Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Multivariate data driven prediction of COVID-19 dynamics: Towards new results with temperature, humidity and air quality data

Dunfrey P. Aragão, Emerson V. Oliveira, Arthur A. Bezerra, Davi H. dos Santos, Andouglas G. da Silva Junior, Igor G. Pereira, Prisco Piscitelli, Alessandro Miani, Cosimo Distante, Jordan S. Cuno, Aura Conci, Luiz M.G. Gonçalves

Abstract

Since the start of the COVID-19 pandemic many studies investigated the correlation between climate variables such as air quality, humidity and temperature and the lethality of COVID-19 around the world. In this work we investigate the use of climate variables, as additional features to train a data-driven multivariate forecast model to predict the short-term expected number of COVID-19 deaths in Brazilian states and major cities. The main idea is that by adding these climate features as inputs to the training of data-driven models, the predictive performance improves when compared to equivalent single input models. We use a Stacked LSTM as the network architecture for both the multivariate and univariate model. We compare both approaches by training forecast models for the COVID-19 deaths time series of the city of São Paulo. In addition, we present a previous analysis based on grouping K-means on AQI curves. The results produced will allow achieving the application of transfer learning, once a locality is eventually added to the task, regressing out using a model based on the cluster of similarities in the AQI curve. The experiments show that the best multivariate model is more skilled than the best standard data-driven univariate model that we could find, using as evaluation metrics the average fitting error, average forecast error, and the profile of the accumulated deaths for the forecast. These results show that by adding more useful features as input to a multivariate approach could further improve the quality of the prediction models.

1. Introduction

After its first case reported at the end of 2019 (Guan et al., 2020) the SARS-CoV-2 virus has rapidly spread around the world, causing a serious acute respiratory syndrome that is simply referred as COVID-19. Currently, it is known that most of the human-to-human transmission of COVID-19 occurs through contact, either by touching a contaminated region and bringing the hands in contact with the eyes, nose, or mouth, and/or through contact with suspended aerosol particles in the air surrounding contaminated people, that are expelled through coughing, sneezing and even talking (Who - coronavirus disease, 2020).

The number of new daily cases and deaths caused by COVID-19 increased in some countries around the world at the end of 2020 and beginning of 2021, with a significant higher number of new daily deaths from COVID-19 compared to the first wave of COVID-19 in some locations. This can be clearly observed in Fig. 1, which shows the higher number of weekly deaths from COVID-19 in Europe, North America and Africa, which starts around 2020 October 1st.

To forecast the behavior of COVID-19 pandemic, approaches such as SIR, SEIR, SIRASD epidemiological models and machine learning methods as Long Short-Term Memory (LSTM) and Modified Auto-Encoder (MAE) have been compared in previous work (Pereira et al., 2020). Clustering the countries by some features prior to training a set of models, one for each cluster, seems to prevent mistakes when designing...
Environmental Research 204 (2022) 112348

D.P. Aragão et al.

2

Fig. 1. COVID-19 deaths in Europe, North America and Africa - Image extracted from https://ourworldindata.org/.

a prediction model since it is based on dynamics of similar countries. Moreover, this previous study shows that the MAE method is more efficient and produced better results for the COVID-19 epidemiology than the LSTM method and other traditional methods, besides it requires more features in order to get improved results. Actually, there are other works presenting model comparisons, as using a chaotic marine predator algorithm (CMPA) to improve the performance of the adaptive neuro-fuzzy inference system (ANFIS) and to apply its COVID-19 prediction method to confirmed cases using official World Health Organization (WHO) data sets for Russia and Brazil (Al-qaness et al., 2021). A final but not last study (Friedman et al., 2021) compares 386 public global forecast models for COVID-19 mortality. The idea is to help authorities to select the best forecast models so far, since there are many public websites with this information and the predictions can be significantly disparate. Most of the models that passed the selection criteria proposed by the authors used the traditional univariate epidemiological models such as SIR and SEIRD.

The several efforts to predict the spread of the disease in different countries, using mathematical, statistical, and data-driven based approaches, help to understand the epidemiological characteristics and policies to be adopted to reduce the number of deaths caused by the disease and slow the spread of the COVID-19 pandemic. In South America, for example, the number of daily deaths in the second wave, as of 2021 March 20th, were much higher than in the first wave. Moreover, there was an attempt to reduce the number of daily deaths in Asian countries, while Oceania (Who - coronavirus disease, 2020) has it partially controlled (Fig. 2).

One research direction regards the effects that air pollution has on COVID-19 death toll and cases, since it is known that air pollution can cause several disorders, including chronic respiratory diseases, stroke and cardiovascular problems. Since the beginning of the pandemic, several studies investigate the hypothesis that the effects of air pollution could increase death counts and transmissibility of COVID-19 (Travaglio et al., 2021). Two consequences may arise from the effects of a polluted air: either increasing the transmissibility of the virus (short-term) or causing health conditions that increase the lethality of COVID-19 (long-term) (Barcelo, 2020). Hence, it should be considered the hypothesis that it is possible to correlate pollution, temperature, and humidity with COVID-19 deaths, as verified next.

In the first phases of the current SARS-CoV-2 pandemic, multiple studies suggested that outdoor airborne transmission could play an important role during COVID-19 outbreak in northern Italy (Conticini et al., 2020; Setti et al., 2020). The main hypothesis was that virus-laden aerosol could interact with atmospheric particles creating clusters with pre-existing particles acting as carriers, enhancing the persistence of the virus in atmosphere (Carducci et al., 2011). This lead to the suggestion that atmospheric particulate matter concentrations is a kind of proxy to track virus dispersion in the atmosphere. There are not, up to now, specific data on the interaction of SARS-CoV-2 with pre-existing particles. However, it is known that atmospheric aerosols could contain biological material (bacteria and viruses) in certain conditions (Verreault et al., 2008; Després et al., 2012) and that the interaction between viruses and atmospheric particles could influence (increasing or decreasing) their infectivity (Groulx et al., 2018). Other studies found that concentration and size distribution of both virus-laden aerosol and pre-existing particles strongly influence the probability of interaction (Belosi et al., 2021). Hence, some work on studying filters for purifying the air for indoor environments are also found, specifically for reducing aerosol particulate matter (PM) and volatile organic compounds (VOCs) (Fermo et al., 2021).

On the long-term scenario, we highlight three works (Travaglio et al., 2021; Comunian et al., 2020; Konstantinoudis et al., 2021) showing a positive correlation between air pollution (mostly Nitrogen dioxide (NO2) and Particulate Matter 2.5 μm (PM2.5)) and number of deaths and cases. Conversely, other works show short-term negative correlation for specific pollutants and region due to the improvement of air quality in mid 2020 as a result of mitigation measures early in the pandemic (Zangari et al., 2020; Naqvi et al., 2021; Rodríguez-Urrego and Rodríguez-Urrego, 2020). On the short-term scenario, a few studies try to relate the air pollution with the spread of COVID-19, since it is known that virus and bacteria can survive longer in the air when they get together with particle matter, specifically Particulate Matter of 10 μm (PM10) and PM2.5 (Verreault et al., 2008; Després et al., 2012; Groulx

Fig. 2. Weekly confirmed COVID-19 deaths in populated continents - Image extracted from https://ourworldindata.org/.
The relationship between temperature for new daily cases and new COVID-19 deaths is also highlighted in some studies as another weather impact. A thoroughly discussion on the effects of meteorological parameters over COVID-19 spread has been done, such as to analyze the effects of temperature changes on 166 countries and its relationship with COVID-19 (Wu et al., 2020). In addition, studies on the effects of weather conditions indicate that these characteristics could affect the spread of COVID-19 in the local territory (Brasey et al., 2020). The relationship between temperature and humidity, which may include a significant change in COVID-19 transmission, and the most appropriate temperature for viral transmission and resistance on contaminated surfaces and the duration of the presence of the virus, is also examined in other works (Ma et al., 2016; Wang et al., 2020; Chan et al., 2011; Diao et al., 2016).

Nevertheless, a case study in Brazil shows that temperatures and humidity are highly correlated to the spread of COVID-19 (Auler et al., 2016). It is noticed that the tropical climate may be favourable to the COVID-19 transmission, as occurs in most Brazilian cities. Furthermore, a work based on Brazilian cities (Prata et al., 2016) explores the linear and nonlinear relationship between the average temperature and confirmed cases as reported by the use of a generalized additive model (GAM) that shows a negative linear relationship with the number of confirmed cases and a flattened curve spreading at around 25.8 degrees Celsius. Also, the impact of meteorological conditions across the 27 capitals of Brazilian provinces was analyzed in the first month of the disease outbreak. The presented results take into account temperature, population density and flights as features for COVID-19 forecasts (Peqorno et al., 2020).

Hence, as our main contributions here, by taking into consideration the results of the above works, this paper first tries to confirm that environmental variables plays a role on the COVID-19 spreading, verifying some evidence that there is a correlation with COVID-19 deaths in the most affected Brazilian cities. Further, we use environmental features such as AQI (Air Quality Index), temperature, and humidity as input to a multivariate predictive model in an attempt to generate better predictions. We compare the performance of the best data-driven univariate model and the best multivariate model that we could find over several training attempts. The idea of our current work is to use the K-means clusters to identify similar AQI curves from different countries. This allows us to use a based transfer learning approach across each cluster of our model. Actually, in this work, we apply the LSTM method together with the clustering approach can improve our previous data driven approach, providing a more accurate forecast, which is currently not limited only to daily death cases as done in our previous work (Pereira et al., 2020).

2. Materials and methods

Different scenarios could be adopted regarding the granularity level form the geographic perspective (communities, cities, provinces, countries etc). A more granular approach yields better reliability and less variability for each location in spite of higher computational effort and data availability.

On the other hand, a less granular approach, such as adopting countries’ data, generalizes information through the use of uncertainty metrics such as average, maximum and minimum. The main advantage of such less granular approach regards in a richer data availability available in global datasets. However, we emphasize that, depending on a country’s size, information such as air quality may not adequately represent the country’s overall behavior.

Nonetheless, to demonstrate the feasibility and validity of this study, we conduct a primary investigation using the less granular approach. The first step is to acknowledge how country data perform and cluster based on similarities. Methods like K-means are an instrument for investigating dimensions and generating similarity clusters among assembly members. Following that, a smaller granularity analysis should be carried out.

2.1. COVID-19 deaths dataset

On the global perspective, we use the COVID-19 dataset available at the World Health Organization (WHO) website (Who - coronavirus disease, 2020). Similarly and consistently with the WHO dataset, we use data collected by Wesley Cota (Cota, 1590) to carry out the analysis for Brazilian cities and states. Both datasets starts at January 22, 2020.

The WHO dataset comprises 274 different localities, including countries and major cities. China, for example, was divided into 33 provinces, which were added to the aforementioned number of localities. On these cases, we sum the number of cases and deaths of the provinces to achieve a set of unique values for the whole country. We apply the aforementioned process to the following countries: Australia, Canada, China, Denmark, France, the Netherlands, and the United Kingdom, resulting in a total number of 192 localities.

2.2. Air quality index (AQI)

The AQI is an index that describes the status of air quality. It is a dimensionless value that takes into account six criteria pollutants: PM<sub>2.5</sub>, PM<sub>10</sub>, SO<sub>2</sub>, CO, NO<sub>2</sub> and O<sub>3</sub>. The computation is performed based on the Equation (1).

\[
AQI_P = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C_{wp} - C_{low}) + I_{low}
\]  

where, \(AQI_P\) is the air quality index for each pollutant; \(I_{high}\) and \(I_{low}\) are the high and low indexes breakpoints; \(C_{high}\) and \(C_{low}\) are the high and low concentration breakpoints; and \(C_{wp}\) is the measured concentration of the pollutant. Having the AQI for each pollutant computed, the maximum value is selected as the final index value. Mathematically,

\[
AQI = \max(AQI_{p1}, AQI_{p2}, AQI_{p3}, ..., AQI_{pn})
\]  

It is important to point out that each country has its own definition of the AQI. This includes choosing only some air quality parameters (for example, only CO and NO<sub>2</sub>) and changing the ranges of safety margins for these parameters to be used in the AQI Equation (1). In our approach, we have used the average of the AQI parameters available, for each country and date. This was necessary because countries have neither the same AQI parameters nor the same volume of data available.

2.3. World Air Quality Index project

The dataset available from the World Air Quality Index project (WAQD) (World air quality index p, 2021), which corresponds to a set of data referring to the air quality of over 130 countries, covering 2000 major cities, and updated three times a day starting in January of 2020. The dataset contains minimum, maximum, median, and standard deviation values for each of the air pollutant types, which are: CO, NO<sub>2</sub>, O<sub>3</sub>, SO<sub>2</sub>, PM<sub>10</sub>, PM<sub>2.5</sub>, as well as meteorological data, including humidity and temperature.

Furthermore, the WAQI dataset contains data that has been mislabeled, implying that several countries are missing data on various components. Only meteorological data is available in some countries. The AQI estimation bias for a significant percentage of countries in the periods where COVID-19 patterns were present in the entire year 2020 reflects this.

Because of that lack, the filtered data contemplate a reduced amount of 95 countries and 615 cities. Besides, some countries keep data for multiple major cities. For example, in Brazil, data is kept for the cities of
São José dos Campos, Vitória, and São Paulo. We looked for the capital’s country first in these cases, and if we couldn’t find it, we looked for the city with the largest population.

Also, we notice that the cities with daily air pollutants records in the WAQI dataset correspond to the countries’ capitals or major urban centers. Furthermore, that the numbers of confirmed COVID-19 cases in these large urban centers are the primary sources of data to the country’s amount values, with a strong correlation between the numbers in the capital or largest urban center and the country’s amounts numbers. As a result, we took this approach, taking into account registered numbers, curve trending behavior, seasonality, and residual variables.

2.4. Air quality dataset from satellite data

One of the main global monitoring tools is satellites, which are capable to acquire information of almost all Earth’s surface. Focusing on the interest of analyzing air quality, we have used the data generated by the SENTINEL-5 mission, which is part of the European Earth Observation Program (Copernicus) (Copernicus open access hu), under the management and coordination of the European Commission (EC). It is focused on air quality and provide information about the main data products cloud, HCHO, SO\(_2\), O\(_3\), Aerosol layer height, NO\(_2\), CO, CH\(_4\).

The SENTINEL-5 acquires data in form of bands around the Earth based on geographical location and interval of days. Then, the data are processed as numerical matrices and, using cartographic databases, we delimit the analysis of air quality to a region such as a country or smaller ones such as states and cities.

2.5. Air quality dataset from PM filters

As mentioned in Section 2.3, in addition to the WAQI (World air quality index p, 2021) there are also other proven datasets for making data available from PM\(_{2.5}\) sensors. These sensors check the amount of exposure to air pollution in daily average micrograms/cubic meters. It can be used to obtain the number of suspended particulate matter in a unit volume of air within 0.3–10 \(\mu\)m (PM\(_10\)). In Brazil, through the WAQI platform we can see that most of the installed sensors are in the south/southeast and midwest regions. And that the regions that make data available for these platforms also have their own platform to share data with the population, which is the case of CETESB - State of São Paulo [33], JEAP - State of Rio de Janeiro [34], CELEPAR - State of Paraná [35] and IEMA - State of Espirito Santo [36].

With the use of Particulate matter sensors, it is possible to get real-time data from regions where one wants to monitor and compare the data obtained from the sensors with other air pollution visualization methods, such as satellite data. It is also possible to obtain data from closed environments, such as educational, industrial, shopping centers or any other closed environment. For its installation, mainly in Brazil, there is a technical standard from the Ministry of the Environment [37], as if it is installed wrongly, it can provide false data to anyone who may use them. In this work, in order to show our approach we use a subset of these data, conforming the São Paulo City regulations.

2.6. Feature engineering

Based on country data from the WAQI and WHO datasets, we decided to work with a similar number of countries. As a result, we only kept countries that had information on at least five pollutant features (out of a total of six pollutant features that were measured). Fig. 3 illustrates the workflow.

At the end of this features and countries selection, we organize 7 dataframes, each of which contains information on 63 countries, and the data corresponds to:

- COVID-19 daily deaths Dataframe;

Finally, the country features dataframes are fed into the AQI calculation, which uses a six-level scale for each day as specified by the US-EPA 2016 standard, with: Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous, which represents scales for health consequences and cautionary statements. The COVID-19 daily deaths were used with the AQI six-level dataframe in the forecast experiments.

The AQI is noteworthy because of certain cautionary statements for active children and adults, as well as people with respiratory disorders such as asthma. The five-level are classified as:

- Good - None. 0 to 50 as a value range;
- Moderate - should limit prolonged outdoor exertion. Range of values: 51 to 100;
- Unhealthy - should limit prolonged outdoor exertion. Set of values divided into two ranges: 101 to 150 (more related to sensitive people) and 151 to 200 (population in general);
- Very Unhealthy - should avoid all outdoor exertion; everyone else, especially children, should limit outdoor exertion. Set of values divided into two ranges: 201 to 250 (very danger mainly to sensitive people) and 251 to 300 (very danger to population in general);
- Hazardous - everyone should avoid all outdoor exertion. A value of 301 and higher.

2.7. Clustering data

Natural clusters in the AQI of the countries’ output were identified using the unsupervised K-means algorithm. Using this tool, we can find countries with similar pollutant air patterns that could be used according to the place where the model will be applied. In this way, we can use only a specific group, decreasing the number of data and improving the model generalization. The algorithm’s number of \(k\) determines the number of clusters required to group the elements.

K-means is an unsupervised approach that takes a set of \(n\) observations and splits them into \(k\) groups, identifying related trends in each
group depending on the n-nearest observation’s mean [38]. The idea is to organize each point to its current nearest center, update the clustering centers by calculating the mean of the member points, and repeat the process before convergence conditions [38]. The algorithm can be applied in a variety of ways, including arranging counties. However, since the method is unsupervised, it should be combined with an un-regulated neural network for pattern extraction.

2.8. LSTM based Neural Networks

One or more explicit knowledge re-feeding into the structure is common in Recurrent Neural Network (RNN). However, when combined with the backpropagation algorithm, this type of function can cause “blow up” or vanishing gradients. In first case, the gradient explosion can result in an oscillation over the neural weight values. On the other hand, the vanish may turn impracticable the network fitting for assimilating the requisite time dependency. To handle with these problems [39], it is proposed the Long Short-Term Memory (LSTM).

In contrast to traditional RNN, the LSTM preserves information for a larger number of historical data entries while maintaining the importance of recent data, by the use of two internal states, which are the cell state and the hidden state, making the information transport to the network neurons ahead.

An LSTM is an architecture of the RNN family that conveys both long and short-term information. To accomplish such complex information retrieval from sequential sources, the LSTM is built based on the use of gates that controls the flow of information among the cells. To accommodate the output of the LSTM cells to the predictions of the number of deaths range, we employ a linear layer at the output of the model. The graph of this approach is shown on Fig. 4.

The data is captured, processed, and analyzed using the Python 3.7 platform, Pandas 1.2 framework, and Matplotlib (release 3.4.0) for plotting statistical graphics. The LSTM neural network was implemented and tested using Tensorflow version 2.4.

3. Experiments and results

In order to validate our approach, a series of setups and experiments have been done. The first experimental part is related to data regularization and clustering, in order to be used by the LSTM network. Then, this clustered data is used for the training of LSTM, whose parameters and specifications were better adjusted. Next section depicts the experiments with clustering, and the LSTM, with their particularities. Finally, the application on the evaluation of different feature configurations is performed in order to verify the correlation.

3.1. Clustering with K-means

To understand the behavior and reaction related to changes in the air quality variation, our idea is to perform the K-means to clustering over the AQI shape curves, related to the behavior time-based on the trend and seasonality of the temporal series. Therefore, before applying the K-means algorithm, each country’s AQI is subjected to a 30-day moving average, mitigating the effects of a random value and short-term fluctuations on the AQI over a specified time frame. After, each curve was normalized, assisting in the establishment of clearly defined elements and attributes, providing a comprehensive catalog of data, and converting that data into a uniform format ranging from 0 to 1.

Subsequently, we submit the data to Dynamic Time Warping (DTW) and Euclidean metrics to evaluate the algorithm. The Euclidean distance metric fits better once we applied it to the normalized data, showing that the clustering was based not on its AQI absolute values.

On the K-means algorithm, we set the k value for this analysis as 9 and it was submitted for the process of dividing 63 country using the Euclidean distance calculated. Fig. 5 depicts both cluster k-numbers and the countries that made them up.

Cluster 1 (Fig. 6a), cluster 6 (Fig. 6b), and cluster 7 (Fig. 6c) were chosen to understand the grouping made from the behavior of the AQI curves, as shown in Fig. 6. As can be seen in these comparisons, there are several patterns in the prevalence of air pollutants trends and seasonality for each country over specific periods, and COVID-19 daily deaths provide important data on air quality.

As mentioned, the clustering was based on the trend and seasonality of the curves. Even though the K-means produced clusters based on similar shapes and it may permit that some countries grouped on a specific cluster may not stand on the same expected AQI distribution values. As shown in Fig. 7, through descriptive statistics, the generated clusters can contain countries that contrast in data distribution, in which the expected value (mean value) of AQI from different countries may not be the same interval values.

Once K-means clustered the elements based on the distribution and regulation of the nearest centroids for attribution of the found point to this cluster, the approach allows for the organic selection data, as well as the labeling and referring of centroids to newly inserted data. This results in using for explanation in terms of understanding how residents react to the behavior of the local curve, in a place with a specific AQI, which residents are already subjected to this air quality. Any other country that comes into one of these clusters will be able to submit a neural network transfer learning, which will allow for a better grasp of the data analysis and understanding.

Lastly, we can better explain the behavior of the data by the exploratory data analysis methods histograms with Kernel Density Estimation (KDE) using Fig. 7. Notice that we added a kernel centered about each data point for every single data point and we eventually sum these together trying to get the probability density function f̂(x) that describes well the randomness of the data. When the AQI distribution has an upper tail (tail at the end of the distribution, at the top) more pronounced than the lower tail (tail at the bottom of the distribution, at the bottom), the distribution has a negative asymmetry, as shown in Fig. 8a. It will have a positive asymmetry if the reverse is true (Fig. 8b). The distribution is otherwise symmetric (most apparent to the normal distribution curve) (Fig. 8c). As a result, not necessarily we’re dealing with outlier data, but rather data with the asymmetric distribution, also known as skewed data.

3.2. Evaluating the prediction models for the city of São Paulo

Due to the small size of the training dataset, we decided to train the model with 90% of the data and the remaining 10% of the data is used to evaluate the predictive skill of the model. We use data from March 27, 2020 to June 03, 2021, which accounts for 435 days, or 62 standard weeks of data (Sunday to Saturday). The train dataset is composed of the
first 56 weeks and the test dataset is the last 6 weeks. To alleviate trends and seasonality we use as input the biweekly (14 days) moving average for the COVID-19 deaths and cases time series.

Forecast of future deaths is performed on a weekly basis. The first week of the training dataset \((d_0, \ldots, d_6)\) is used as the input of the model to predict the next one \((d_7, \ldots, d_{13})\). The input/output samples are organized using a sliding window of 1, so the next input sample is \((d_1, \ldots, d_7)\) and the expected output is \((d_8, \ldots, d_{14})\). This approach increases the number of training examples from 56 to 392 samples.

The train dataset has size \(n\) and the test dataset has size \(j\). \(W\) is the total dataset \((\text{train + test})\) tensor and \(W_i\) is week \(i\) from the dataset. We used a walk-forward approach to validate the models after training. In this approach, the last week of the training dataset \(W_5\) is used as input of the forecast model to predict the next one \(\tilde{W}_1\). We then use \(W_{n-1}\) to predict \(\tilde{W}_2\), \(W_{n-2}\) to predict \(\tilde{W}_3\) and so on. The output of this process is a tensor of predicted data \(\tilde{W}\) of size \(j\). Therefore, it is possible to compare the forecast of the model against real data in the test dataset and estimate the forecast performance for multiple future weeks. The Root Mean Square Error (RMSE) between real data from the test dataset \((W_{n-1}, W_{n-2}, \ldots, W_{n+j})\) and the respective predicted weeks \((\tilde{W}_1, \tilde{W}_2, \ldots, \tilde{W}_j)\) is used as a performance metric for the models. A small RMSE average for the predicted weeks indicates good generalizability and small RMSE for the whole dataset indicates good fitness. This evaluation approach was used in both the univariate and multivariate cases.

Due to the stochastic nature of Neural Networks weights optimization, it is good practice to train multiple versions of the same model. It could be the case that models with similar performance on training could result in significantly different forecasts. To increase the robustness of this process we first performed a simple grid-search on the hyperparameters of the network to find the most suitable training setup for each input configuration. The training configuration setup is a vector \([c_1, c_2, c_3, d_1, d_2, d_3, d_4, \text{dropout rate}, \text{batch size}, \text{epochs}]\), where \(c_1\) is the number of input days, \(c_2\) is the number of units at each layer, \(c_3\) is the number of epochs, \(d_4\) is the batch size, \(d_5\) is a dropout rate and \(c_6\) indicates if the input series are normalized. We tested 8 configurations, changing the number of units of each layer \((50 \text{ and } 100)\), the batch size \((16 \text{ and } 32)\) and if the inputs are normalized or not \((\text{None and MinMax})\). The training for each set of hyperparameters has only 10 trials of 100 epochs each, and the average performance of configuration is stored and compared with other configurations. This grid-search process was performed for each of the eight input configurations. After selecting the most suitable hyperparameters for each input configuration, we performed a more in dept training, with 500 epochs and 50 trials. The network architecture is as described in Section 2.8 and the evaluation of the models found is presented below.

### 3.2.1. Univariate approach

For the univariate case we used the deaths time series (input) to predict the deaths time series (expected output). This case is relatively simple, the network only has problems fitting the curve if the time series has high autocorrelation. The best training configuration found in the grid-search was \([7, 50, 100, 16, 0.0, \text{None}]\). The average RMSE for the whole dataset and for the test dataset is shown in the boxplots D of Fig. 9. The forecast curve from the best model found during the 50 models trial is shown in the first two graphs of Fig. 10 (biweekly moving average and accumulated deaths).

### 3.2.2. Multivariate approach

For the multivariate approach we decided to perform three experiments, gradually adding more features to the input and observing the predictive skill of the models found. The three input configurations tested are: deaths and Aqi \((D + A)\); deaths, Aqi, temperature and humidity \((D + A + T + H)\) and deaths, cases, Aqi, temperature and humidity \((D + C + A + T + H)\). Fig. 9 show the average RMSE for these cases, both for the whole dataset and for the test dataset and the curves predicted by the models can be observed in Figs. 10 and 11.

### 3.3. Discussion about the results

The experiments provided some interesting results. First, about the evaluation, we decided to plot both the RMSE for the whole dataset (the entire curve in Fig. 10) and also for only the test dataset (only the red part of Fig. 10). The first RMSE represents the fitting performance of the model, namely, the capacity of the model to approximate data from the training. The second RMSE represents the generalization performance of the model, i.e., how the model performed on data not seeing during training. Another important information to evaluate the model performance is the accumulated deaths. We use an approach that is similar to the one that has been used to compare the performance of several forecast models (Friedman et al., 2021). The idea is to see to what percentage of the real accumulated COVID-19 deaths the prediction models gets after the forecast. We improved this approach by also showing the entire curve, to see if the deaths accumulated by the forecast also follow the profile of the real accumulated curve (second plot of Fig. 10).

The motivation for evaluating both the fitting error and the forecast error can be seen in the upper graph of Fig. 10. This graph shows the predicted curve of the best model for the univariate case. Despite having a good prediction error of only 8.42, this model wasn’t able to converge to a good fitting solution. This can be partially explained through the boxplot D of Fig. 9. Notice that this boxplot has low standard deviation.
which is an indication of the poor quality of the input space created by only using one feature. Since the search space is less diverse, the training is more effective in finding sub-optimal solutions with similar performance, and this partially explains why this case had the best overall expected forecast performance (Fig. 9b).

Next we provide an analysis for the multivariate cases. The boxplots presented in Fig. 9a clearly shows that the average fitting performance of the models increases when more input features are added (lower mean RMSE). The case where deaths, air quality, temperature and humidity (D + A + T + H) are used as input has the best overall fitting performance. And, as it can be seen in Fig. 9b, it also generates the model with the best forecast performance, even though in the average it lost to the univariate case. A problem that we noticed is that the input space is so appropriate for the network architecture that it ends up generating models that overfit, i.e., that fit the curve perfectly for the training dataset but that has a poor performance on data that it does not see during training (this means that it does not generalize). The forecast shown in the upper graph of Fig. 11 is the best overall model, which has a good trade-off between fitting and forecast performances. Even though the predictive error is a bit higher than some other models, this model also has a near perfect overall accumulated COVID-19, being only $-0.06\%$ away from the real accumulated value after the forecast. So we can say with confidence that this model is the most skilled that we found, that is, it has superior performance in the 7 day forecast for the COVID-19 deaths time-series of the city of São Paulo.

As a final remark, the network does not improve its performance by adding COVID-19 number of cases time-series in the mix (D + C + A + T + H). It takes much longer to converge and also does not converge to better solutions than the case without the cases time-series. Although intuitively, the use of the cases time-series as input would help the forecast, since more people with COVID-19 means that more people can die of COVID-19. However, this does not translate into a better performance of the neural network function approximation. An explanation is that the number of cases time-series has a high linear correlation with

---

**Fig. 6.** Countries groups created by K-means algorithm - AQI curves for 1-, 6-, and 7-k clusters.
the number of deaths time-series, and, in terms of the neural network training, this does not add much information to the search space, despite making it more complex (with one more dimension). It is also possible that changing the architecture of the network can lead to better models in this case, and this investigation can be performed as future works.

One of the goals of our project is to have a common repository (in...
construction at http://ncovid.natalnet.br), which will host all information, including not only codes however also the results of our achievements in form of graphs and web pages, which will be made available to the community. A preview is at www.natalnet.br/covid that is working currently with previous data to this work (we are on the way to update it). Besides, code and data used in this research can be found in https://github.com/Natalnet/ncovid-air-paper.

4. Conclusion

In this work we have verified the influence of environmental factors as temperature, humidity and air quality, on the forecast of the COVID-19 pandemic behavior. For that we have used a data driven approach, which is based on a simple model using RNN with LSTM cells. The model has been applied on COVID-19 data from Brazil reporting the daily deaths, the daily temperature, humidity, and an index for air quality.

From the experiments we could observe some interesting issues. First, as expected, we noticed that the effects of isolation and lockdown causes instantaneous modifications on data that the temporal average approach can not explain nor reduce its problems. Perhaps the annual average can help in this issue, geographically speaking. Yet, results from the experiments using multiple feature configurations could indicate that a data-driven approach can be used to strengthen the hypotheses of correlation between air pollution and COVID-19 fatalities.

As noticed, the number of groups used in the K-means is 9, for the reported results. We also tested with other numbers, however the results

---

Fig. 10. Forecast for the biweekly moving average of COVID-19 deaths and accumulated COVID-19 deaths for the city of São Paulo. These are the best models for each input configuration.
were not as satisfactory as for this number, which has the best performance. In future work, we intend to investigate and apply the results of K-means; after adding a new location and inserting it into a cluster, we can identify a previously trained model and use it in forecasting tasks.

This new set of features has been shown to have severe influence on the COVID-19 dynamics, so we figured out from here that it should be taken into consideration by any method that tries to understand the behavior of the pandemics, mainly the ones based on AI that have been developed in our previous research (Pereira et al., 2020). Our future work is exactly in this direction, with the feature set improved by the current work. We will revisit the use of MAE, LSTM, and other deep learning approaches on COVID-19 pandemic for predicting its behavior on a long time manner.

**Funding**

This work has been developed with partial support of Coordination for the Improvement of Higher Education Personnel (CAPES-Brazil), grants number 001 and 88815.056890/2020-01, and by National Research Council (CNPq-Brazil) grant number 311640/2018-4.

**Availability of data and material**

Code and data used in this research can be found in https://github.com/Natalnet/ncovid-air-paper.
Credit authors statement

All authors have made substantial contributions to the conception or design of the work; the acquisition, analysis, or interpretation of data; the creation of new software used in the work; drafted the work or revised it critically for important intellectual content; and approved the version to be published, as detailed next. Dunfry P. Aragao: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review and editing; Emerson V. Oliveira: Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft; Arthur A. Bezerra: Data curation, Investigation, Visualization, Writing - original draft; Davi. H. Santos: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, Writing - review and editing; Andreogus G. Silva-Junior: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft; Igor G. Pereira: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft; Jordan S. Cuno: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - review and editing; Alessandro Miani: Conceptualization, Data curation, Funding acquisition, Methodology, Resources, Supervision, Writing - review and editing; Luiz M. G. Goncalves: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review and editing. In addition, all authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Declaration of competing interest

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

References

Al-Qaeness, M.A., Saba, A.I., Elsheikh, A.H., Elaziz, M.A., Ibrahim, R.A., Lu, S., Hemdan, A.A., Shamsungan, S., Ewees, A.A., 2021. Efficient artificial intelligence forecasting models for COVID-19 outbreak in Russia and Brazil. Process Saf. Environ. Protect. 149, 399–409. https://doi.org/10.1016/j.prosafenv.2021.10.007.
A. C. Adler, F. A. Casarso, V. G. da Silva, L. F. Fires, Evidence that high temperatures and intermediate relative humidity might favor the spread of COVID-19 in tropical climate: a case study for the most affected Brazilian cities. Sci. Total Environ. 729, doi:10.1016/j.scitotenv.2020.139090.
Barcelo, D., 2020. An environmental and health perspective for covid-19 outbreak: meteorology and air quality influence, sewage epidemiology indicator, hospitals disinfection, drug therapies and recommendations. Journal of Environmental Chemical Engineering 8 (4), 104066.
Belosi, F., Conte, M., Gianelle, V., Santachiara, G., Contini, D., 2021. On the concentration of sars-cov-2 in outdoor air and the interaction with pre-existing atmospheric particles. Environ. Res. 193, 110603.
Brassey, J., Heenegan, C., Mahltani, K.R., Aronson, J.K., 2020. Do weather conditions influence the transmission of the coronavirus (SARS-CoV-2)? Tech. rep., The Centre for Evidence-Based Medicine. www.cebm.net/oxford-covid-19/.
A. Carducci, I. Federigi, M. Verani, Covid-19 airborne transmission and its prevention: waiting for evidence or applying the precautionary principle?, Atmosphere 11 (7). doi:10.3390/atmos110711701. URL https://www.mdpi.com/2073-433X/11/7/11710.
Chan, K.H., Peiris, J.S., Lam, S.Y., Poon, L.L., Yuen, K.Y., Seto, W.H., 2011. The effects of temperature and relative humidity on the viability of the SARS coronavirus. Advances in Virology. https://doi.org/10.1155/2011/734690.
Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., Du, M., Liu, M., 2020. Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. Sci. Total Environ. 729, 139051. https://doi.org/10.1016/j.scitotenv.2020.139051.

Zangari, S., Hill, D.T., Charette, A.T., Mirowsky, J.E., 2020. Air quality changes in New York city during the covid-19 pandemic. Sci. Total Environ. 742, 140496.