Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 219 (2023) 1453–1461

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 2022

Unsupervised analysis of COVID-19 pandemic evolution in brazilian states: Vaccination Scenario

Victor Cassão[a*], Domingos Alves[b], Ana Clara de Andrade Mioto[a], Mariana Tavares Mozini[b], Renan Barbieri Segamarchi[b], Newton Shydeo Brandão Miyoshi[b]

[a] *São Carlos School of Engineering, University of São Paulo*
[b] *Ribeirao Preto Medical School, University of São Paulo*

Abstract

Brazil is one of the countries with the worst response against the pandemic scenario of coronavírus. At the beginning we were on average with 4000 deaths in a 24 hours period. In the course of this situation, large amounts of health and medicine datasets were being generated in real time, requiring effective ways to extract information and discover patterns that can help in the fight against this disease. And even more important is to monitor the progress of prophylactic measures and whether they are being effective in reducing the spread of the virus. Thus, the aim of this study is to analyze how the coronavirus has different ways to evolve in each Brazilian state with the influences of the vaccination process. To achieve this goal, the time series Clustering Technique based on a K-Means variation was applied, with the similarity metric Dynamic Time Warping (DTW). We produced this study using the data reported by the Ministry of Health in Brazil, referring to deaths per 100k inhabitants and all vaccination data available. Our results indicate an unevenly occurring vaccination and the need to identify other associated patterns with human development indices and other socio-economic indicators, being this the first analysis developed in the country, under the goals above.

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

* Corresponding author. Tel.: +55 16992873393.
E-mail address: victorcassao@usp.br

1877-0509
10.1016/j.procs.2023.01.435

1. Introduction

Brazil is one of the countries with the worst response against the novel coronavirus in the world. In the beginning of April 2021, it registered for the first time, more than 4.000 deaths in the last 24 hours [1], and nowadays, in February 2022, it has an amount of 27.492.904 confirmed cases and 638.673 deaths, representing more than 6.65% and 10.96% in cases and deaths, respectively, worldwide [2].

Even starting the vaccination in January 2021, with health professionals, indigenuous population and general population over 60 years older [3], Brazil had many problems related to the vaccination process advancement, guided by, overall, anti-science and anti-vaccine movements.

Brazil has efficient public politics created all over its recent history. One of the most famous and well succeeded is the National Immunization Program (in portuguese Programa Nacional de Imunização - PNI), from 1973, it has the finality of being a national vaccination public policy, able to control, eliminate and eradicate vaccine preventable diseases[4]. Coordinated by Ministry of Health and executed by Public Health System (SUS - in portuguese Sistema Único de Saúde), the PNI has in its structure a shared model with Stadual and Municipal Health Secretaries, being national and international recognizement given to its effectiveness in many diseases eradication, e.g. polio, helping to child mortality ratio reduction and improvement in brazilians life expectancy[5]. Due to the country's extension, the PNI has been an important tool for social and regional inequalities reduction, enabling vaccination to all over the population, independent of its location, coveraging all life stages and granting to everyone the basic health rights access. Given the efforts of health professionals, PNI and SUS supports, and the help of many areas of the society, the vaccination process has made huge advances. In February 2022, it already had an amount of 378.011.196 of applied doses in all of the 5.570 cities in the country[1].

Analyzing the pandemic's evolution is an extremely complex task, because it evolves several factors that directly influence its behavior. This article aims to extend the analysis of [6], that implements an unsupervised analysis approach, through a time series clustering technique based on a similarity metric between them. In the paper, the Dynamic Time Warping(DTW) was used as the similarity metric, and the deaths per 100k inhabitants as the pandemic feature. This study has checked that the pandemic has evolved in different ways between each Brazilian state, having as characteristic, a clustering based on the time series shape. Thereby, this paper aims to apply the same technique, including all vaccination data available, to understand and identify how the pandemic has evolved after the vaccination process starts.

2. Methods

2.1. Dataset

Every data used was taken from a Github's repository project[7] that compiles, validates and checks information about Covid-19 health indicators and publishes for community usage. The data is retrieved from official sources(Brazilian Ministry of Health at the federal level and State Health Secretaries at the stadual level) where it receives daily updates.

The dataset makes available information about: date, state, deaths and cases (cumulative and daily numbers), deaths and cases per 100k inhabitants and for current Brazil's situation, it also has available information about vaccination on first, second, single and third doses.

For the article's purposes, it used the deaths per 100k inhabitants and vaccination coverage of first and second doses(Calculated based on the current population amount of each state, also available in the repository).

2.2. Data preprocessing

To make results comparable with the period before the vaccination started, an adjustment was made in the used features in the clustering process. As the starting point, only data after the first applied vaccine in the country was considered, and as consequence, the number of deaths per 100k inhabitants was reseted to zero, adding up their values according to daily reports after that time.

During December 2021, the database storing all information about the Covid-19 pandemic in Brazil, passed by a blackout in data, starting to share mismatching information with the current situation at that time, lasting up to more than a month to return to normality. To make a consistent dataset, without wrong or missing information, it was defined until the end of November 2021 as a valid and trustable period for collecting information, avoiding that problem on the database. In short, every used data follows a period of time starting in the first applied vaccine, until the end of November 2021.

A final adjustment was applied to the number of 100k inhabitants, normalizing its value to bring it to the same scale when compared to vaccination coverage.

2.3. Time Series Clustering

Time series are a type of dynamic data organization structure in which the features vary according to a certain time interval, minutes, hours, days or months, which is the main point that differs when compared to the static data types (which has no variation on the time axis)[9]. Due to its variation as a function of time, each data contained in a time series is collected chronologically[10], having a structure that is very sensitive to modifications in the order of data, changing the behavior of the observed phenomenon, resulting in a wrong analysis.

Clustering algorithms belong to the group of unsupervised machine learning algorithms, therefore, they operate on unlabeled databases. In general, they are efficient algorithms in finding hidden patterns in the database, minimizing a similarity function for data contained in the same group and maximizing for data in other groups [10,11]. K-Means is one of the most popular clustering algorithms in the literature, being able to process large databases in a relatively short time. Its algorithm works iteratively, grouping n data into k groups, being defined at the beginning the initial value of the clusters, and in the following iterations minimizing a distance function between each datapoint and its closest cluster. The standard implementation of K-Means for static databases is quite efficient, however, in the case of time series, this implementation needs changes for its operation according to the dynamic data of the time series, especially regarding the distance metric used to calculate how close, or far, they are between each other.

As described by [12] the Euclidean distance did not prove to be an adequate metric for time-series data because it does not consider the order of the elements in the data, losing important information related to the analyzed phenomenon, especially in relation to the data obtained through the time axis. Because of this, a new metric called Dynamic Time Warping (DTW) was used to replace the Euclidean distance. This distance works by taking two time series and, using an auxiliary distance matrix, calculating all possible non-linear paths between them. From all these calculated paths, the shortest path will be selected, which will represent the relative distance between the two time-series, also being able during this process to align data points all over the series, even with different lengths[13].

In the work carried out in [14], a multivariate analysis of the evolution of Covid-19 in Brazil was carried out using a dataset containing 10 different features of health, geographic and other social indicators. A factor analysis was applied to the dataset used to reduce the number of original variables in order to extract only the most important information, applying at the end, a K-Means clustering algorithm.

A similar and interesting approach to the one used in this project was also adopted by [15] where the authors carried out an analysis of the evolution of Covid-19 in the US states in order to identify relevant patterns related to the spread of the disease in different regions of the country. In this work, a hierarchical clustering technique of time-series was used, having as a similarity metric to calculate their similarity: the Dynamic Time Warping. Predictions of the evolution of the pandemic were also made based on mathematical models (Logistic, Gompertz and SIR model).

A hierarchical clustering among countries is presented in [16] 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.

The work in this article was carried out applying a time-series clustering using K-Means variations and the Dynamic Time Warping as the distance metric. The algorithm was developed using the TSLearn library that brings a time-series clustering algorithm to run in this analysis.

2.4. Tools and technologies

The developed code was created using the programming language Python and auxiliary libraries, Matoplotlib for graphic visualizations, TSLearn for time series preprocessing and clustering, Scikit Learn for basic machine learning and preprocessing algorithms and Pandas for data cleaning and manipulations.

2.5. Knowledge Discovery in Databases(KDD)

The Knowledge Discovery in Databases is a well-known tool for useful information extraction that, initially, is implicit or unknown in the database's content[17]. This methodology makes possible the understanding of how data are organized and how their relationship works, allowing through its steps a well knowledge for the right handling of the information.

For this article's purpose, the exploratory process was useful for appropriate data extraction, data cleaning, correct period of time selections, as well, data transformations in right patterns according to data format, making possible correct assessment and interpretations.

2.6. Brazilian states knowledge

To better clarify the mentions on brazilian states, right below, follows a brief description about the name's abbreviation of the 26 Brazilian states and the federal district, largely used in graphics plotted in this article: 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

The results for a data grouping with 6 clusters and the deaths per 100k inhabitants, vaccination coverage for first and second doses, and a time range staring on the beginning of vaccination process until the end of November 2021, has resulted in 3 clusters with 3 states (clusters 0, 1 and 3) 1 cluster with 4 states (cluster 5) 1 cluster with 6 states (cluster 2) and cluster 4, having 8 states. The cluster's final results are summarized at Table 1.

Table 1. Clustering results.

State	Cluster	Deaths per 100k inhabitants	Vaccination Coverage First Dose	Vaccination Coverage Second Dose

DF	0	213.55	74%	62%
ES	0	186.78	75%	61%
SC	0	192.22	78%	64%
AC	1	111.48	62%	45%
AM	1	178.27	63%	48%
RO	1	253.90	66%	52%
CE	2	156.22	74%	62%
MT	2	258.13	71%	52%
PB	2	138.60	76%	57%
PI	2	128.69	75%	58%
RN	2	120.59	73%	59%
SE	2	144.58	73%	59%
AP	3	113.72	57%	35%
PA	3	108.07	60%	39%
RR	3	189.21	56%	37%
GO	4	241.99	72%	54%
MG	4	199.58	77%	63%
MS	4	246.58	71%	61%
PE	4	105.67	74%	57%
PR	4	273.39	77%	63%
RJ	4	236.32	74%	59%
RS	4	228.45	77%	66%
SP	4	223.35	81%	73%
AL	5	110.63	69%	50%
BA	5	117.73	71%	52%
MA	5	79.52	62%	45%
TO	5	161.94	65%	48%

For a better visualization of the obtained data, mainly in this case, where we have a strong relationship in the times series evolution and not only for the final absolute numbers, below are the graphs of the final result of the 27 Brazilian states distributed in 6 clusters. Each graph represents the evolution during the established time period of each feature used in the clustering process.

Figure 1 shows the division of clusters according to the evolution of deaths per 100k inhabitants. The clusters, in general, have shown the characteristics of the technique about shape-based grouping. The states of MT, RR, PE grouped in clusters 2, 3 and 4, respectively, showed a greater intensity in their variation when compared to the others, but still maintaining the characteristic of cluster's shape.

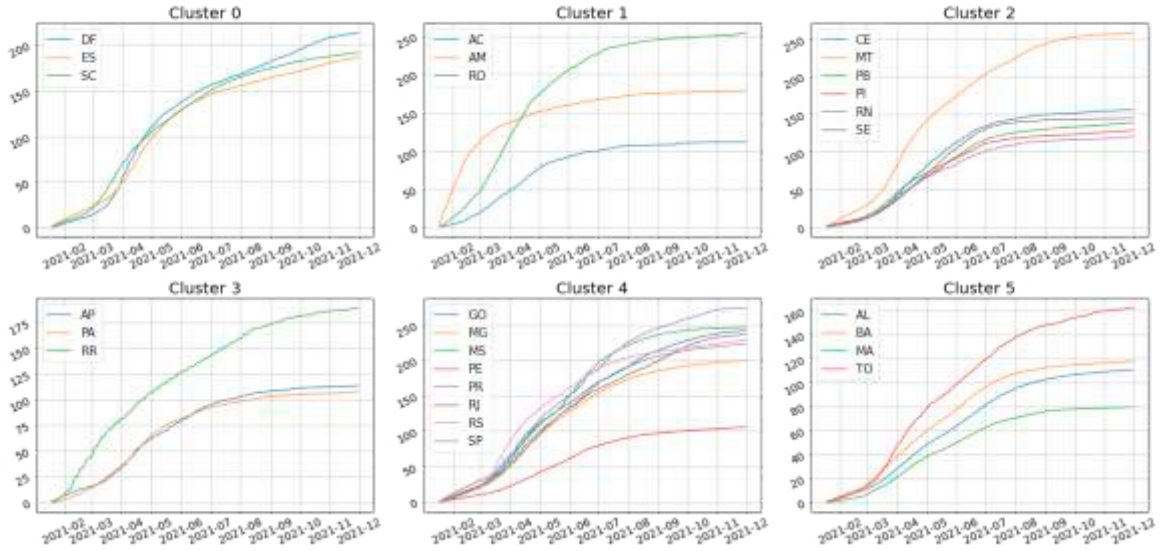

Fig 1. Evolution of deaths per 100k in identified clusters

Figure 2 shows the evolution of the increase in vaccination coverage of the first dose in the states. Analyzing the graph, it is possible to see a similar behavior between the states grouped together during the analysis period. It is also possible to verify the current situation of vaccination coverage, identifying states with low and high numbers in their vaccination coverage and, mainly, how this evolution has occurred.

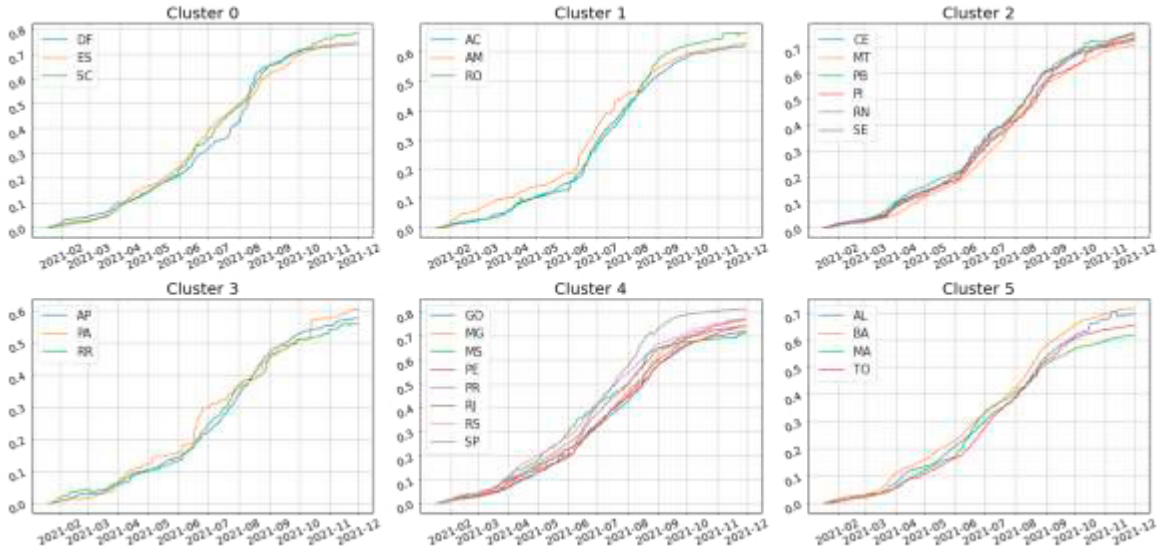

Fig 2. Evolution of first dose in identified clusters

For figure 3, we have data related to vaccine coverage of the second dose, where it is possible to verify a low initial progression, justified by the minimum time needed between the application of the two doses, which can have a high variation depending on the manufacturer.

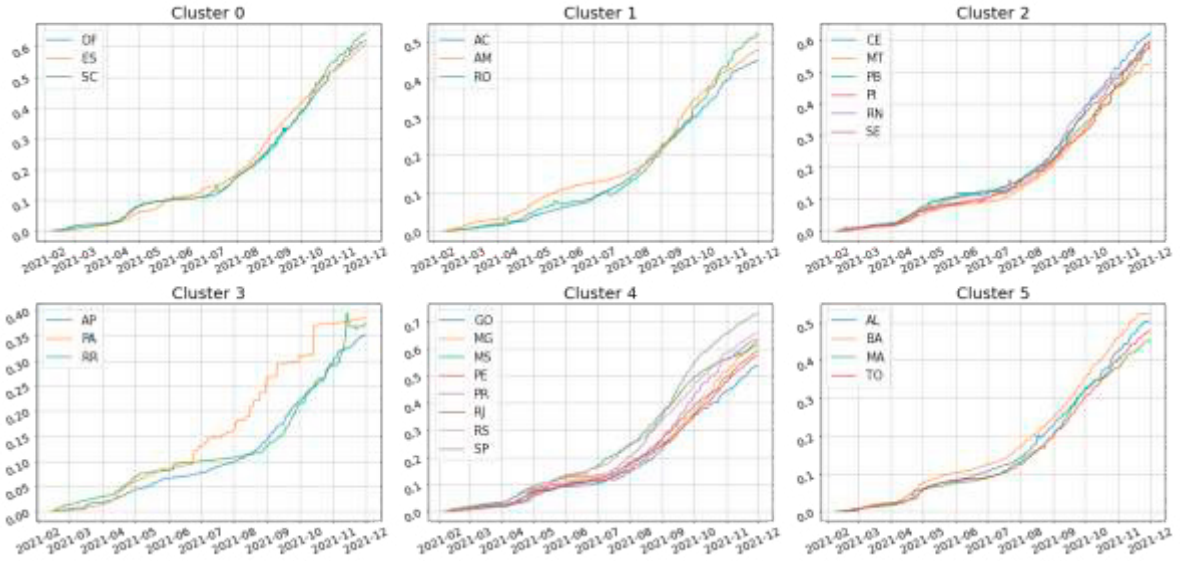

Figure 3. Evolution of second dose in identified clusters

4. Discussion

In the analysis, it is possible to verify that agroupment followed a strong relationship between the vaccination in states, causing a clearly initial split in states that, in the end of the process, reached at least 70% on vaccination coverage in the first dose. Clusters 1 (AC, AM, RO), 3 (AP, PA, RR) e 5 (AL, BA, MA, TO) were the states below this reference value, with cluster 3 being the worst performance, as in the first as in the second dose, followed by clusters 1 and 5. Above this value, there are clusters 0 (DF, ES, SC), 2 (CE, MT, PB, PI, RN, SE), 4 (GO, MG, MS, PE, PR, RJ, RS, SP).

In cluster 0, states had a slow evolution start, and at the end of March 2021, they had an increase in the vaccination numbers, causing a significant change in the rising of its slope, stabilizing numbers at the end the period.

The states of cluster 1 had a different start between the states of AC and RO and the state of AM, for the first two, the evolution of vaccination occurred slowly until the end of April 2021, where there was a small increase until the beginning of June 2021. Since the beginning of the AM state, there was a small rapid increase in the number of vaccines, up to around 4%-5%, rising steadily, but slowly, until the beginning of June 2021 After that, the three states experienced a significant increase, until mid-September 2021, stabilizing the numbers until the end of the period.

The evolution of cluster 3 states started slowly, with two main peaks in the increase in vaccination numbers. The first, at the end of March 2021, contributing with a small increase until reaching a rate of approximately 10%, and the second, at the beginning of June 2021, contributing to the main and significant increase in the vaccination coverage of the states of this cluster, until the beginning of October 2021, where the values stabilized, until the end. For cluster 3 states, the beginning is very slow, with ups and downs, with a gradual increase from mid-March 2021, until it stabilizes in early October 2021, followed by a low increase until the end of the period.

For cluster 4, the states had not very high growth, but constant growth. At the end of March 2021, there was a significant increase in the number of vaccines applied, maintaining a growth similar to an exponential function until the beginning of September 2021, where the numbers stabilized until the end of the period.

For cluster 5, states had a slow start with a small increase after mid-March 2021, with a second most significant increase only in early June 2021. BA and AL states had higher numbers in mid-March 2021 than the other states, but still having a similar behavior at this point and, mainly, in its evolution after that date, with a flattening of the curve in the middle of October 2021 ahead.

Regarding the analysis of the number of deaths per 100k inhabitants, the same behavior was observed, based on the shape and slope of the curves. Clusters 0, 2 and 4 show a similar behavior at the beginning of the analysis period, with a linear growth until the first half of March 2021. From that date, each cluster showed an increase in the number of deaths and had a different evolution from that. In cluster 2, the states experienced an increase from the second half of March 2021 until the beginning of June 2021, where there was a sharp reduction in mortality, flattening the curve until the end of the analysis period. The state of MT presented a high number compared to the average of the group, however, its format and behavior followed the same pattern as the other states. Regarding cluster 4, the flattening of the curve only occurred in early November 2021, maintaining this behavior until the end of the period. In cluster 0, the behavior of the states was similar to cluster 4 from the second half of March 2021, with the main difference being a very clear non-flattening, as in the other clusters, until the end of the analysis period.

Clusters 3 and 5 presented lower numbers at the beginning when compared to the others, having a similar behavior until the beginning of March 2021. From that date, cluster 5 had a smaller variation in the number of deaths until the beginning of the flattening of the curve in July 2021. In cluster 3, there was a small variation in the evolution of the data, until the flattening of its curve after August 2021. The state of RR maintained a similar behavior to the other states in the cluster, presenting a greater variation during its evolution.

5. Conclusion

Brazil is a continental country where there is a large socio-economic difference among Brazilian states. Vaccination is the best strategy to combat the coronavirus spread and our goal was to understand how vaccination occurred in Brazilian states. We carried out an unsupervised analysis, using the DTW technique to understand this scenario and the impacts for the population.

As observed in the previous work[6], the time series clustering technique using Dynamic Time Warping as a similarity metric, that results in a grouping according to the similarity of the shape and slope of the time series curves, related to this characteristic, the clusters have a grouping based on their similarities throughout the evolution of the observed phenomenon.

As shown in the results, vaccination occurred very unevenly in Brazilian states. There is a variation between 55% and 80% in Brazilian states and we were able to identify the clusters that demonstrate the states that had a good vaccination campaign among those that did not. Deaths per 100k inhabitants show a greater variation, this happening because of the numerous population differences.

It is still necessary that new analyzes be carried out, correlating these numbers with human development indices and other socio-economic indicators to better explain the results. Despite this, our work is an innovative effort to understand how vaccination has evolved in Brazil and its impact on different states during the Covid-19 pandemic.

6. References

[1] Alves, Domingos, et. al., Estimativa de casos de Covid-19.Portal Covid-19 Brasil. https://ciis.fmrp.usp.br/covid19.

[2] Neiva, M.B., et al. Brazil: the emerging epicenter of COVID-19 pandemic. Rev. Soc. Bras. Med. Trop, 53 (2020).

[3] Domingues, Carla M.A.S, et al. Programa Nacional de Imunização: a política de introdução de novas vacinas, https://periodicos.unb.br/index.php/rgs/article/view/3331

[4] Domingues, Carla M.A.S; Teixeira, Antônia M. S.: Coberturas vacinais e doenças imunopreveníveis no Brasil no período 1982-2012: avanços e desafios do Programa Nacional de Imunizações.

[5] Domingues, Carla M.A.S, et al.: 46 anos do Programa Nacional de Imunizações: uma história repleta de conquistas e desafios a serem superados. Cad. Saúde Pública 36 (2020).

[6] Cassão, Victor, et. al.: Unsupervised analysis of COVID-19 pandemic evolution in brazilian states. Procedia Computer Science 196, 655-662 (2022).

[7] Cota, Wesley. Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level, https://doi.org/10.1590/SciELOPreprints.362.

[8] Neiva, M.B., et al. Brazil: the emerging epicenter of COVID-19 pandemic. Revista da Sociedade Brasileira de Medicina Tropical, 53 (2020).

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

[10] Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T. (2015) Time-series clustering – A decade review. Inf Syst 2015;53:16–38. https://doi.org/10.1016/j.is.2015.04.007.

[11] Na S, Xumin L, Yong G. (2010) Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, IEEE; 2010, p. 63–7. https://doi.org/10.1109/IITSI.2010.74.

[12] 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.

[13] Niennattrakul, Vit, Ratanamahatana, Chotirat A.: On clustering multimedia time series data using k-means and dynamic time warping. International Conference on Multimedia and Ubiquitous Engineering (MUE'07), 733-738 (2007).

[14] Nascimento MLF. (2020) "A multivariate analysis on spatiotemporal evolution of Covid-19 in Brazil." Infect Dis Model. 5:670–80.

[15] 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.

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

[17] Frawley J. William, et al. Knowledge Discovery in Databases: An Overview. Relational Data Mining, Springer, 28-47 (2001).