Use of econometrics and machine learning models to predict the number of new cases per day of COVID-19

Roberto F. Silva[1], Bruna L. Barreira, Fernando Xavier[1], Antônio M. Saraiva[1], Carlos E. Cugnasca[1]

[1]Departamento de Engenharia de Computação e Sistemas Digitais – Escola Politécnica Universidade de São Paulo (USP) – São Paulo – SP – Brazil

{roberto.fray.silva,brunalyuger,fernando.xavier}@gmail.com, {saraiva, carlos.cugnasca}@usp.br

Abstract. *The COVID-19 pandemics will impact the demand for healthcare severely. It is essential to continually monitor and predict the expected number of new cases for each country. We explored the use of econometrics, machine learning, and ensemble models to predict the number of new cases per day for Brazil, China, Italy, and South Korea. These models can be used to make predictions in the short term, complementing the epidemiological models. Our main findings were: (i) there is no single best model for all countries; (ii) ensembles can, in some instances, improve the results of individual models; and (iii) the ML models had worse results due to the lack of data.*

1. Introduction

Pandemics can be described as diseases that spread on through a large number of people in a short period. This results in a sudden increase in demand for health services, which usually are not prepared for these occasions [Ferguson et al., 2020]. The Sars-Cov-2 or COVID-19 pandemic, which started in 2019 in Wuhan, China, spread worldwide in less than 2 months, and cases can be found in practically all countries in the world.

COVID-19 seriously affects the lungs and the upper respiratory system, leading a percentage of the infected people to demand special care in hospitals. This percentage changes with age: 2.5% for 18-49 years old, 7.4% for 50-64 years old, and more than 12% for people with more than 65 years [Garg, 2020]. These may stay at a hospital bed using respirators for up to 16 days [Ferguson et al., 2020]. Therefore, for this period, no other patient can use these resources, resulting in an increase in the possibility of deaths due to lack of proper care. Allocating patients and resources for pieces of equipment are essential to reduce the impacts of the disease.

Currently, there are no vaccines available for COVID-19, as it is a new disease. Several research projects are being conducted to find possible vaccines, but this is expected to take up to 18 months. Therefore, a group of methods called non-pharmaceutical interventions (NPI) must be adopted, in an attempt to "flatten the curve", or reduce the size of the peak on the demand of hospital beds and specialized equipment such as respirators [Ferguson et al., 2020].

Several epidemiological models, such as the Susceptible Infected Recovered (SIR) model, are used to estimate the curve that describes the evolution of a disease in a specific population [Weiss, 2013]. Those models are essential for long term prediction

and resource allocation, but their results in the short term depend on the assumptions adopted.

Time-series analysis models are data-driven and suited for trend and seasonality detection. For this reason, we believe that they may help decision-makers to make decisions in the short term. These models can complement the results of epidemiological models, with a focus on the short term. Nevertheless, the small size of the dataset in the case of COVID-19 limits the implementation of data-hungry methods, such as convolutional neural networks and long short-term memory networks.

For this reason, we have implemented and compared econometrics and machine learning (ML) models that demand fewer data points to identify and extract patterns. We evaluated four main research questions: *(i) What is the model the best predicts the number of new cases per day of COVID-19 on the following countries: Brazil, China, Italy, and South Korea?; (ii) Is there a model that best describes the number of new cases per day for all those countries?; (iii) Do econometrics models provide a better prediction than ML models?; and (iv) Does the use of ensembles improve the results of those models?*. To the best of our knowledge, this is one of the first attempts to evaluate several models and ensembles for pandemics evolution in different countries.

Therefore, the main objective of this work is to evaluate the use of econometric (ARIMA and SARIMA), ML boosting models that can train on few data (AdaBoost and Gradient Boosting Regressor), and ensemble models for predicting the number of new cases in four countries: Brazil, China, Italy, and South Korea. The econometric models are state of the art models in econometrics, and the ML models are state of the art models in ML boosting models. We have chosen the boosting models because, in general, these demand less data for pattern identification than deep neural networks.

We considered the period from 01/23/2020 (beginning of the disease in China) until 03/22/2020 (when quarantine measures were established in many states in Brazil). We chose to evaluate the prediction of new cases per day, which is one of the most relevant features for decision making in terms of policy-making.

This work is organized as follows: Section 2 describes important works in the healthcare domain related to the ARIMA and SARIMA models, and the AdaBoost and Gradient Boosting Regressor (GBR) models; Section 3 describes the methodology that was adopted; Section 4 contains the main results for each model and discussions on these results; and Section 5 contains the final remarks and concludes the paper.

2. Theoretical foundations

This section describes the main concepts and important works related to: the ARIMA and SARIMA econometric models (Section 2.1), and the AdaBoost and GBR ML models (Section 2.2).

2.1. ARIMA and SARIMA models

The ARIMA and SARIMA models are state of the art models in econometrics, with applications on several domains that involve time series analysis and prediction. They were used in several research pieces on the healthcare domain and provided interesting results, both on their variate and multivariate forms.

According to [Soebiyanto, Adimi, and Kiang, 2010], the ARIMA model identifies trends in time series data using three components: an autoregressive

component, a differencing component, and a moving average component. It has three hyperparameters: p (autoregressive order), d (differencing order), and q (moving average order). Its notation is ARIMA(p,d,q). The SARIMA is a version of the ARIMA model that contains seasonality components for each of the orders (P, D, and Q), as well as a seasonal order (S). Its notation is SARIMA(p,d,q,P,D,Q,S).

The ARIMA models have been used for evaluating the Influenza epidemics for some time, as illustrated by the research by [Domínguez et al., 1996]. These authors investigated the weekly number of cases and deaths in six towns in the Barcelona province using several univariate ARIMA models configurations. They have concluded that the models provided insights for detecting the epidemic activity of the disease. Nevertheless, more information is needed for deciding on control measures.

[Soebiyanto, Adimi, and Kiang, 2010] evaluated the impact of climatological features on the prediction of seasonal transmission of Influenza on warm regions (Hong Kong, and Maricopa, Arizona, USA). These authors evaluated the use of ARIMA, ARIMAX, SARIMA, and SARIMAX models for weekly prediction of confirmed cases in each region. The ARIMAX and SARIMAX models also included environmental features, besides the confirmed cases time series.

[Soebiyanto, Adimi, and Kiang, 2010] have concluded on their work that: (i) the ARIMAX had the best result for Hong Kong; (ii) the SARIMAX had the best result for Maricopa; (iii) the predictions for 1-step ahead were satisfactory; and (iv) a prediction of more than one step could provide more value for decision making.

[Proprou, Jaroensutasinee, and Jaroensutasinee, 2006] used several configurations of the univariate ARIMA model to model and predict the monthly dengue hemorrhagic fever cases in southern Thailand in the months between January and August 2006. The models were trained on data from 1994 to 2005, and their main results were: (i) ARIMA models had satisfactory results; and (ii) the results have the potential to provide insights for policy-making.

[Han et al., 2011] analyzed the increase of narcolepsy in China after the 2009 H1N1 winter Influenza epidemic, utilizing linear and ARIMA models. They have identified that the epidemic led to an increase of 3 times in the number of reported cases, and discussed how the disease could influence the increase in narcolepsy.

In this section, several research pieces that used ARIMA and SARIMA models were presented, demonstrating the usefulness of this technique for the analysis of disease time series. The main contributions of our work for this area are to study both models for datasets that are considerably smaller than those in the literature and compare them with ML boosting models. In the next section, we describe some of the important research pieces in the literature using ML models.

2.2. AdaBoost and GBR models

The AdaBoost and GBR models are state of the art models in regression tasks on several domains. In the following paragraphs, we analyze important works that use ML for analyzing disease outbreaks, as well as the growing trend of using ML models due to the abundance of data, powerful hardware, and advanced techniques and models.

The work by [Nilashi et al., 2017] presented a complex model for disease prediction using an ensemble of classification and regression decision trees (CART), several preprocessing techniques, and a fuzzy rule-based model. It was composed of

four components: (i) use of expectation-maximization clustering on the dataset; (ii) use of principal component analysis for reducing dimensionality; (iii) use of CART for identifying patterns; and (iv) use of a fuzzy rule-based model to make the prediction. The authors observed improved results on the MAE and R2 on most datasets analyzed.

An ensemble can be defined as a model that gathers several predictors' results and makes a final prediction [Satillana et al., 2015]. We implemented simple ensembles that combine the results of the individual models with equal weight on the final prediction. The primary rationale behind using ensembles is that, as the models focus on different patterns on the data, combining multiple models' results could improve the overall prediction result [Satillana et al., 2015].

AdaBoost and GBR are models that use the boosting technique, which can be defined as an ensemble that combines the results of several models in a sequential manner [Satillana et al., 2015]. The main advantage of those models is the possibility of reducing the bias of the individual models [Shafaf and Malek, 2019].

According to [Satillana et al., 2015], the AdaBoost model has the advantage of being able to learn local rules in the data. This can also be observed for other decision trees ensembles, such as GBR. Nevertheless, the models need more data for training compared to models such as ARIMA and SARIMA and may present fitting problems.

[Satillana et al., 2015] proposed a model that uses AdaBoost to forecast weekly Influenza cases in the USA based on several datasets: searches on Google, tweets, hospital visit records, and a participatory surveillance system. This model was compared with stacked linear regression and support vector regression. One important finding was that the proposed ensemble model outperformed all the individual models. The authors observed that the AdaBoost model showed the lowest RMSE and MAPE and presented reasonable predictions for up to three-week forecasts.

We refer the reader to the work by [Shafaf and Malek, 2019] for an in-depth analysis of ML applications for emergency medicine. According to these authors, ML can improve the prediction and early detection of diseases. These models can help decision-makers to obtain insights related to disease progression. Three out of the twenty works reviewed by these authors used ensembles of decision trees. Two of them used Gradient Boosting for classification, with satisfactory results.

[Zhang et al., 2019] compared several models to predict blood pressure rates based on several physiological data. They have compared ridge and lasso regressions, elasticnet, support vector regression, k-nearest neighbor, and the GBR models. Their results show that the GBR had better accuracy for calculating diastolic (64%) and systolic (70%) pressures, while also being considerably fast (0.1s).

3. Methodology

The methodology used in this paper was composed of seven steps:

1. Data gathering from official databases, for the period from 01/23/2020 until 03/22/2020. We gathered data for the following countries: Brazil[1], South Korea[2], China[3], and Italy[4]. The data gathered contained the following features: total number of

1 https://www.kaggle.com/unanimad/corona-virus-brazil
2 https://www.kaggle.com/kimjihoo/coronavirusdataset#Case.csv
3 https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
4 https://www.kaggle.com/sudalairajkumar/covid19-in-italy#covid19_italy_region.csv

cases, the total number of deaths, new cases, new deaths in the day, and recovered patients. For Brazil, we gathered data for the number of recovered patients from the Johns Hopkins University CSSE [Dong, Du, and Gardner, 2019][5]. Additionally, we used the total population per country from the World Bank database[6] to calculate the health indicators;

2. Data preprocessing. We preprocessed the datasets (one for each country) to remove missing data and identify possible outliers. After several experiments considering different treatments for outliers, we chose to consider the whole datasets without removing any data points. The removal of data points resulted in worse results, mainly because the dataset is considerably small, and removing data points would make it harder for the models to identify patterns and trends. We then calculated the following health indicators for all datasets: prevalence, incidence, and death rate. We separated the datasets into three subsets: train (81% of the full dataset), validation (9% of the full dataset), and test (10% of the full dataset);

3. Exploratory data analysis. We conducted an exploratory data analysis to evaluate: (i) the autocorrelation in the data, using the PACF and ACF plots, as well as the Dickey-Fuller test; and (ii) the behavior of each feature in the time series, considering statistical measures (mean, mode, standard deviation) and visualizations, such as box plots and line plots;

4. Implementation of the econometrics models and hyperparameters analysis. We have implemented two econometrics models: ARIMA and SARIMA, using the SARIMAX class from the statsmodels library. For both models, only one time series can be considered. Therefore, both the inputs and outputs were the absolute new cases feature. We used the train and validation subsets for hyperparameters analysis and final model training, and the third subset for model evaluation. We then implemented the models for each dataset. The hyperparameters evaluated for the ARIMA were the autoregressive (p), differentiation (d), and the moving average (q) components. The hyperparameters evaluated for the SARIMA were the same as for the ARIMA plus the seasonal components (P, D, Q, and S). We used a grid search to find the best models, with values for each component from 0 to 3, which were defined experimentally. The metric used for evaluation was the mean absolute error (MAE);

5. Implementation of the ML models and hyperparameters analysis. We have implemented two ML boosting models: AdaBoost and GBR, using the scikit-learn library. Unlike the models implemented in Step 4, these models consider all the features in the dataset as inputs. Therefore, we expected that they would capture more information and, therefore, provide a better prediction. As in Step 4, we used the train and validation subsets for hyperparameters analysis and final model training, and the third subset for model evaluation. We implemented the models for each dataset. The hyperparameters chosen for the analysis, based on several experiments, were: (i) for the AdaBoost, the learning rate (2, 5, and 10), loss function (linear, square, and exponential), and number of estimators (5, 10, and 15); and (ii) for the GBR, the learning rate (0.02, 0.05, and 0.1), max_depth (2 and 5), number of estimators (10, 20, and 30). We used a grid search to find the best models. The MAE of the prediction was used for evaluation;

6. Implementation of ensemble models. Three ensemble models were

5 https://github.com/CSSEGISandData/COVID-19
6 http://wdi.worldbank.org/table/2.1#

implemented: (i) econometrics models ensemble; (ii) ML models ensemble; and (iii) an ensemble of all models. The first one considered an average of the predictions of the ARIMA and SARIMA models as its prediction for each data point. The second one considered the average of the predictions of the AdaBoost and GBR models as its prediction. The third one considered an average of the predictions of all models as its prediction;

7. Models comparison. The final models from Steps 4, 5, and 6 were evaluated for each dataset to find the most suitable model for each country, based on the MAE on the test subsets;

We used the following Python libraries: Pandas, Statsmodels, Scikit-Learn, Matplotlib, NumPy, and Seaborn. The implementation was done using a Google Colab TPU. The datasets and the code are available on an open Github repository[7].

4. Results and discussion

This section describes the main results of the research and is divided into four subsections: 4.1 contains the exploratory data analysis; 4.2 contains the results of the econometrics models and their ensemble; 4.3 contains the results of the ML models and their ensembles; 4.4 contains the results of the comparison of all models implemented.

4.1. Exploratory data analysis

After preprocessing the data, we have conducted three analyses: (i) an analysis of each feature of the dataset, focusing on their behavior on the different datasets; (ii) an analysis of the autocorrelation of the new cases feature for each dataset; and (iii) an analysis of stationarity of the new cases feature for each dataset, using the Augmented Dickey-Fuller (ADF) test.

Figure 1 illustrates the charts of the total cases and new cases per country (on the left side) and the ACF plots for each country (on the right side). We decided to analyze the number of new cases per day for two reasons: (i) it is the recommendation for time series analysis that are not stationary, that they should be differentiated; and (ii) a prediction of potential new cases for the next period could provide valuable information for decision-makers. The first point is vital as most of the econometrics models (such as ARIMA and SARIMA) are designed for stationary data, as these time series can, in theory, be predicted.

All datasets analyzed present difficulties for time series analysis models: (i) the datasets for Brazil and Italy are very small, containing 27 data points each; (ii) the new cases for Italy is growing fast and changing its shape; (iii) the dataset for China has an outlier that is more than 10x higher than the mean value (it is not a wrong value); and (iv) the methods used by the countries to measure the number of infected people vary.

As the healthcare indicators (incidence, prevalence, and death rate) all depend on the raw data to be calculated, they correlate with these raw features. Nevertheless, in this work, we decided to use all indicators for training the ML models, as they might provide additional information for identifying patterns on the data. We also observed that, in general, these models had good results on predicting death rate, as it does not change drastically from day to day, as is the case of new cases per day.

7 Link for the GitHub open repository

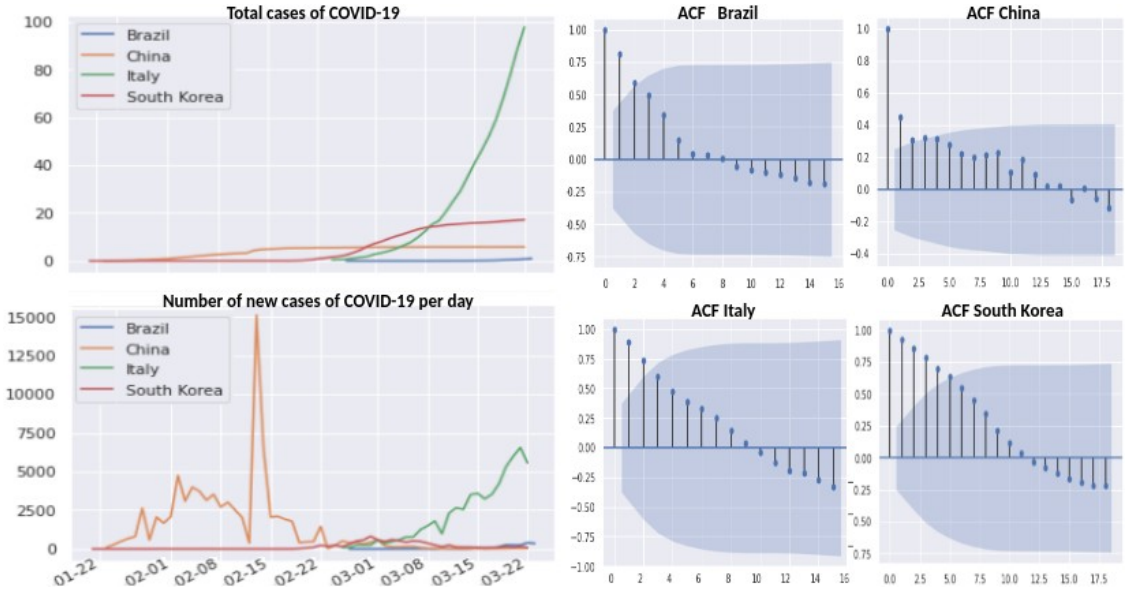

Figure 1. On the left side: (i) top: total cases of COVID-19; (ii) bottom: number of new cases per day. On the right side (using new cases feature): (i) top left: ACF for Brazil; (ii) top right: ACF for China; (iii) bottom left: ACF for Italy; (iv) bottom right: ACF for South Korea.

The second analysis that was conducted was related to the autocorrelation of the new cases feature for each dataset. The ACF plots for all countries is illustrated on the right side of Figure 1. All plots show a degree of autocorrelation. This was expected according to the case of a disease spreading on a population.

Lastly, the ADF test also indicated that all of the countries had a non-stationary series for new cases per day. For this reason, especially for the econometrics models, the series must be differentiated until they present a stationary behavior. The next Section contains the results of the implementation of these models.

4.2. Econometrics models

It is important to note that the main difference between the ARIMA and SARIMA models is the estimation, by the second one, of the effect of seasonality on the data. Even though the analysis conducted on Section 4.1 did not show direct seasonality effects, we believe it is essential to test both types of models, as some seasonal effects may be identified on the larger datasets (China and South Korea).

Table 1 contains the values of hyperparameters that originated the best models for the ARIMA and SARIMA techniques, as well as their ensemble. The ARIMA model provided the best results for Italy (MAE: 682.37), while the SARIMA provided the best results for Brazil (MAE: 163.53). Nevertheless, it is interesting to note that, for both China and South Korea, the ensemble resulted in a lower MAE.

As for the values of the hyperparameters for the ARIMA model, the results were diverse. Nevertheless, none of the best models used zero for its p or q values. As for the best SARIMA models, all identified a seasonal component. Except for Italy, the seasonal component identified by the model improved the MAE of the final ensemble. Therefore, we conclude that: (i) seasonal models should be tested in similar problems with a small dataset, even when they do not present a seasonality that can be detected

by the traditional seasonal decomposition methods; and (ii) model tuning should be conducted for each dataset.

Table 1. Variables to be considered on the evaluation of interaction techniques. The best values are in bold.

Country	ARIMA		SARIMA		Ensemble ARIMA + SARIMA
	Hyperp. values (p,d,q)	MAE	Hyperp. values (p,d,q,P,D,Q,S)	MAE	
Brazil	2,0,2	280.03	**2,2,2,1,2,1,2**	**163.53**	169.24
China	1,2,1	45.70	2,2,1,2,2,1,1	54.44	**21.50**
Italy	**1,1,2**	**682.37**	1,2,1,1,1,2,2	744.92	1285.47
South Korea	2,0,1	53.34	2,0,2,2,0,1,1	34.71	**26.43**

We believe that the main factor impacting the high MAE for Italy was its fast increase in the number of confirmed cases. This increases the difficulty of the models in identifying trends and can be explained by two main factors: (i) the phenomenon of sub notification of cases and wrong diagnoses at the beginning of the disease spread; and (ii) the lack of testing to estimate the real number of infected people. In the next Section, we analyzed the results for the ML models.

4.3. ML models

As was described in Section 3, we used all features as inputs for the ML models implemented in this work. The primary rationale behind this choice was to try to improve the amount of information captured by the model, resulting in better predictions.

Table 2 contains values of hyperparameters for the best models for the AdaBoost and GBR techniques, as well as their ensemble. For the AdaBoost models, the linear loss and a learning rate of 2 resulted in the best models. As for the number of estimators, these varied by country. In the case of the GBR, all of the best models presented max features of 1. Two of them used a learning rate of 0.1, and two of them used 10 estimators. Therefore, we can conclude that the model tuning should be conducted for every dataset, as was observed for the models in Section 4.2.

The AdaBoost model resulted in the lowest MAE for Brazil (MAE: 51.66) and Italy (MAE: 1748), and the use of GBR provided the best results for China (MAE: 186.19) and South Korea (MAE: 41.98). An essential difference with Section 4.2 is that none of the ensembles resulted in a lower MAE than the best individual model. Further research is being conducted to understand the causes of these results better, as they may provide helpful insights for other time series analysis problems with small datasets.

We believe that one of the following two reasons may be the primary explanation for these results: (i) the patterns identified by both models may be more similar than the ones identified by the ARIMA and SARIMA models; or (ii) the GBR may be identifying patterns that are not relevant, due to the small dataset. Nevertheless, the use of ensembles may still provide interesting insights on the predictions (and even improve them), as was observed in Section 4.2. The use of a larger dataset could lead to

this improvement, as these models are data-hungry.

Table 2. Results on the test subset of the AdaBoost, GBR, and ensemble of AdaBoost and GBR models for each country. The best values are in bold.

Country	AdaBoost		GBR		Ensemble AdaBoost +GBR
	Hyperp. values (lr, loss, n_e)	MAE	Hyperp. values (lr, max_f, n_e)	MAE	
Brazil	**2, linear, 5**	**51.66**	0.01, 1, 10	197.48	63.52
China	2, linear, 15	344.33	**0.1, 1, 30**	**186.19**	274.60
Italy	**10, square, 5**	**1748.00**	0.05, 1, 10	4101.95	3315.33
South Korea	2, exp., 10	55.14	**0.1, 1, 20**	**41.98**	57.90

As was observed in Section 4.2, the high MAE for Italy may be due to the fast increase in the number of confirmed cases. This impacted more heavily on the ML models analyzed, as these need more data than the econometrics models analyzed to identify patterns. In the next Section, we conducted a comparison of all the models implemented.

4.4. Models comparison

In this section, we address the four research questions. Table 3 contains the results of all models on the test subset for each country. To answer the first question, we can infer that the best models for predicting the number of new cases per day, among the ones analyzed, are: for Brazil, the AdaBoost (MAE: 51.66); for China, the ensemble of econometrics models (MAE: 21.50); for Italy, the ARIMA (MAE: 682.37); and for South Korea, the SARIMA (MAE: 34.71).

Table 3. Results of all models on the test subset for each country. The best values are in bold.

Country	Econometrics models			ML models			Ensemble of all models
	ARIMA	SARIMA	Ensemble	AdaBoost	GBR	Ensemble	
Brazil	280.03	163.53	169.24	**51.66**	197.48	63.52	65.67
China	45.70	54.44	**21.50**	344.33	186.19	274.60	146.40
Italy	**682.37**	744.92	1285.47	1748.00	4101.95	3315.33	1897.51
South Korea	53.34	34.71	**26.43**	55.14	41.98	57.90	41.24
Average	265.36	**249.40**	375.66	549.79	1131.90	927.84	537.71

Answering the second question, there was no single model that obtained the best

results across the four datasets. This may be due to several reasons: (i) the dynamics of the disease spread in each country, impacted by different behaviors, Government measures, population density, population age structure, and HDI, among others; (ii) the strategy for testing adopted by the country, which differed considerably among the analyzed countries; (iii) the stage of disease spread in each country, which directly affects the rate of increase of infected people; (iv) sub notification or wrong diagnosis, which may differ between the countries; among others.

To answer the third question, econometrics models were able to better capture the trends and predict the number of new cases per day for China, Italy, and South Korea. It is interesting to observe that the econometrics models ensemble better suited the data for China (MAE: 21.50) and South Korea (MAE: 26.43), while the ARIMA model better suited the data for Italy (MAE: 682.37). Nevertheless, as Italy's data increased considerably in the period considered, all models had higher MAEs than for the other countries.

Based on these results, we can infer that for small datasets with higher variability (such as in the case of China and Italy, as described in Section 4.1), econometrics models should obtain better results. Considering the simple average of the models' MAEs across all countries, the SARIMA model provided the best results (MAE: 249.40). On average, all three econometrics models (ARIMA, SARIMA, and their ensemble) had better results than the ML models and the ensemble of all models. The model that performed the worst was the GBR (MAE: 1131.90), while the AdaBoost was the best model only for Brazil (MAE: 51.66).

This observation can be mainly explained by the following factors: (i) the ML models, as described in Section 2.3, demand more data for identifying patterns; (ii) these methods were not designed to deal specifically with time-series data; and (iii) the features used (besides the new cases per day time series) may have introduced noise on the pattern detection.

To answer the fourth research question of this paper, we can infer that the use of ensembles may improve the results compared to using the models independently, as was observed for China and South Korea. The main factor that explains this observation is that, as the models identify different patterns, an average of their predictions may improve the overall results. This is the basis of several implementations of ensemble models. Therefore, we recommend, for similar prediction problems, to evaluate both the models independently and ensembles of these models.

Although many more experiments are needed to be able to conclude which model is the best for the different scenarios, we can infer that the more the models differ among themselves on the method they use for identifying patterns and making predictions, the more significant is the potential for ensembles to provide better predictions than the individual models. An exception would be the case in which some models have significantly worse results than others, as was observed for the ML models for China and Italy.

It is also important to note that, to further improve the quality of the models' predictions, research is being conducted to consider other variables such as: the number of tests per country (and how these may influence on the results), the number of doctors in relation to the population of the countries, the number of intensive care unit beds available over time, and several other important socio-economic variables, such as access to health services and access to basic sanitation.

The main contributions of this work are: (i) to evaluate the different state of the art models for time series analysis in small datasets; (ii) to compare econometrics and ML models for predictions pandemics spread; and (iii) to evaluate the use of ensembles on small datasets. We believe that our results can be useful for both researchers to identify new exciting areas to improve predictions of disease spread and for practitioners, to consider additional models to aid in their decision-making processes.

The main limitations of this work were: (i) the lack of available data, as it was conducted at the beginning of the pandemics spread; (ii) difficulty on comparing the data from different countries, as the methods used to estimate the number of infected people vary by country (for example, South Korea uses extensive testing, while Brazil only tests people that enter the hospitals with the disease's symptoms); and (iii) the lack of knowledge of the dynamics of the disease spread on different environments and countries (such as Brazilian slums, which may have a much larger spread than other environments).

We believe that these limitations are in place whenever a new disease starts spreading, so we encourage other researchers to continue exploring econometrics and ML models, intending to provide additional useful information for decision-makers in the short term.

5. Conclusions

In this paper, we analyzed different econometrics and ML boosting models for predicting the number of new cases of COVID-19 per day for four different countries. We also evaluated: (i) the potential for using econometrics and ML models on this task; (ii) the potential of using ensembles of econometrics and ML models, as well as of all models implemented; and (iii) which model resulted in the lowest MAE for each country.

Our main findings were: (i) there is no single model that better predicts the number of new cases per day for all countries; (ii) there is potential for using ensembles, especially of econometrics models; (iii) as ML models are more data-hungry than the econometrics models, they had worse results, probably due to the small size of the datasets; and (iv) that the implemented models could be used to provide additional information for decision-makers.

The same methodology can be applied for other diseases, especially if they are in their initial spreading period, in which the dynamics of the disease are only partially known. We believe that the use of time series analysis models could provide valuable information on the possible behavior of the features in the short term, complementing more long term epidemiological prediction models.

Future works are related to: (i) re-evaluating the models with larger datasets, which will improve the quality of the predictions; (ii) implementing supervised ML models that are state of the art for time series analysis, such as convolutional neural networks and long short-term memory networks, with a larger dataset; (iii) analyzing the use of unsupervised ML models for clustering the countries with similar characteristics, considering both healthcare indicators, data from the disease spread, population density, and HDI; and (iv) evaluating the use of the predictions of the analyzed models with the epidemiological models for improving decision-making.

Acknowledgements

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001, Itaú Unibanco S.A. through the Itaú Scholarship Program, at the Centro de Ciência de Dados (C2D), Universidade de São Paulo, Brazil, and also by the National Council for Scientific and Technological Development (CNPq).

References

Dominguez, A., Muñoz, P., Martínez, A., Orcau, A. (1996) "Monitoring mortality as an indicator of influenza in Catalonia, Spain". *Journal of Epidemiology & Community Health*, v.50, n.3, p.293-298.

Dong, E., Du, H., Gardner, L. (2020) "An interactive web-based dashboard to track COVID-19 in real time". *The Lancet Infectious Diseases, Correspondence*, p.1-2.

Ferguson, N. et al. (2020) "*Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand*", 2020, p.1-20.

Garg, S. (2020) "Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—COVID-NET, 14 states, March 1–30". *Morbidity and Mortality Weekly Report*, v.69. Available on: https://www.cdc.gov/mmwr/volumes/69/wr/mm6915e3.htm. Accessed on: 04/24/2020.

Han, F. et al. (2011) "Narcolepsy onset is seasonal and increased following the 2009 H1N1 pandemic in China". *Annals of neurology*, v.70, n.3, p.410-417.

Nilashi, M., bin Ibrahim, O., Ahmadi, H., Shahmoradi, L. (2017) "An analytical method for diseases prediction using machine learning techniques". *Computers & Chemical Engineering*, v.106, p.212-223.

Promprou, S., Jaroensutasinee, M., Jaroensutasinee, K. (2006) "Forecasting Dengue haemorrhagic fever cases in Southern Thailand using ARIMA models". *Dengue Bulletin*, v.30, p.99-106.

Santillana, M., Nguyen, A.T., Dredze, M., Paul, M.J., Nsoesie, E.O., Brownstein, J.S. (2015) "Combining search, social media, and traditional data sources to improve influenza surveillance". *PLoS computational biology*, v.11, n.10, p.1-15.

Shafaf, N., Malek, H. (2019) "Applications of Machine Learning Approaches in Emergency Medicine; a Review Article". *Archives of academic emergency medicine*, v.7, n.1, p.1-9.

Soebiyanto, R.P., Adimi, F., Kiang, R.K. (2010) "Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters". *PloS one*, v.5, n.3 p.1-10.

Weiss, H.H. (2013) "The SIR model and the foundations of public health". *Materials matematics*, n.3, p.1-17.

Zhang, B., Ren, J., Cheng, Y., Wang, B., Wei, Z. (2019) "Health data driven on continuous blood pressure prediction based on gradient boosting decision tree algorithm". *IEEE Access*, v.7, p.32423-32433.