Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 196 (2022) 655–662

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

Unsupervised analysis of COVID-19 pandemic evolution in brazilian states

Victor Cassão[a*], Domingos Alves[c], Ana Clara de Andrade Mioto[b], Filipe Andrade Bernardi[b], Newton Shydeo Brandão Miyoshi[a]

[a]*Barão de Mauá University Center, Ribeirão Preto/SP, Brazil*
[b]*Bioengineering Postgraduate Program, University of São Paulo, São Carlos, Brazil*
[c]*Ribeirao Preto Medical School, University of Sao Paulo, Ribeirao Preto, Brazil*

Abstract

Extracting information and discovering patterns from a massive dataset is a hard task. In an epidemic scenario, this data has to be integrated providing organization, agility, transparency and, above all, it has to be free of any type of censorship or bias. The aim of this paper is to analyze how coronavirus contamination has evolved in Brazil applying unsupervised analysis algorithms to extract information and find characteristics between them. To achieve this goal we describe an implementation that uses data about Covid-19 spread in Brazilian states (26 states and the federal district), applying a Time Series Clustering technique based on a K-Means variation, using Dynamic Time Warping as a similarity metric. We used data reported by the Ministry of Health in Brazil, referring to deaths per 100k inhabitants, during 452 days from the first reported death in each state. Two analyzes were performed, one considering 3 clusters and the other with 6 clusters. Through these analysis, 3 patterns of responses to the pandemic can be observed, ranging from one of greater to lesser control of the pandemic, although in recent months all clusters showed a highly increase in the number of deaths. The identification of these patterns is important to highlight possible actions and events, as well as other characteristics that determine the correct or incorrect public decision-making in combating the Covid-19 pandemic.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the CENTERIS –International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

Keywords: Time Series Clustering; Dynamic Time Warping; Unsupervised Analysis; Covid-19;

*Victor Cassão. Tel.: +551699287-3393.
E-mail address: victorcassao@gmail.com

1877-0509 © 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the CENTERIS –International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021
10.1016/j.procs.2021.12.061

1. Introduction

1.1. Contextualization

In the end of 2019, in the city of Wuhan, in China, was reported the first clinical case of a new disease. The preliminary analysis indicates that the disease was caused by a virus with a strong relationship with the virus of Severe Acute Respiratory Syndrome(SARS-2002) and the Middle East Respiratory Syndrome(MERS-COV-2012). Some studies have shown that this novel coronavirus demonstrated 96% of similarity with bat's coronavirus from the species Rhinolophus affinis [1]. Its origin is still unknown, but some studies have shown a possibility of a mutation between bat's coronavirus and an unknown origin coronavirus, transmitted to humans by an intermediate host [2].

The first positive case of the novel coronavirus (SARS-COV-2) in Brazil was reported by a traveler returning from Italy on February 25th, 2020, with the first death confirmed on March 17th, 2020. After the first case, Brazil has had an increase in total confirmed cases, especially in the southeast region, that has represented more than one-third of total cases of the country at that time [3]. That region includes the city of São Paulo and Rio de Janeiro, two of the most important Brazilian's urban centers, that in the first months became the main hotspot of the disease, starting to create a local transmission between capital, metropolitan cities and advancing to small cities in the countryside.

It is possible to analyze the virus dissemination based on all available data provided by health departments. In an epidemic scenario, this data has to be integrated providing organization, agility, transparency and, above all, it has to be free of any type of censorship or bias. It should reflect the current status of the virus dissemination to provide to health managers reliable parameters to make good decisions.

Extracting information and discovering patterns from a massive dataset is a hard task. But, there are some techniques that can help the extraction of those characteristics. The unsupervised analysis provides many techniques that perform well when used in this scenario. The clustering technique is a well-known algorithm extremely efficient to find patterns in the dataset, making it possible to visualize and understand their characteristics. In general, it works in an unsupervised way, dividing similar data into the same groups, without any previous knowledge about the data[4].

1.2. Related Work

A proposed methodology was described in [5] to analyze the Covid-19 evolution in all states of the United States. The authors presented a multiple Time Series analysis, applying a hierarchical clustering technique using the Dynamic Time Warping(DTW) as a distance metric to discover patterns and find similarities about the disease's spread in different regions of the country. It's also done a prediction of the pandemic evolution based on mathematical models (Logistic, Gompertz and SIR model).

In [6] a multivariate analysis of Covid-19 evolution in Brazil is proposed using a dataset containing 10 different features classified as Health, Geographic and Social indicators. A factor analysis is applied in this dataset to reduce the number of original variables, extracting the most important information, and then, applying the classical clustering algorithm (K-Means) to group similar data and analyze their characteristics to understand the spread.

A hierarchical clustering among countries is presented in [7] showing the results when this technique is applied using three different features. In this case, when applied using the number of cases, number of cases per million of inhabitants and the number of cases per million and per country's area. The results have shown that each clustering process grouped countries in different ways, according to each country's particularity.

1.3. Objectives

The aim of this paper is to analyze how coronavirus contamination has evolved in Brazil, among the 26 states and the federal district, applying unsupervised analysis algorithms to extract information and find characteristics between them. This paper brings preliminary results from an initial stage project.

The paper is organized as follows. In Section 2 there is a brief description of used datasets, theoretical references and used technologies, like libraries and programming languages. In Section 3 there are all the results. In Section 4, a conclusion about the initial results and a discussion about future works.

2. Materials and Methods

This section describes the dataset used as the basis for the analysis, the technological aspects such as programming language and software libraries, the machine learning techniques used in the project, and the research plan developed.

2.1. Dataset

All used data in this work was taken from a Github's repository [8] that compiles public Covid-19 health indicators about the epidemic, retrieved from official sources like the Brazilian Ministry of Health at the federal level and State Health Secretaries at the stadual level, validating and checking the data integrity to share with the community.

There are some available files in CSV (Comma Separated Values) format, representing different types of dataset. As the objective here is to analyze the pandemic evolution between states, the chosen dataset was the one containing only information about Brazilian states, disregarding cities, initially. This repository receives daily updates since the first reported case in february, 25th, 2020. The present data, after filtering for the purpose's analysis, has the following indicators: date, state, deaths, new deaths, total cases, new cases, deaths per 100k inhabitants and total cases per 100k inhabitants. In this paper was used the time series of the number of deaths per 100k inhabitants daily reported for each brazilian state.

2.2. Unsupervised Analysis

The unsupervised analysis is a machine learning approach that aims to classify data without any kind of labeling dataset. There are a lot of different algorithms for unsupervised analysis, and each one has its specific application, better and worst use cases, but in general, they are applied for pattern recognition.

Clustering algorithms are a kind of unsupervised analysis where their main function is to aggregate similar data in groups, based on a similarity metric. The same identified group will have a high similarity ratio, however, different groups have a low similarity ratio. A famous clustering algorithm is K-Means, which consists in partitioning N elements in K groups, or clusters, where N is the total of elements being analyzed and K, the total of clusters. The value of "K" can be defined by the user experience or using some available techniques to calculate a possible value, for example, the elbow or the silhouette method.

K-Means defines an initial value to the cluster's central point, also known as medoid, and generally, the choice is to pick up a random value between the range of the data. The algorithm will associate each datum to the nearest medoid based on a distance metric, usually Euclidean distance, and then, recalculate the medoid considering the new data associated, until it converges to a value or the iterations ends.

As described in section 2.1, the dataset used here is a type of time series dataset. Differently from static data, time series has a feature that changes over time [4]. In other words, a given phenomenon varies through a specific period of time. It is necessary to be careful to manipulate this kind of data, mostly, because the data is organized in a specific order that represents how that phenomenon has occurred, and any change in that order, even minimal, will implicate a different behavior. Because of that, the clustering process of a time series dataset is a complex task. This section brings a description of a possible implementation with K-Means, using a different similarity metric.

As described above, a time series is a dataset ordered in a specific sequence that represents the behavior of a given phenomenon over a period of time. The problem [9] of a time series clustering using the Euclidean distance as a metric is exactly the incapacity to consider and analyze those time shifts. Particularly, this metric is unable to understand the time changes, ignoring the most important data information from the time axis. This is a serious problem because it twists the similarity calculus between two series.

DTW is, like the Euclidean distance, a similarity measure, but more recommended to time series because of its capacity of time-shift recognition. There are many data mining applications, like classifications, clustering, anomaly detection and others [10].

The DTW technique works by taking two time series and, using a distance matrix, calculating all the possible non-linear paths between them [11]. From all these calculated paths, the minimum path will be selected, representing the relative distance of those two series. The method is able to align series data, even with different lengths.

The TSLearn library, further described in section 2.3, already has this implemented algorithm. It works by using a K-Means variation, that instead of the Euclidean distance, uses the DTW as similarity metric to agrupate similar series.

2.3. Technologies

All developed code was created in the programming language Python. For graphic visualizations, the Matplotlib library was used. For time series clustering, it was used the clustering module, from the TSlearn library. Scikit Learn was used for basic machine learning algorithms such as K-Means. Pandas was used for data cleaning and transformations.

2.4. Research and Analysis Workflow

The research plan and analysis workflow was defined based on a Knowledge Discovery in Databases (KDD) approach. It is an extraction process of useful information from raw data [12]. Fig. 1 describes how the dataset exploratory process was divided in two main steps: Data Preprocessing and Time Series Clustering. The data preprocessing step was conducted by the dataset creation, taking the raw data and calculating the correct period of time. The clustering step was responsible for the time series clustering algorithm application, using the DTW as similarity metric, resulting in the clustered data. Both steps, further explained in sections 2.5 and 2.6, respectively.

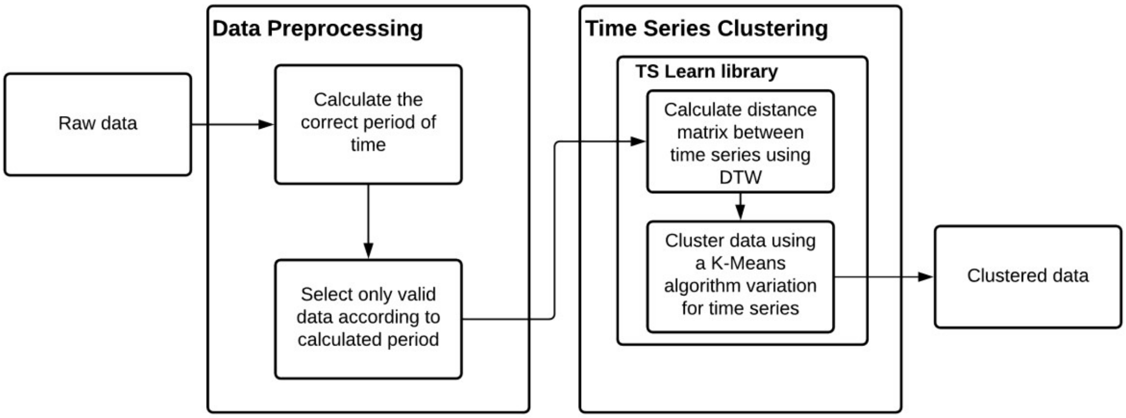
Raw data
Data Preprocessing
Calculate the correct period of time
Select only valid data according to calculated period
Time Series Clustering
TS Learn library
Calculate distance matrix between time series using DTW
Cluster data using a K-Means algorithm variation for time series
Clustered data

Fig. 1. Research workflow.

2.5. Data preprocessing

The time series dataset was made using the deaths per 100k inhabitants over a specific period of time, having the starting point at the first confirmed case in each state. As the first cases did not occur on the same day, it was needed to calculate the delay between the first and last confirmed case to analyze states evolution based on the same period of time. This adjustment aligns all state's situations, once we are observing now, to the evolution in days, not by dates. At the interval of the first and last positive case in states, São Paulo and Roraima respectively, 25 days had passed by. That difference was subtracted off the current day, to result in a closed period of 452 days.

2.6. Time Series Clustering

The clustering step was divided into two pieces given by two different values of k. The first value was chosen as a way to map the epidemic spread based on the dataset clustered into 3 groups of control (k=3), in a try to group the response against the pandemic evolution in Brazil in 3 relative levels: bad, medium and good. The second value was defined by a heuristic, calculated by the upper value of the square root of the total series in the dataset, returning k=6.

2.7. Brazilian States Knowledge

Here follows a brief description about abbreviation of the 26 Brazilian states and the federal district, largely used in graphics plotted in this paper: Acre (AC), Alagoas (AL), Amazonas (AM), Amapá (AP), Bahia (BA), Ceará (CE), Distrito Federal (DF), Espírito Santo (ES), Goiás (GO), Maranhão (MA), Minas Gerais (MG), Mato Grosso do Sul (MS), Mato Grosso (MT), Pará (PA), Paraíba (PB), Pernambuco (PE), Piauí (PI), Paraná (PR), Rio de Janeiro (RJ), Rio Grande do Norte (RN), Rondônia (RO), Roraima (RR), Rio Grande do Sul (RS), Santa Catarina (SC), Sergipe (SE), São Paulo (SP) and Tocantins (TO).

3. Results

In Fig. 2, there are the results of the time series clustering with k=3. This k value, apparently, shows a well defined cluster 2 with the states of Alagoas(AL), Bahia(BA), Maranhão(MA). Although, clusters 0 and 1 strongly indicate that they are overloaded. For example, the first cluster has three states at the bottom outlier, Minas Gerais(MG), Paraná(PR), Santa Catarina(SC), it seems they could be in a different, maybe isolated, cluster, because of their different shape from the others.

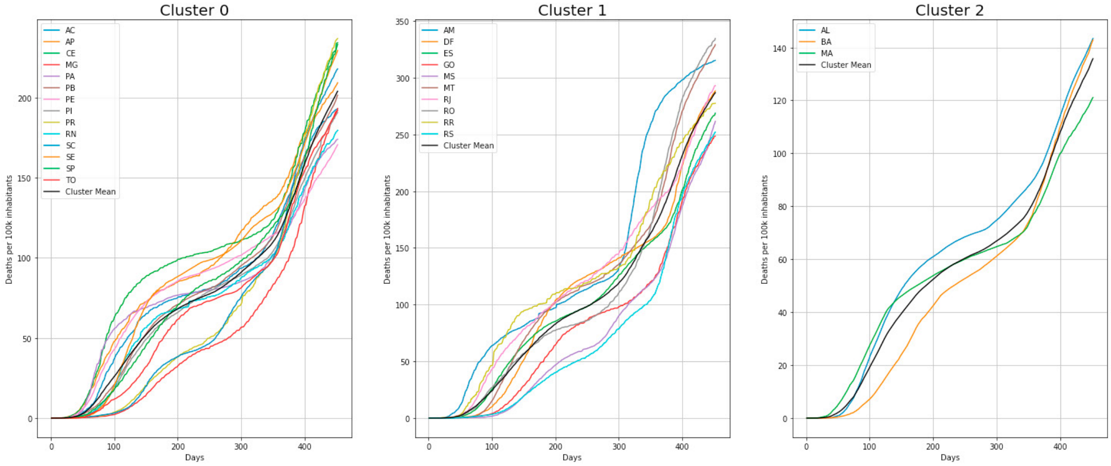

Fig. 2. time series clustering with k=3.

Fig. 3, k=6 brings a slightly better grouping than k=3. States that before were overloaded in a single cluster, are now divided more precisely. The states of Pará(PA), Pernambuco(PE) and Rio Grande do Norte(RN) were grouped at the same cluster with other eleven states, and here, they are grouped in an isolated one.

The states of Alagoas(AL), Bahia(BA) and Maranhão(MA), even changing the number of clusters, kept grouped together. It is possible to consider there are interesting characteristics between them, able to explain their spreading evolution.

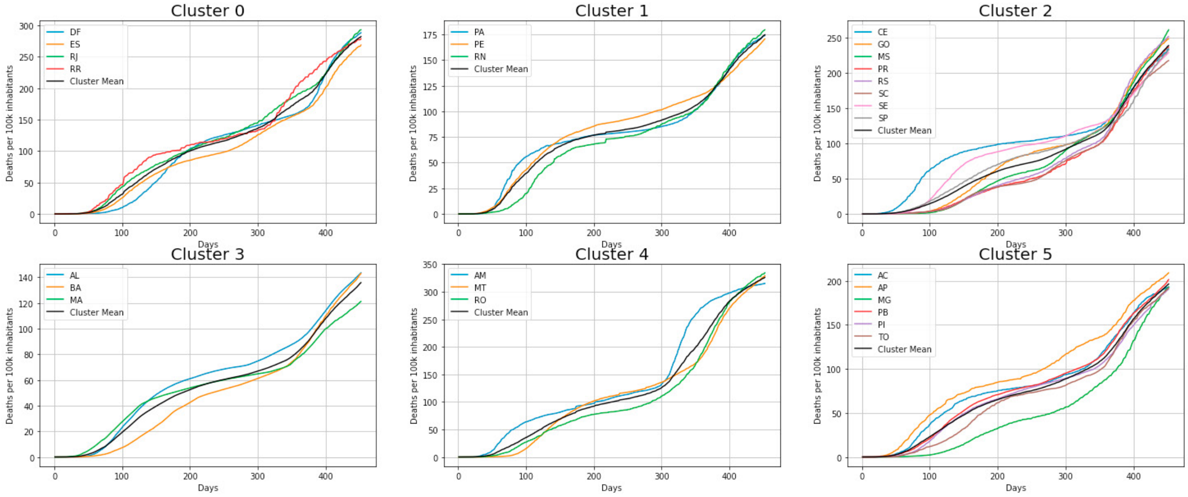

Fig. 3. time series clustering with k=6.

4. Discussion

In this work, we seek to identify how Brazilian states responded to the Covid-19 epidemic based on clustering in relation to the number of deaths per 100k inhabitants. We chose this information as we consider it the most reliable in relation to other available measures that may suffer from a greater degree of underreporting. The usage of a metric that considers the population, makes possible comparisons between different states. So, the use of the number of deaths per 100k inhabitants was an assertive choice. If the absolute death reports had been used, the state of São Paulo, for example, could be in an isolated cluster due to having the highest absolute death numbers.

The first analysis, carried out considering k=3, we can observe that the main differences between the groups are regarding the total number of deaths per 100k inhabitants and also regarding the slope of this rise. We verified that cluster 2, in this case, is a group composed of only 3 states and were the states least affected by the pandemic so far, with an average of at most 140 deaths per 100k inhabitants, but without any sign of pandemic control; cluster 0 is the intermediate in this aspect, where most of the states are and also where there is greater variability of curves and finally cluster 1, where the states most affected by the pandemic are, reaching an average of 280 deaths per 100k inhabitants. In this analysis the number of states per cluster varies a lot. Seeking a better division and grouping of states, we performed an analysis with k=6.

In the analysis considering k=6, following the heuristic to determine the ideal number of clusters, we can observe 4 clusters with 3 or 4 states (clusters 0, 1, 3 and 4), cluster 5 with 6 states and the largest cluster being the 2 with 8 states. We can see in Fig. 4 how the average behavior of these clusters was regarding the control of the pandemic. We noticed that they all maintained a very similar behavior in the first month of the pandemic, but soon after it started to differentiate. Clusters 4 and 0 were the states that had the highest number of deaths per 100k inhabitants throughout the entire pandemic, with 4 having the highest slope and therefore the fastest increase in the number of infected. Cluster 3 were the states that managed to better control the pandemic, maintaining a lower number of deaths per 100k inhabitants and also with a lower inclination.

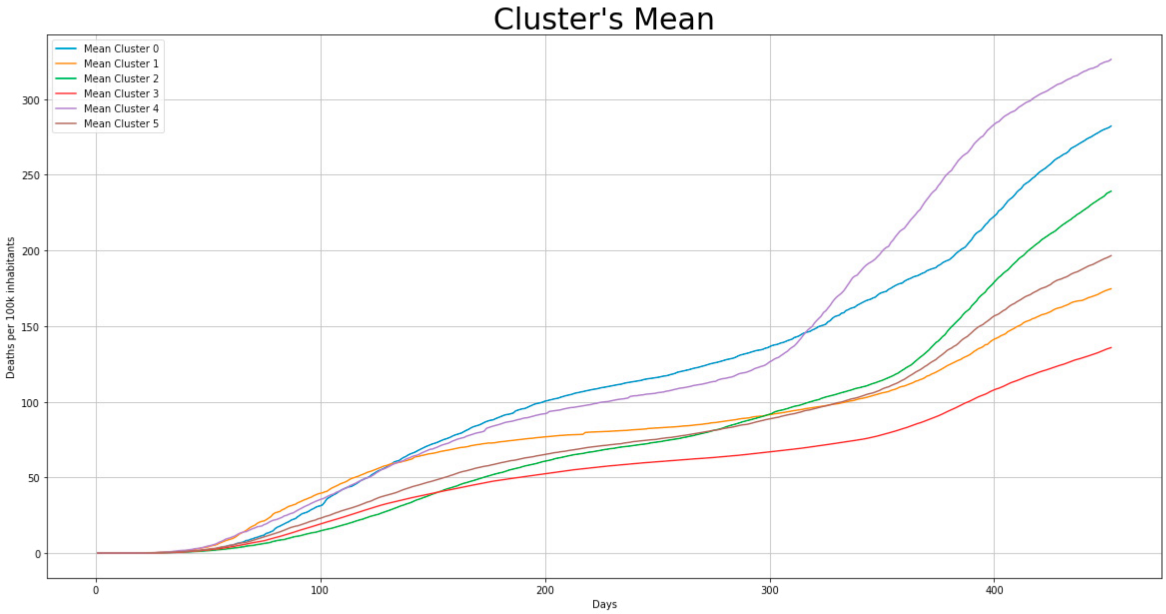

Fig. 4. comparison of clusters mean.

5. Conclusion

The time series clustering has shown an important tool to analyze data and find behavioral similarities between the pandemic evolution in brazilian states. To measure the similarity between series, the Dynamic Time Warping(DTW) was used as a similarity metric, due to its time shift recognition capacity. Initially, the number of deaths per 100k inhabitants was defined as the parameter to be analyzed. The number of clusters was defined by two initial values, 3 and 6, the first, chosen in a try to fit the pandemic evolution in 3 groups of control, and the second, given by an heuristic.

Applying the clustering algorithm to k=3, states were grouped in 2 overloaded and 1 well defined cluster. Considering k=6, states were best distributed, with more well defined clusters, having 3 states being grouped together for both k values, indicating a high similarity behavioral in pandemic evolution in those states (The states are AL, BA, MA and their cluster mean had the lowest numbers of deaths per 100k inhabitants). Even with these lower numbers, the pandemic has never been completely controlled, once Brazil has never had a public politics of social distancing and masks wearing supported by the federal government, contributing to this scenario, not only in this specific case but all over the country.

Despite this evidence and considering a country with continental dimensions like Brazil, further analysis is still needed to identify the reasons for these differences in the pandemic evolution in brazilian states. As a future work, a good improvement will be adding other parameters in the clustering process to bring more information about states, just like done in [5]. Some states have specific particularities that even using a parameter that considers the total population living there, doesn't mean that all demographic factors were really considered. States like MG and AM have a big green area and their analysis need to be more carefully conducted. Another improvement that can be done in future work, is to automate the process of finding the ideal number of clusters through a hierarchical clustering approach [7]. It will be very interesting to compare the hierarchy between states and which kind of characteristics is possible to extract from them.

The results obtained here can give a way to find which actions and strategies against Covid-19 were adopted, or not, by brazilian states to reach this scenario. This analysis can give a background to guide in the future, new actions to help how to proceed and what are the best responses for similar cases.

6. Acknowledgments

We would like to thank Barão de Mauá University Center for the financial support.

7. References

[1] Nogueira, José Vagner Delmiro. (2020) "CONHECENDO A ORIGEM DO SARS-COV-2 (COVID 19)." Revista Saúde e Meio Ambiente 11(2) : 115-124.

[2] Duarte, Phelipe Magalhães. (2020) "COVID-19: Origem do novo coronavírus." Brazilian Journal of Health Review 3 (2) : 3585-3590.

[3] Candido, Darlan S., et al. (2020) "Evolution and epidemic spread of SARS-CoV-2 in Brazil." Science 369 (6508): 1255-1260.

[4] Liao, T. Warren. (2005) "Clustering of time series data—a survey." Pattern recognition 38(11): 1857-1874.

[5] Rojas, Ignacio, Fernando Rojas, and Olga Valenzuela. (2020) "Estimation of COVID-19 dynamics in the different states of the United States using Time-Series Clustering." medRxiv.

[6] Cota, Wesley. (2020) "Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level.".

[7] Ratanamahatana, Chotirat Ann, and Eamonn Keogh. (2004) "Everything you know about dynamic time warping is wrong." in Third workshop on mining temporal and sequential data (vol. 32). Citeseer, Seattle, USA.

[8] Niennattrakul, Vit, and Chotirat Ann Ratanamahatana. (2007) "On clustering multimedia time series data using k-means and dynamic time warping." in International Conference on Multimedia and Ubiquitous Engineering (MUE'07). IEEE Seoul, Korea (South).

[9] Zarikas, Vasilios, et al. (2020) "Clustering analysis of countries using the COVID-19 cases dataset." Data in brief 31: 105787.

[10] Petitjean, François, Alain Ketterlin, and Pierre Gançarski. (2011) "A global averaging method for dynamic time warping, with applications to clustering." Pattern recognition 44(3): 678-693.