916-2.

[14] Open Knowledge Brasil OPB, *COVID-19 2.0 Transparency Index Report*, howpublished = Available at transparenciacovid19.ok.org.br, 2020.

[15] E. C. Sabino, et al. "Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence", *The Lancet*, vol. 397, issue 10273, pp. 452-455, 2021.

[16] F. A. L. Marson, and M.M. Ortega, "COVID-19 in Brazil", *Pulmonology*, vol. 26, number 4, pp. 241–244, 2020.

[17] J. Mattos, and E. Silva, and P. M. Neto, and Renato Vimieiro, "Clinical risk factors of ICU & fatal COVID-19 cases in Brazil" *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning* pp. 33-40, year 2020.Porto Alegre, RS, Brazil. XDOI 10.5753/kdmile.2020.11956.

[18] M. Tokarnia, "Um em cada 4 brasileiros não tem acesso à internet, mostra pesquisa", *Agência Brasil*, 2020.Available at https://agenciabrasil.ebc.com.br/economia/noticia/2020-04/um-em-cada-quatro-brasileiros-nao-tem-acesso-internet. Accessed January 11, 2022.

[19] L. V. e Silva, et al. "COVID-19 Mortality Underreporting in Brazil: Analysis of Data From Government Internet Portals", *Journal of Medical Internet Research*, vol. 22, number 8, August, 2020.

[20] R. G. da Silva, and M. H. D. M. Ribeiro, and V. C. Mariani, and L. dos S. Coelho, *Chaos, Solitons & Fractals*, vol. 139, October, 2020.

[21] Saulo B. Bastos and Daniel O. Cajueiro, "Modeling and forecasting the early evolution of the COVID-19 pandemic in Brazil", *Scientific Reports*, vol.10, number 1, November, 2020.

[22] I. G. Pereira, and J. M. Guerin, and A. G. S. Júnior, and G. S. Garcia, and P. Piscitelli, and A. Miani, and C. Distante, and L. M. G. Gonçalves, "Forecasting COVID-19 Dynamics

in Brazil: A Data Driven Approach", *International Journal of Environmental Research and Public Health*, vol. 17, number 14, July, 2020.

[23] M. H. D. M. Ribeiro, and R. G. da Silva, and V. C. Mariani, and L. dos S. Coelho, "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil", *Chaos, Solitons & Fractals*, June, vol. 135, 2020.

[24] M. de B. Braga, et al. "Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon", *PLOS ONE*, pp. e0248161, vol. 16, number 3, 2021.

[25] E. Z. Martinez, and D. C. Aragon, and A. A. Nunes, "Short-term forecasting of daily COVID-19 cases in Brazil by using the Holt's model", *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 53, 2020.

[26] E. Schulz, and M. Speekenbrink, and A. Krause, "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions", *Journal of Mathematical Psychology*, vol. 85, pp. 1– 6, 2018. ISSN 0022-2496.

[27] F. Pedregosa, and G. Varoquaux, and A. Gramfort, and V. Michel, and B. Thirion, and O. Grisel, and M. Blondel, and P. Prettenhofer, and R. Weiss, and V. Dubourg, and J. Vanderplas, and A. Passos, and D. Cournapeau, and M. Brucher, and M. Perrot, and Edouard Duchesnay, "Scikit-learn: Machine learning in Python", *The Journal of Machine Learning research*, vol .12, pp. 2825–2830, 2011.

[28] D. Duvenaud, "Automatic Model Construction with Gaussian Processes", Pembroke College, University of Cambridge, 2014.

[29] C. M. Bishop, "Pattern Recognition and Machine Learning", 2006, Springer.

[30] N. S. Gorbach, and A. A. Bian, and B. Fischer, and S. Bauer, and Joachim M Buhmann, "Model selection for Gaussian Process Regression", *German Conference on Pattern Recognition*, pp. 306–318, 2017.

[31] C. Bergmeir, and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation", *Information Sciences*, pp. 192–213, vol. 191, Mayo, 2012.

[32] X. He, et al. "Temporal dynamics in viral shedding and transmissibility of COVID-19", *Nature Medicine*, pp. 672–675, number 5, vol. 26, April, 2020.

[33] F. Hutter, and Lars Kotthoff, and Joaquin Vanschoren, "Automated machine learning: methods, systems, challenges", *Springer*, 2019. Isbn 978-3030053178.

[34] R. S. Olson, and J. H. Moore. "TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning", *Automated Machine Learning*, pp. 151–60, 2019.

[35] A. C. Cameron, and F. A.G. Windmeijer, "An R-squared measure of goodness of fit for some common nonlinear regression models", *Journal of Econometrics*, vol 77, number 2, pp 329-342, 1997.

[36] D. Chicco, and M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation", pp e623, vol. 7, *PeerJ Computer Science*, July, 2021.

[37] M. A. Zöller, and Marco F. Huber. "Benchmark and Survey of Automated Machine Learning Frameworks", *Journal of Artificial Intelligence Research*, pp. 409–472, vol. 70, 2021.

[38] I. Rahimi, and F. Chen, and A, H. Gandomi, "A review on COVID-19 forecasting models", *Neural Computing and Applications*, 2021.

[39] M. C. Castro, and S. Gurzenda, and C. M. Turra, and S. Kim, and T. Andrasfay, and Noreen Goldman, "Reduction in life expectancy in Brazil after COVID-19", *Nature Medicine*, pp. 1629–1635, vol. 27, no 9, June, 2021.

[40] M. Al-Raeei, and M. S. El-Daher, and O. Solieva, "Applying SEIR model without vaccination for COVID-19 in case of the United States, Russia, the United Kingdom, Brazil, France, and India", *Epidemiologic Methods*, vol. 10, february, 2021.

[41] O. P. Neto, et all. "Mathematical model of COVID-19 intervention scenarios for São Paulo—Brazil", *Nature Communications*, vol.12, number 1, january, 2021.

[42] S. Ahmetolan, and A. H. Bilge, and A. Demirci, and A. Peker-Dobie, and Onder Ergonul, "What Can We Estimate From Fatality and Infectious Case Data Using the Susceptible-Infected-Removed (SIR) Model? A Case Study of Covid-19 Pandemic", *Frontiers Media*, vol. 7, 2020.

[43] C. E. Rasmussen, and C. K. I. Williams, "Gaussian Processes for Machine Learning", *Massachusetts Institute of Technology Press*, 2006.

[44] R. M. A. Velásquez, and J. V. M. Lara, "Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression", vol. 136, July, *Chaos, Solitons & Fractals*, 2020.

[45] F. Ahmad, and S. N. Almuayqil, and M. Humayun, and S. Naseem, and W. A. Khan, and K. Junaid, Prediction of COVID-19 Cases using Machine Learning for Effective Public Health Management", *Computers, Materials & Continua*, vol. 66, number 3, pages 2265–2282, 2021.

[46] Y. Alali, and F. Harrou, and Y. Sun, "Optimized Gaussian Process Regression by Bayesian Optimization to Forecast COVID-19 Spread in India and Brazil: A Comparative Study", *2021 International Conference on ICT for Smart Society (ICISS)*, August, 2021. https://doi.org/10.1109/iciss53185.2021.9532501.

Elloá B. Guedes A Computer Science PhD, Guedes is an associate professor at Amazonas State University in Manaus, Amazonas, Brazil. Co-founder of the Intelligent Systems Laboratory, she currently leads the institution's Intelligent Systems Research Group, working on research and development of solutions and applications based on Machine and Deep Learning.

André Igor Nóbrega da Silva Graduated in Electrical Engineering from the Federal University of Campina Grande in 2021. Currently a Msc. student at the Campinas State University Computer Scicence Institute in the area of Information Engineering and Artificial Intelligence.

Paulo R. Lins Júnior Graduated in Electrical Engineering with specialization in Telecommunications from the Federal University of Campina Grande (2006), Master in Electrical Engineering from the Federal University of Campina Grande (2008) and Ph.D. in Electrical Engineering from the same university (2013). Currently, he works as a full professor at the Federal Institute of Education, Science and Technology of Paraíba - IFPB, Campina Grande campus, at Informatics Department, teaching in the higher courses of Telematics Technology and Computer Engineering and acting as a leader and researcher from the Research Group on Communications and Information Processing - GComPI, and as a permanent professor at the Graduate Program in Information Technology - PPGTI, also at the IFPB, with an interest in works involving modeling and performance analysis of communication networks, information processing, analysis techniques, mining and data modeling and machine learning.

Edmar C. Gurjão was born in Campina Grande, Brazil in 1974. Graduated in Electrical Engineering from Universidade Federal da Paraíba (1996), master in Electrical Engineering from Universidade Federal da Paraíba (1999) and PhD in Electrical Engineering from Universidade Federal de Campina Grande (2003). Visiting professor at Notre Dame University (USA) in 2012. Actually is professor of Electrical Engineering Department at Universidade Federal de Campina Grande and in the Master Degree Program in Science and Technology in Health at Universidade Estadual da Paraíba. Experience in Electrical Engineering with emphasis in Compressed Sensing, Software Defined Radio, Cybersecurity, and Signal Processing and its applications. Senior Member of IEEE and member of the Brazilian Society of Telecommunications (SBrT). Co-author of Introduction to Signal and Systems (in Portuguese) 2015, and Digital Signal Processing, Momentum Press, 2018.