International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-024-00542-9

REGULAR PAPER

K-means DTW Barycenter Averaging: a clustering analysis of COVID-19 cases and deaths on the Brazilian federal units

Jonatas Silva do Espirito Santo[1,2] · Jackson Santos da Conceição[1,2] · Lilia Carolina Carneiro da Costa[1,3] · Rosemeire Leovigildo Fiaccone[1,3] · Marcos Ennes Barreto[1,4] · Maria Yury Ichihara[1] · Anderson Ara[1,5]

Received: 20 December 2023 / Accepted: 21 March 2024

Abstract

A challenge faced while monitoring the COVID-19 pandemic in Brazil is the identification of patterns of incidence and mortality, which can help prioritize interventions to avoid excessive disease transmission and associated deaths. This study aimed to identify epidemiological patterns concerning the evolution of the pandemic among Brazilian federal units (states). The proposed methodology is based on a combination of non-hierarchical k-means clustering and dynamic time warping (DTW), used to measure distances among time series, with the subsequent use of the DTW Barycenter Averaging (DBA) algorithm to calculate cluster centroids for time series of variable lengths. The dataset used is a time series consisting of the number of new cases and deaths per epidemiological week, and the number of cumulative cases and deaths until a given epidemiological week for each of the 27 Brazilian federal units. Six groups of Brazilian federation units were formed based on the similarities between the prevalence and incidence curves. The results demonstrate efficiency with respect to the characterization of both COVID-19 cases and rates of mortality.

Keywords COVID-19 · Clustering · K-means · DTW Barycenter Averaging

1 Introduction

In Brazil, the COVID-19 pandemic began on February 26, 2020, following the first case notification in state of São Paulo [1]. During the first 105 epidemiological weeks following the introduction of SARS-CoV2, approximately three waves occurred, characterized by periods of growth, followed by stabilization and then decreasing numbers of cases and deaths [2, 3]. The first wave associated with the variant of concern Alpha occurred between epidemiological weeks 9 (23/02/2020 to 29/02/2020) and 45 (01/11/2020 to 07/11/2020). The second wave, more pronounced than its predecessor, as a result of the Gama variant, the introduction of the Delta variant and more frequent social gatherings coinciding with end-of-year celebrations and relaxed masking mandates, occurred between epidemiologi-

Jackson Santos da Conceição, Lilia Carolina Carneiro da Costa, Rosemeire Leovigildo Fiaccone, Marcos Ennes Barreto, Maria Yury Ichihara and Anderson Ara have contributed equally to this work.

✉ Jonatas Silva do Espirito Santo
jonates@gmail.com

Jackson Santos da Conceição
jsantos1013@gmail.com

Lilia Carolina Carneiro da Costa
liliacosta@ufba.br

Rosemeire Leovigildo Fiaccone
fiaccone@ufba.br

Marcos Ennes Barreto
m.e.barreto@lse.ac.uk

Maria Yury Ichihara
maria.yury@fiocruz.br

Anderson Ara
ara@ufpr.br

[1] Centre of Data and Knowledge Integration for Health (CIDACS), Gonçalo Moniz Institute, Oswaldo Cruz Foundation, Salvador, Bahia 41745-715, Brazil

[2] Superintendence of Economic and Social Studies of Bahia, Government of bahia, Salvador, Bahia 41745-002, Brazil

[3] Department of Statistics, Federal University of Bahia, Salvador, Bahia 40170-110, Brazil

[4] Data Science Institute, London School of Economics and Political Science, London, Greater London 610101, England

[5] Department of Statistics, Federal University of Paraná, Curitiba, Paraná 81531-980, Brazil

Published online: 14 April 2024

cal weeks 46 (08/11/2020 to 14/11/2020) and 51 (19/12/2021 to 25/12/2021). A third wave was observed between epidemiological weeks 52 (26/12/2021 a 01/01/2022) in 2021 and 8 (20/02/2022 a 26/02/2022) in 2022, during which time the Omicron variant predominated [2, 4]. The greater transmission capability of this variant caused an alarming impact in the number of cases and deaths within the context of belated vaccination efforts around the country, which were initiated on January 17, 2021.

The present study considers the COVID-19 pandemic period between 25/02/2020 and 01/01/2022, during which more than 22.2 million cases were confirmed in Brazil, leading to approximately 620,000 deaths, according to available data by [5]. The study aims to identify epidemiological patterns among the states in an attempt to better understand related factors and provide support for the adoption of timely preventive measures by health authorities. More comprehensive knowledge on the impact of the COVID-19 pandemic is of great interest to both government as well as civil society notorious. Multivariate clustering methods, also known as unsupervised machine learning algorithms, are an important tool to aid in the grouping of similar data. In particular, the aggregation of observations in accordance with characteristic similarities, as well as the identification of differences among groups [6], are both procedures that can help depict behaviors in different geographic regions with respect to COVID-19 cases and deaths. To this end, the patterns exhibited during the pandemic can be characterized by classifying and describing groupings among the cases and deaths that occurred in each Brazilian federation unit (FU). Clustering methods employ measures of distances and/or (dis)similarities among data. The literature contains several reports on distance measures, with Euclidean distance being the most prevalent [7]. However, this distance measurement method is sensitive to distortions along the time axis, and, when using this metric to measure distance between two time series, temporal effects may not be recognized. To address this limitation, some studies such as [8–10] have employed dynamic time warping (DTW), a widely used metric capable of measuring distances between time series for classification and clustering purposes.

Introduced by a community of researchers working on voice processing in the 1970s, and originally described in [11, 12], DTW is a dynamic programming algorithm for time-series analysis that provides results as a function of dissimilarity, which can be used for distance measures in both unsupervised and supervised machine learning algorithms. Thus, distance is capable of capturing temporal dependencies in analyzed data, as well as enabling the calculation of distance between two time series with differing lengths. However, this approach requires greater computational power compared to Euclidean distance calculation, which is easier to implement and implies lower computational cost. Some approaches have been developed to reduce the processing time associated with the DTW algorithm by introducing restrictions on the diagonal of the distance matrix in which the DTW algorithm is executed, such as the Itakura parallelogram [13], the SakoeChiba band [12] and the Ratanamahatana-Keogh band [14]. DTW makes it possible to implement unsupervised learning methods to classify time series into groups, such as hierarchical or k-medoids clustering. For time-series means calculations, which may serve as centroids, and thus permit the use of the k-means method, several proposals have been developed. Niennattrakul and Ratanamahatana [15, 16] demonstrated that these approaches may generate imprecise averaging, which impacts convergence. Petitjean et al. [17] proposed DBA (DTW Barycenter Average), which provides a global average from a time series based on DTW, thereby constituting a more consistent alternative to previous methods. In their paper, Petitjean et al [18] used DBA to implement the k-means algorithm but did not investigate the impact of the initial series selection on the centroid calculation, nor consider the behavior of their method in time series with different lengths. Jang et al. [19] proposed the use of the k-means clustering algorithm based on DTW to generate reliable reference patterns in the application of accelerometer-based gesture recognition. Anh and Thanh [20] evaluated the performance of methods to assess clustering using k-means for time-series data using DTW distance analysis, proposing an efficient method based on Barycenter Averaging (DBA) to calculate time-series means, as well as medians to determine initial centroids for k-means clustering.

More recently, [18] studied the effect of using DBA in a nearest centroid algorithm by adapting the nearest-neighbor classification algorithm to employ DTW distance with less computational cost. Forestier et al. [21] used DBA to generate time-series means to augment sparse data sets. Cuturi and Blondel [22] proposed a method denominated soft-dtw, which provides differentiable loss function that is essential to machine learning algorithms, such as neural networks. Leodolter et al. [23] proposed an incremental way of calculating DTW distances for different sets of time-series pairs with significantly reduced computational cost. Furthermore, [24] employed a type of unsupervised artificial neural network, self-organizing Kohonen maps (SOM).

The present investigation utilized a variation on the traditional, non-hierarchical clustering method k-means, in which DTW is employed to define distances between time series and DTW Barycenter Averaging (DBA) to define the centroids of the identified groups, so that the units of analysis originate from a longitudinal rather than cross-sectional perspective. In addition, we used a variant of the traditional metric of internal dispersion, herein calculated using DTW and DBA, to assess the homogeneity of the groups and the efficacy of grouping, as well as to select k number of groups. It is worth

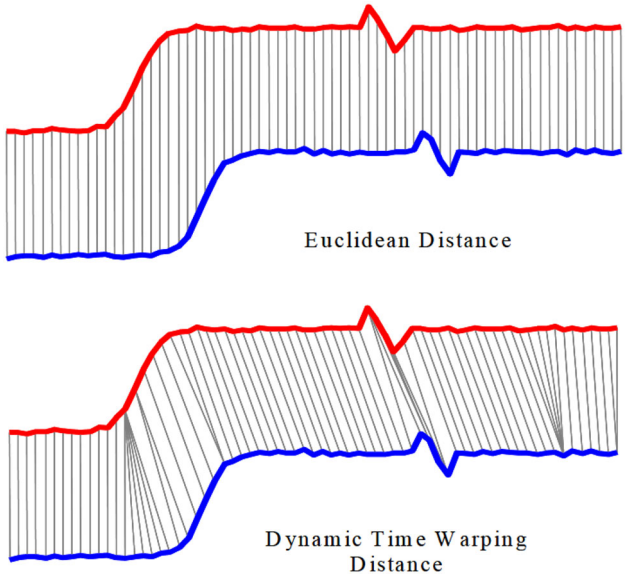


Fig. 1 Alignment of two time series using Euclidean distance and dynamic time warping (DTW) [14]

mentioning that herein k-means clustering was performed based on DTW and DBA in a set of time-series data with differing lengths.

Accordingly, our objective was to present the epidemiological curves of the COVID-19 pandemic grouped by Brazilian federation units using the DTW-DBA clustering method. This article is organized as follows. Section 2 describes the clustering proposed method. Section 3 presents a simulation study, in which the clustering method is applied to artificially generate labeled time-series data sets in order to evaluate the performance of the employed method. Section 4 describes the application of the method on COVID-19 curves among the Brazilian Federation Units. Section 5 closes the paper with the final comments.

2 *K*-means DTW Barycenter Averaging

Most clustering methods require the use of a similarity metric to compare instances, which in this case are time series. The alignment of two time series using Euclidean distance implies one-to-one point scanning, while DTW distance employs many-to-many. Consequently, DTW is able to capture temporal dependencies even when these occur at different times during the analyzed series, making it also possible to measure similarity between series of different lengths, as illustrated in Fig. 1.

Considering $\mathbf{X} = X_1, X_2, ..., X_p$ and $\mathbf{Y} = Y_1, Y_2, ..., Y_q$ as two distinct series, the DTW algorithm uses a matrix of order $p \times q$, $M_{p\times q}$, in which each element $[[m_{(i,j)}]]$ is defined by the distance $\delta(X_i, Y_j)$ between the points X_i and Y_j, $i = 1, ..., p$ and $j = 1, ..., q$. The distance adopted for the construction of the matrix can be a variation of Minkowski distance, also known as L_α, while Euclidean distance has been particularly and widely used in cases when $\alpha = 2$, or Manhattan distance when $\alpha = 1$. It is worth noting that for matrix construction in univariate time series, the distance between any two points X_i and Y_j can be summarized as:

$$\delta(X_i, Y_j) =\mid X_i - Y_j \mid .$$

Using the distance matrix, a path is generated from the starting point $(1, 1)$ to point (p, q), modified in relation to the diagonal - which would be the same path followed by Euclidean distance if there were a square matrix of order p. The path $W = w_1, w_2, ..., w_k, ..., w_r$, with w_k corresponding to each position $(i, j)_k$ in the matrix, is chosen in order to minimize the DTW distance between the time series $\mathbf{X} = X_1, X_2, ..., X_p$ and $\mathbf{Y} = Y_1, Y_2, ..., Y_q$,

$$\text{DTW}(\mathbf{X}, \mathbf{Y}) = \min_W [\sum_{k=1}^{r} \delta(w_k)].$$

DTW calculations are performed through the dynamic programming of $\gamma(i, j)$, which represents the cumulative sum of the distance at point (i, j) of the matrix, such that

$$\gamma(i, j) = \delta(i, j) + \min[\gamma(i-1, j), \gamma(i-1, j-1), \gamma(i, j-1)].$$

In the context of DTW, the function of the components of matrix $\gamma(i, j)$ traces a modified path in relation to Euclidean distance, but generally does not stray far from the diagonal. The most commonly used global constraints in the literature are the Sakoe–Chiba band [12] and the Itakura parallelogram [13]. Ratanamahatana and Keogh [14] reviewed the various constraints available and also proposed the Ratanamahatana–Keogh band. These three constraints are illustrated in Fig. 2.

The present investigation utilized the Sakoe–Chiba uniform band, which historically, according to [14], is assumed to provide a window (band) size corresponding to nearly 10% of the length of the series.

Thus, as proposed by [17], DTW Barycenter Averaging was used to determine centroids, in which, based on DTW, provides a consistent way of calculating time-series means. By definition, DBA is a global averaging method that consists of iteratively improving an initial sequence (selected among the set of series from which the centroid will be calculated) with the aim of minimizing the within-cluster (group) sum of square sum of squares (WCSS), i.e., DTW distances between the time-series mean and the set of time series being analyzed

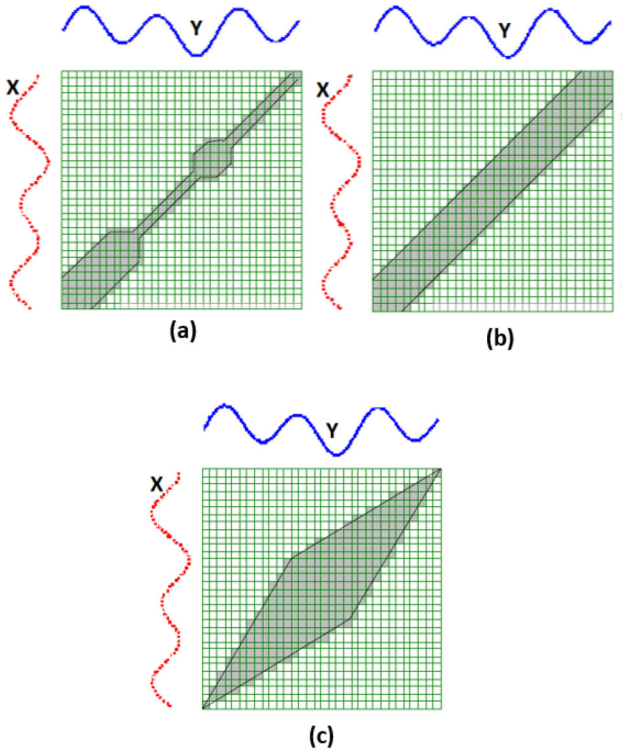


Fig. 2 **a** Ratanamahatana–Keogh Band; **b** Sakoe–Chiba Band; **c** Itakura Parallelogram [14]

[25]. In this way, the updated time-series mean is defined once all barycenters are calculated.

As demonstrated by [17], it is more efficient to use an element of the time series set to start the calculation of the mean rather than to randomly generate a series. In each iteration of the DBA algorithm, two actions are performed:

1. Calculation of the DTW between each time series and the temporary series mean to be updated in order to identify associations between their coordinates.
2. Update each coordinate of the time-series mean using the barycenter of the coordinates associated with it during the first step

It is important to mention that as the size of the resulting time-series mean is equal to the size of the initial series, it is therefore recommended to select the reference series as the one with the longest length in the dataset. As an example, Fig. 3a represents the DBA calculated for the series of accumulated COVID-19 cases per 100,000 inhabitants in the FUs of Bahia and Acre, while Fig. 3b represents the DBA calculated for deaths per 100,000 inhabitants in the FUs of Rio Grande do Sul and Santa Catarina.

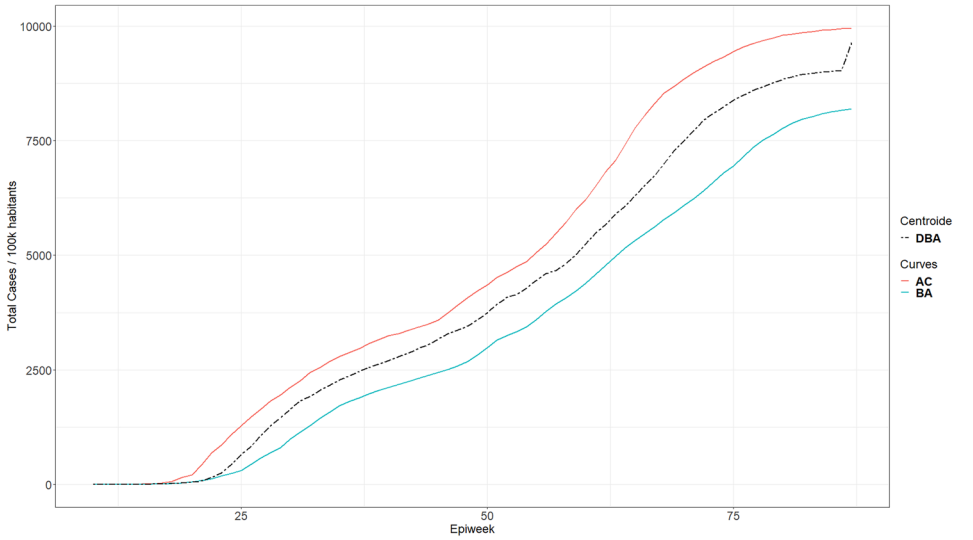
In general, with the objective of grouping a set of time series into k homogeneous groups, with k number of groups constituting a hyperparameter that must be defined a priori by the researcher, the k-means method was used with DTW distance measuring and DBA to determine the centroids of each group in accordance with the pseudo-code described below in Algorithm 1.

Algorithm 1 Dynamic Time Warping K-means

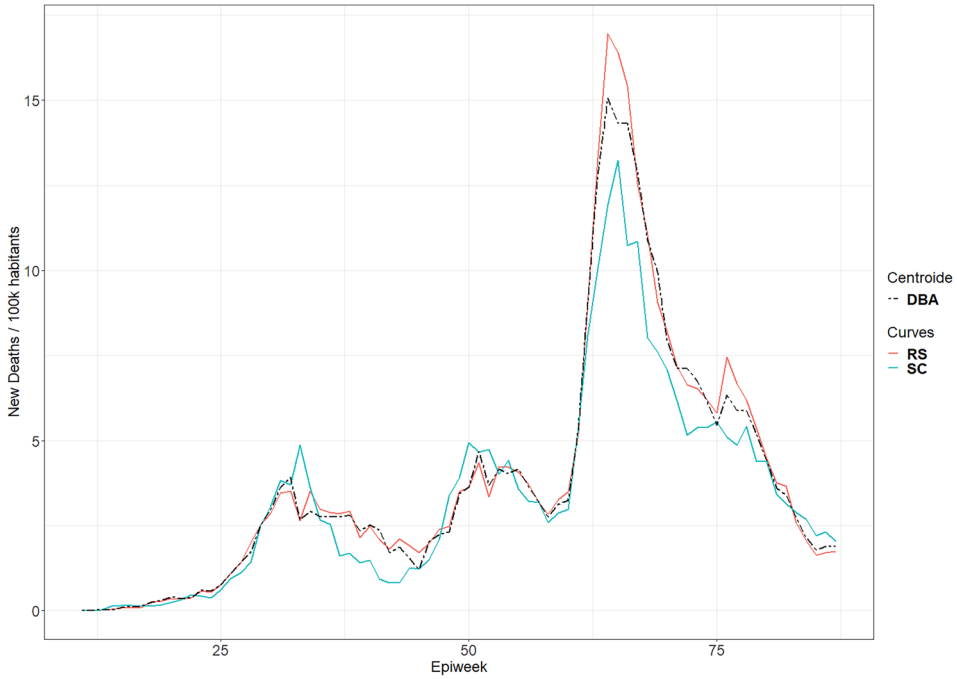
Input: k **// number of groups;**
Input: ϵ **// stop criteria;**
Input: $\mathbb{S} = \mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^n$, in which $\mathbf{S}^j = \{S_1^j, ..., S_{T_j}^j\}, \forall j = (1, ..., n)$**// time series;**
Output: partition G of data

$\mathbf{C}_{global} \leftarrow \text{DBA}(\mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^n)$
$WCSS_{before} \leftarrow 0$
for $j \leftarrow 1$ **to** n **do**
  $WCSS_{before} \leftarrow WCSS_{before} + [\text{DTW}(\mathbf{S}^j, \mathbf{C}_{global})]^2$
end for
for $i \rightarrow 1$ **to** k **do**
  $j \leftarrow sample(1:n)$
  $\mathbf{C}^i \leftarrow \mathbf{S}^j$
  $G_i \leftarrow \emptyset$
end for
while $\phi \geq \epsilon$ **do**
  for $m \leftarrow 1$ **to** n **do**
    for $i \leftarrow 1$ **to** k **do**
      $d_i \leftarrow \text{DTW}(\mathbf{S}^m, \mathbf{C}^i)$
    end for
    $w \leftarrow argmin_i(d_i)$
    $G_w \leftarrow \mathbf{S}^m$
  end for
  $WCSS_{after} \leftarrow 0$
  for $i \leftarrow 1$ **to** k **do**
    $\mathbf{C}^i \leftarrow \text{DBA}(G_i)$
    for $j \leftarrow 1$ **to** n **do**
      if $\mathbf{S}^j \in G_i$ **then**
        $WCSS_{after} \leftarrow WCSS_{after} + [\text{DTW}(\mathbf{S}^j, \mathbf{C}^i)]^2$
      end if
    end for
  end for
  $\phi \leftarrow \mid (WCSS_{before} \div WCSS_{after}) - 1 \mid$
  $WCSS_{before} \leftarrow WCSS_{after}$
end while

Accordingly, k instances of the time series set under analysis must initially be chosen, which will be used as initial centroids. Next, the distance from each observation in the dataset to each of the k centroids is calculated. Then, for each series, the closest centroid (i.e., the one with the shortest DTW distance between the series and the centroid) is verified, and the series is then allocated to this group. After the groups are formed, their centroid is calculated based on the observations allocated to these k groups. Iteratively, the steps of calculating the distance of the observations to the centroid are repeated, while reallocating the observations in



(a)



(b)

Fig. 3 Representation of DTW Barycenter Averaging considering COVID-19 indicator curves in Brazilian Federative Units. **a** Total cases per 100k inhabitants in BA and AC calculated using DBA. **b** New deaths per 100k inhabitants in RS and SC with DBA

groups where the distance to the centroid was the smallest and recalculating the centroid a finite number of times until the most homogeneous groups possible are formed.

To assess the variability of the formed groups, we adapted a measure previously established in the literature [25] that is widely used in cluster analysis applications: the within-cluster sum of square distance (WCSS) display in Eq. 1.

$$\text{WGSS} = \sum_{i=1}^{k} \sum_{\mathbf{S}^j \in \mathbf{C}_i} \text{DTW}(\mathbf{S}^j, \mathbf{C}^i) \tag{1}$$

which $\mathbf{S}^j = \{S_1^j, ..., S_{T_j}^j\}$ represents a time series of length T_j and $\text{DTW}(\mathbf{S}^j, \mathbf{C}^i)$ is the DTW value between the time series $\mathbf{S}^j$ e centroid of the ith group $\mathbf{C}^i$.

A study by [26], in which DTW was utilized to hierarchically group time series, with 70 different time-series data sets used for benchmarking, demonstrated that WCSS, despite being defined in Euclidean space, could be adapted to incorporate distance and centroid measures different from those originally conceived by [25]. Moreover, this measure was found to be efficient when used as a criterion for defining k number of groups.

The algorithm was implemented in R programming language [27], using the dtw package [28] to calculate distances among time series, dtwclust [29] to calculate DBA centroids, and tidyverse [30] for data reading, cleaning, processing and visualization. The computational equipment utilized was a 4th generation Intel®Core™ i7 processor with 16GB of RAM. The R code is available at GitHub https://github.com/cidacslab/Ids-cluster.

3 Simulation studies

3.1 Scenario 1

To evaluate our method's effectiveness, we first conducted a simulation study by generating 100 time series from Gaussian auto-regressive moving average (ARMA) models (1,1), $\mathbf{X}^1, ..., \mathbf{X}^{100}$; these were divided into five groups of 20 series (a, b, c, d, e) with each series containing 100 observations:

$$X_t = \epsilon_t + \phi X_{t-i} + \theta \epsilon_{t-1} \tag{2}$$

with $t = 1, ..., 100$, and the ϕ and θ parameters fixed for each of the five groups. The series was generated via simulations of the models described herein, with implementation in R language using the *arima.sim* function of the package *stats*, part of the core R library [27].

For group a, $\phi = 0.1$ and $\theta = 0.3$; for group b, $\phi = 0.1$ and $\theta = 0.6$; for group c, $\phi = 0.5$ and $\theta = 0.3$; for group d, $\phi = 0.5$ and $\theta = 0.6$; for group e, $\phi = 0.6$ and $\theta = 0.5$.

Thus, the obtained groups had homogeneous series within each individual group, but the series were also heterogeneous considering the five groups. This heterogeneity between groups refers not only to variability, but also to the mean corresponding to each series, as shown in Fig. 4.

Next, we applied the clustering method described in Section , in which the simulated series were submitted to k-means clustering with DTW distances and centroids obtained by DBA, with stop criterion equal to 0.0001, providing convergence with only five interactions. As expected, as illustrated in Fig. 5, the method correctly grouped all 100 time series, forming 5 clusters in which each cluster contained the time series from just one group originally generated in the simulation, thereby demonstrating the method's efficacy.

3.2 Scenario 2

To further evaluate the method using a new set of time-series data with different behavior and complexity than the first simulation, 100 series were generated around zero, with a length of 100, divided into four different groups in relation to the different parameterizations. Each group contained 25 series generated with the same structure, thus ensuring similarity within each group. In the first group, series with **AR**(2) structure were generated, while the second employed **MA**(1). The series in the third group were generated using **ARMA**(2,1), while the fourth group used **ARMA**(2,2):

- **AR(2)**: A second-order autoregressive model, with $|\ \phi_1\ | < 1$ and $|\ \phi_2\ | < 1$.

$$X_t = \sum_{i=1}^{2} \phi_i X_{t-i} + \epsilon_i \tag{3}$$

- **MA(1)**: A model with first-order moving averages, with $|\ \theta_1\ | < 1$.

$$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} \tag{4}$$

- **ARMA(2,1)**: A model of autoregressive terms equal to 2 and moving average terms equal to 1.

$$X_t = \epsilon_t + \sum_{i=1}^{2} \phi_i X_{t-i} + \theta_1 \epsilon_{t-1} \tag{5}$$

- **ARMA(2,2)**: A model with autoregressive terms equal to 2 and moving average terms equal to 2:

$$X_t = \epsilon_t + \sum_{i=1}^{2} \phi_i X_{t-i} + \sum_{i=1}^{2} \theta_i \epsilon_{t-i} \tag{6}$$

The fixed ϕ and θ parameters for each group and the vectors of the parameters were combined among themselves, always in conformity with the unit root pross [31].

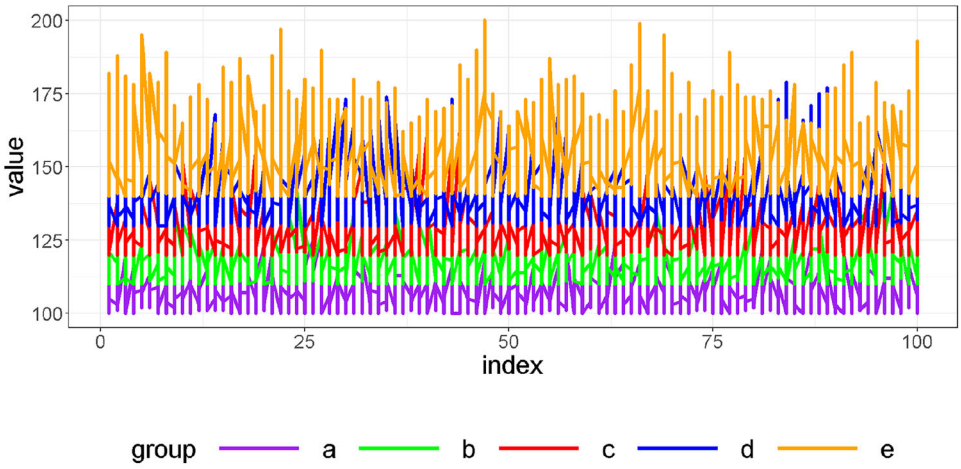
For AR(2) models, $\phi_1 = 0.1$ and $\phi_2 = 0.3$; for MA(1), $\theta_1 = 0.6$; for ARMA(2,1), $\phi_1 = 0.5$, $\phi_2 = 0.1$ and $\theta_1 = 0.6$; for ARMA(2,2), $\phi_1 = 0.3$, $\phi_2 = 0.2$, $\theta_1 = 2.5$ and $\theta_2 = 0.6$.

The behavior of the simulated series is shown in Fig. 6. Figure 6b presents each group separately, clearly illustrating the differences between the groups, despite values varying close to zero. Figure 6a presents a composite image of the simulated series centered around zero.

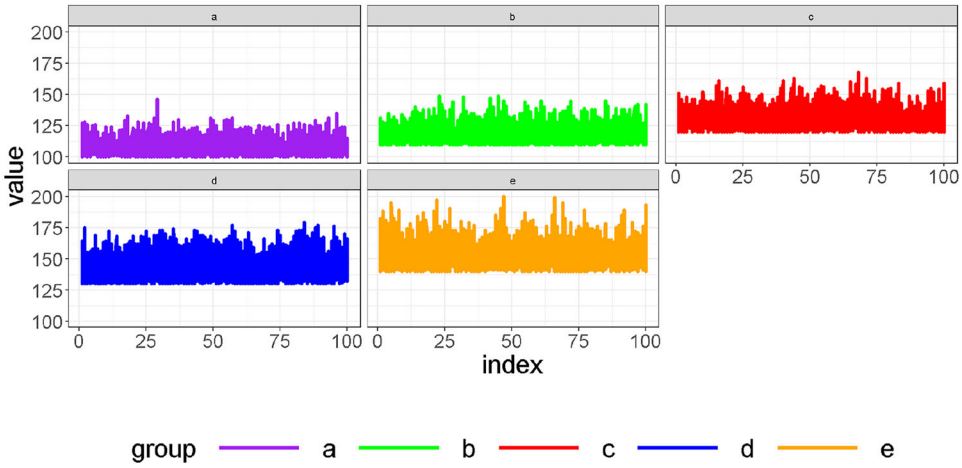
Despite the increased degree of difficulty due to overlapping series values around zero, the clustering method identified four distinct groups that correctly corresponded to the original groups of series (Fig. 7).

3.3 Discussion on simulation studies

The simulation studies consider two different scenarios generated by Monte Carlo sampling with ARMA structures. Each scenario considered 100 time series that were clustered into different numbers of groups. In the both cases, the validation was effective in the sense that the real cluster was correctly classified for each time series. The same result was found to other generated ARMA structures.



(a)



(b)

Fig. 4 Simulation involving five groups of series, presented both jointly (**a**) and separately (**b**)

4 Real data application

The present study analyzed the following time series: Total cases per 100k inhabitants (number of cumulative cases of COVID-19 in a Brazilian federation unit, divided by its total population multiplied by 100), denominated herein as prevalence; total cases per 100k inhabitants (number of new COVID-19 cases in a Brazilian federation unit, divided by its total population multiplied by 100), herein referred to as incidence; a time series of indicators of COVID-19 mortality in Brazilian federation units (states and the Federal District): New deaths per 100k inhabitants (number of new deaths due to COVID-19 in an epidemiological week occurring in a federation unit, divided by its total population multiplied by 100); Total deaths per 100k inhabitants (number of cumulative deaths due to COVID-19 until a given epidemiological week in a federation unit, divided by its total population multiplied by 100).

These time series were constructed using data extracted from the COVID-19 repository made available by [5].

For each of the variables of interest, there were 27 time series, one corresponding to each Brazilian FU. Each point of the time series is the observation of the variable of interest in a given epidemiological week, with the first confirmed case of COVID-19 occurring in the state of São Paulo on February 25, 2020. However, the first case of COVID-19 was only confirmed on March 5, 2020, in the state of Rio de Janeiro, and on March 6, 2020, in the state of Bahia. The

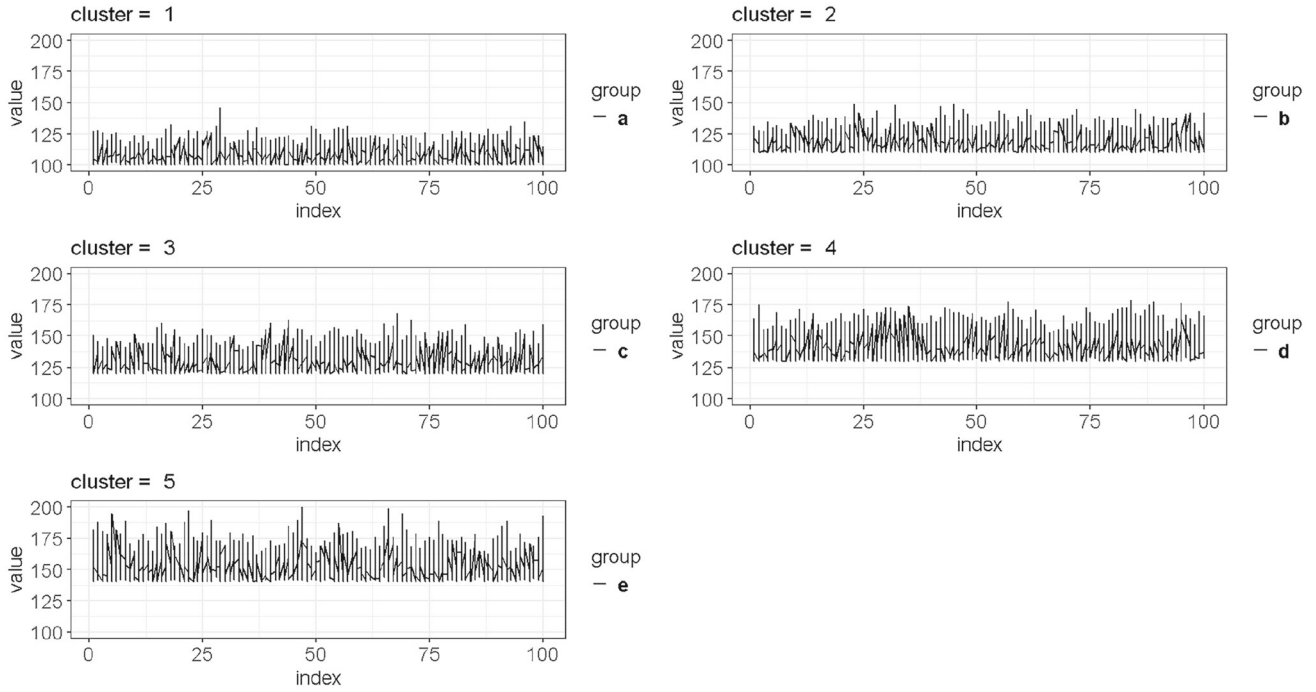


Fig. 5 Clustering of the time-series curves simulated with different parameterizations

first death related to COVID-19 was recorded in the state of São Paulo on March 17, 2020, on March 18 in the Rio de Janeiro, and on March 24, 2020, in the state of Amazonas. Thus, the time series corresponding to each FU began in different epidemiological weeks, regardless of COVID-19 case prevalence and incidence, or even mortality.

In addition to the behavior of the curves across all epidemiological weeks, the levels of the prevalence and cumulative mortality curves, as well as the peaks of the incidence and mortality rate series, constitute relevant information in determining similarity among the observations under analysis.

Accordingly, recognizing the importance of the initial centroid selection step, we employed the following criteria: firstly the maximum values of each series were selected, and k percentiles were then selected from this vector, e.g., if $k = 3$, the 10th, 50th and 90th percentiles were selected. Hence, the curves of the Brazilian FUs associated with the selected quantiles were used as initial centroids.

Next, DTW distance was calculated among the 27 time series and k centroid curves. Then, each of the 27 curves was allocated to the closest group based on the shortest DTW distance between the curve and the k centroids. Finally, for each of the k groups, using the corresponding series, DBA values were calculated, which then served as the new centroids of the group. Through an iterative process of calculating DTW distances between the curves and the centroids, the curves were regrouped and new DBAs were calculated. We used a Sakoe–Chiba band of size 10. The stopping criteria adopted for the k-means algorithm was a WCSS percentage reduction less than $\epsilon = 10^{-16}$ between one iteration and the next.

4.1 Incidence rate of COVID-19 cases in Brazilian Federal Units

Data corresponding to two COVID-19 case indicators, weekly numbers of total cases and new cases per 100k inhabitants, were analyzed for each Brazilian Federation Unit. For each of these two indicators, clustering analysis was performed with k ranging from 2 to 10, with the optimal number of groups selected so that any increase in this number would not significantly reduce WCSS. Thus, in accordance with these criteria and the results presented in Fig. 8a, b, six groups were considered.

Upon analysis of the results depicted in Fig. 9, it becomes evident that the first cluster is formed by Brazilian FUs presenting lower levels of COVID-19 case prevalence at the end of the analyzed period, while the others increase in terms of intensity until the final group consists solely of the State of Roraima, with prevalence rates exceeding 20,000 cases per 100K inhabitants.

Also of note is the similarity among the curves within each group. The obtainment of heterogeneous groups with homogeneity observable within each group was the expected result when applying the clustering method. As seen in Fig. 10, the first cluster, that containing the FUs with the lowest levels of COVID-19 case records per 100K inhabitants, is formed by the states of Alagoas, Maranhão, Pará, Pernambuco and Rio

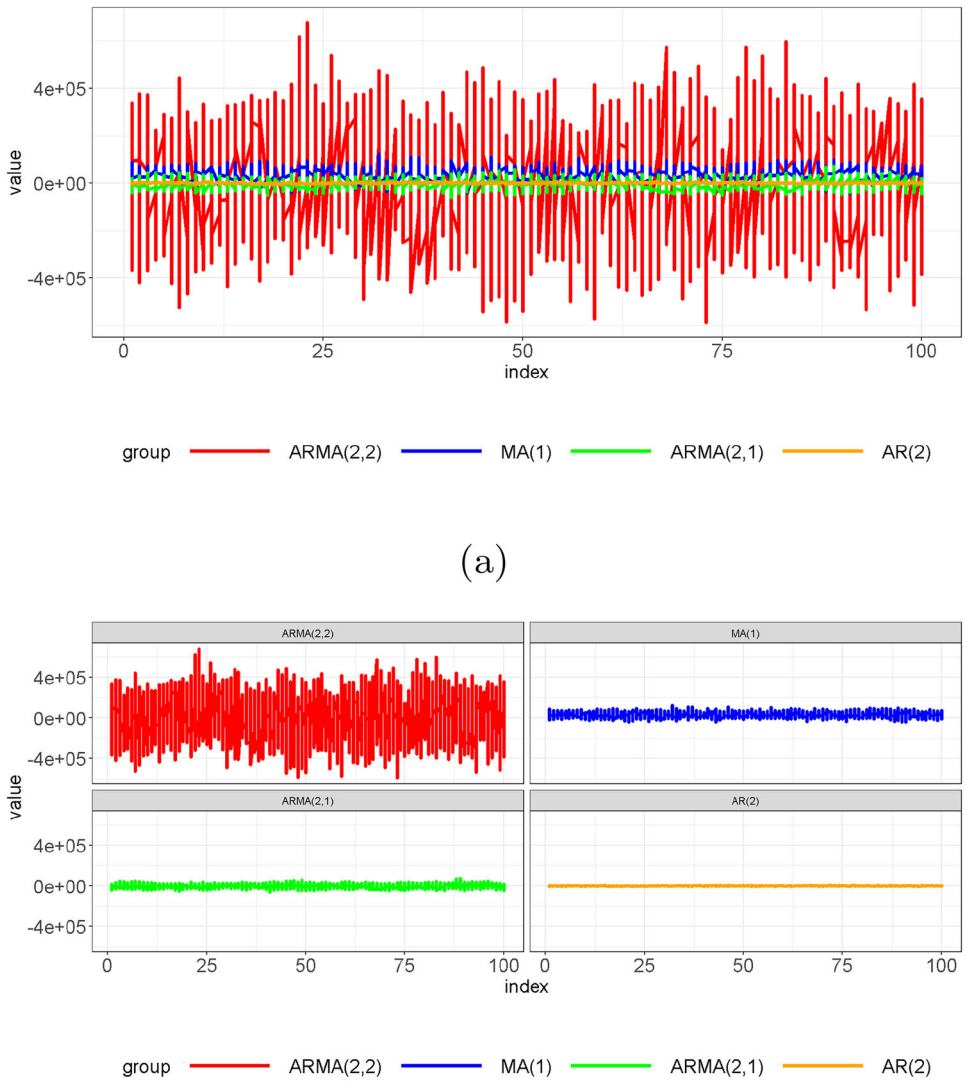


Fig. 6 Second simulated series of four groups, presented jointly (**a**) and separately (**b**)

de Janeiro; the second is formed by Bahia and São Paulo, while the third group consists of the states of Acre, Amazonas, Ceará, Minas Gerais, Paraíba, Piauí and Rio Grande do Norte. The fourth is composed of Goiás, Mato Grosso do Sul, Paraná, Rio Grande do Sul and Sergipe. The fifth cluster comprises Amapá, the Federal District, Espírito Santo, Mato Grosso, Rondônia, Santa Catarina and Tocantins. Isolated from the other states in Brazil, Roraima alone forms the sixth group, with the highest level of recorded COVID-19 cases in Brazil.

Figure 11 illustrates curves corresponding to new weekly cases per 100K inhabitants, demonstrating the evolution of COVID-19 case records among the Brazilian FUs. These curves are more susceptible to alterations due to external factors that affect the evolution of the disease, such as data source. The fact that the data used herein originates from official disclosures by regional Health Secretariats, with data registered according to date of publication, produces significant distortions, such as what is observed in Rio Grande do Norte. Due to a system change, 36,374 new cases were recorded in just one day, June 22, 2021, which produced a marked leptokurtic deformation corresponding to a peak of more than 100 cases per 100K inhabitants in the respective epidemiological week. It is worth mentioning that this event occurred because the data source did not consider the date on which symptoms first appeared as a criterion, but rather the date on which the number of confirmed cases was released.

Figure 12 illustrates clustering among the Brazilian FUs in terms of new weekly cases per 100K inhabitants. The first cluster is formed by the states of Alagoas, Bahia, Maranhão,

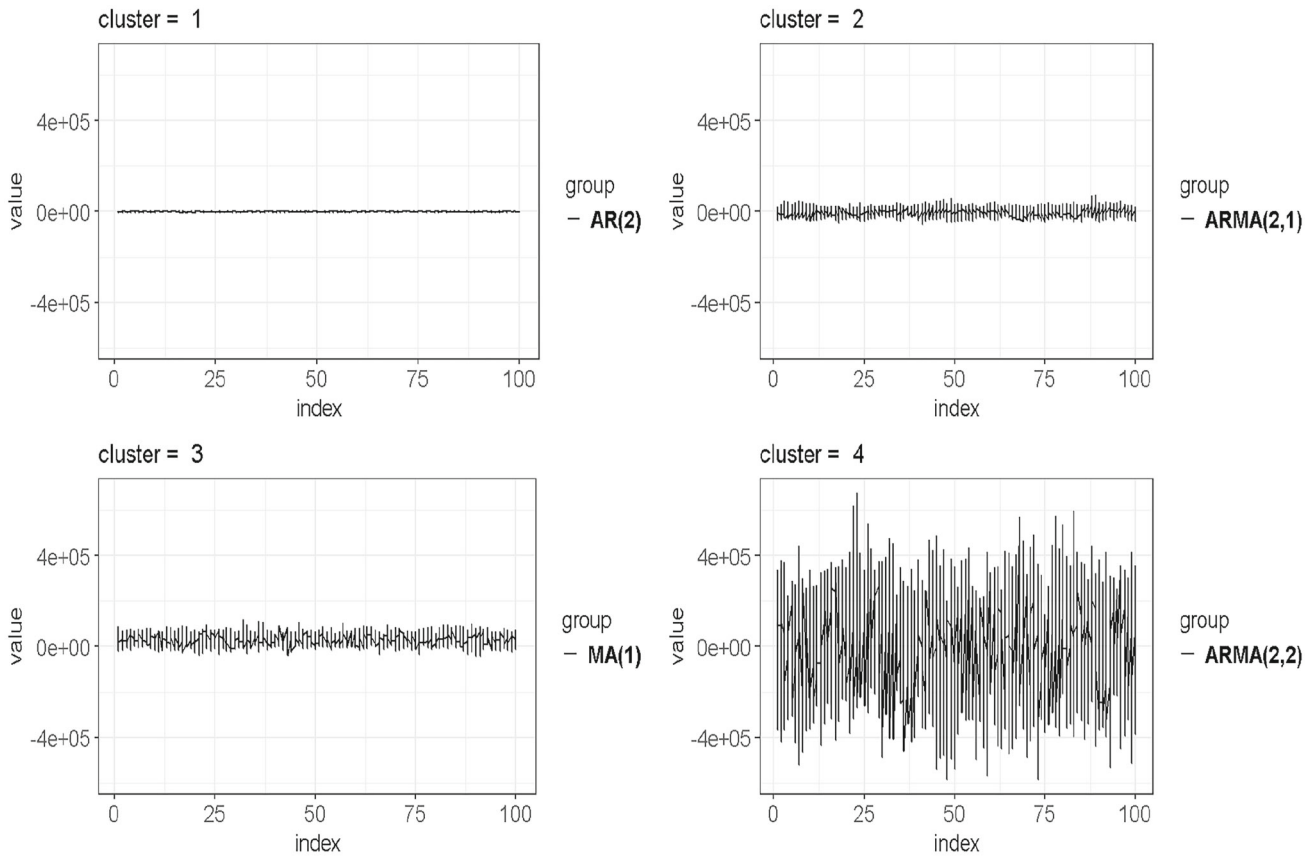


Fig. 7 Grouping of simulated time-series curves with different parameterization

Pará, Pernambuco and Rio de Janeiro, while the second is formed by Acre, Ceará, Goiás, Mato Grosso do Sul, Minas Gerais, Paraíba, Piauí, Rio Grande do Sul and São Paulo. The third cluster contains the states of Amazonas, Espírito Santo, Mato Grosso, Rondônia, Santa Catarina and Sergipe, and the fourth cluster consists of Amapá, the Federal District and Tocantins. The fifth group contains only Paraná and Roraima, while Rio Grande do Norte is the only state in the sixth group.

4.2 COVID-19 mortality per 100k inhabitants in Brazilian Federal Units

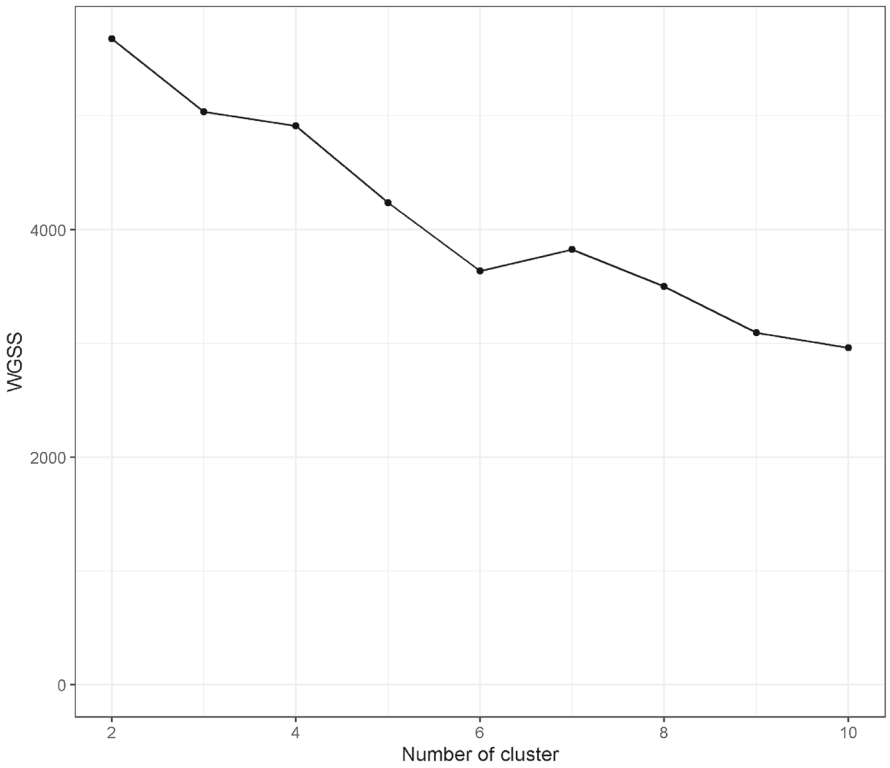
With respect to COVID-19 mortality, we also analyzed the temporal behavior of two indicators, both relativized to the size of each FU's population: New deaths per 100k inhabitants and total deaths per 100k inhabitants. In accordance with the curves illustrated in Fig. 13a, b, six groups were selected to carry out clustering analysis.

Figure 14 shows that, in the analysis of COVID-19 mortality, similarly to case prevalence, the adopted method was capable of efficiently separating the Brazilian FUs into homogeneous groups formed by curves with similar levels of COVID-19 mortality.

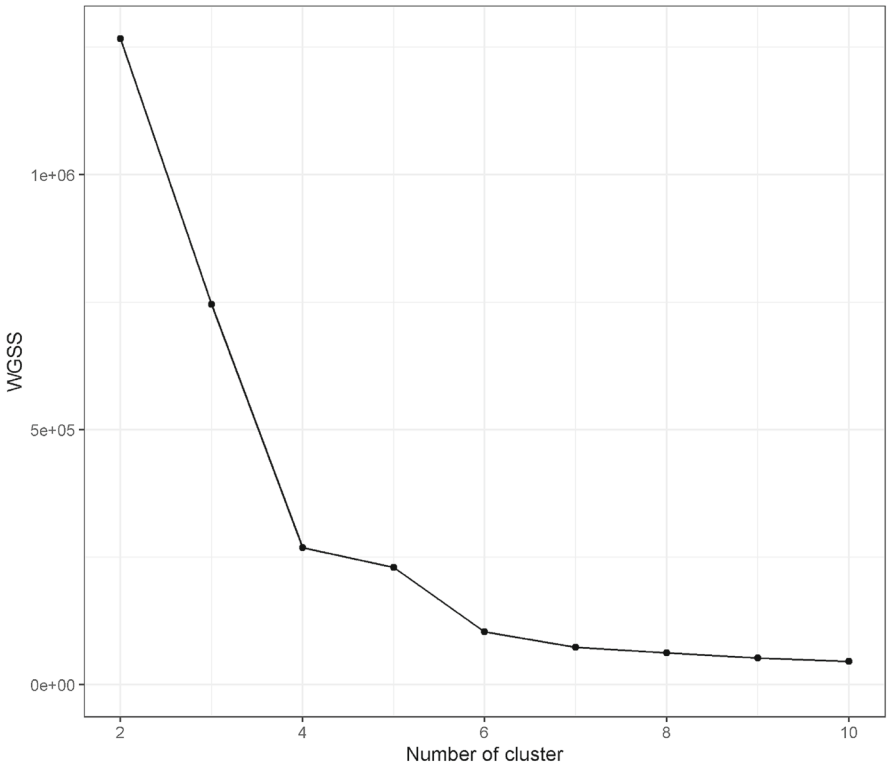
As seen in Fig. 15, the first group consists solely of the state of Maranhão, with the lowest level of COVID-19 mortality recorded during the analyzed period. The second group contains the states of Acre, Alagoas, Amapá, Bahia, Pará, Paraíba, Pernambuco, Piauí and Rio Grande do Norte. In the third group, the states of Ceará, Minas Gerais, Santa Catarina, Sergipe and Tocantins clustered together, while the fourth group contains Amazonas, Espírito Santo, Roraima, Rio Grande do Sul and São Paulo. In the fifth cluster, the Federal District, Goiás, Mato Grosso do Sul, Paraná and Rondônia were grouped together, while the sixth cluster consists of Mato Grosso and Rio de Janeiro.

Setting aside the perspective of accumulated deaths to examine mortality in terms of the evolution of new deaths recorded in each epidemiological week, as shown in Fig. 16, two waves of COVID-19 become evident in Brazil. The first occurs in mid-2020, while the second takes place around the end of the first quarter of 2021, with some peaks in FUs being evident.

The clustering map shown in Fig. 17 reveals that the first cluster is formed by the states of Alagoas, Bahia and Maranhão, while the second is composed of Amapá, Paraíba, Pernambuco, Piauí and Tocantins. The third contains Acre, Ceará, Pará, Rio Grande do Norte and Sergipe. The fourth



(a)



(b)

Fig. 8 WCSS of clustering simulations of incidence and prevalence curves for Brazilian FUs. **a** New cases per 100k inhabitants. **b** Total cases per 100k inhabitants

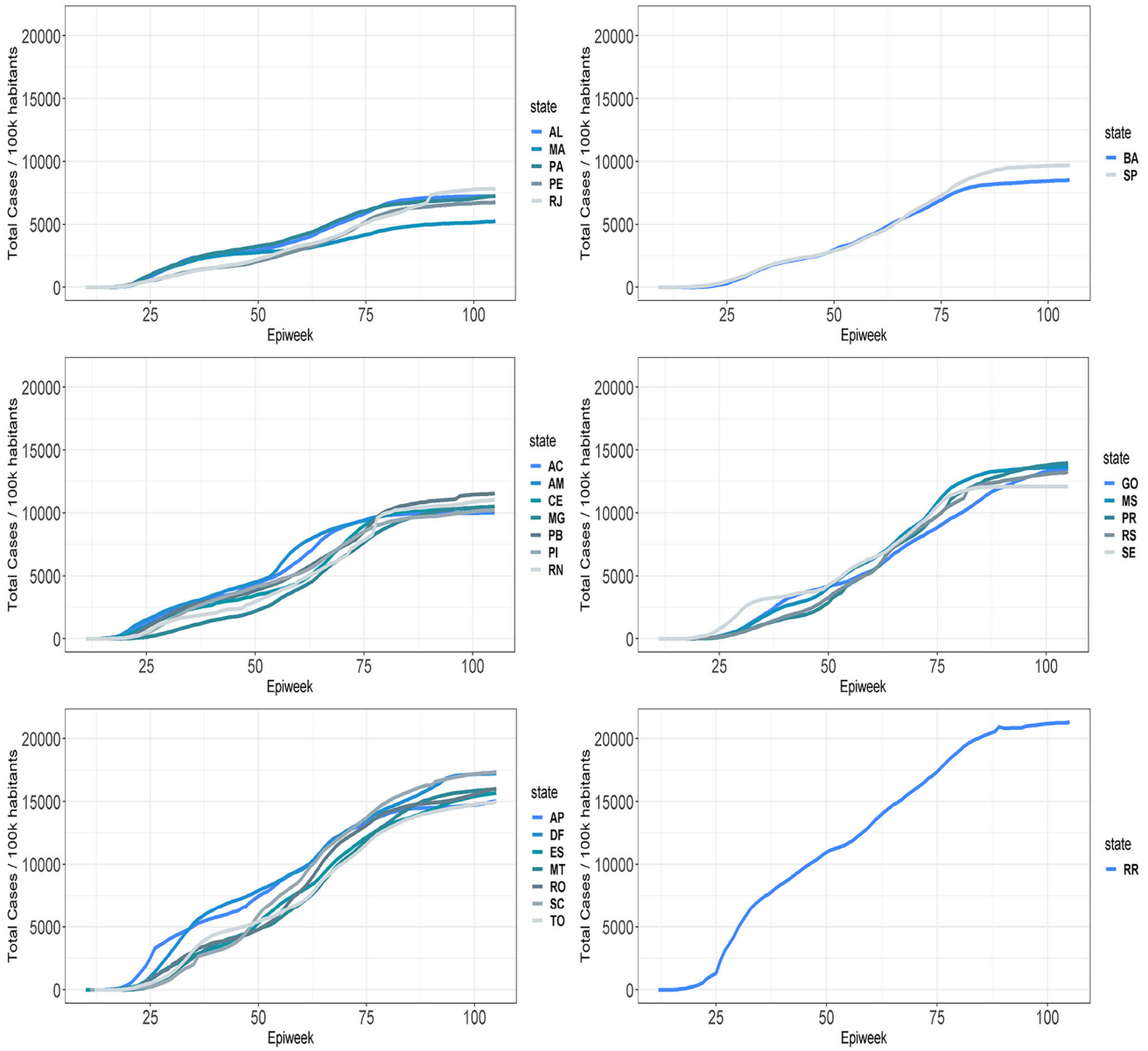


Fig. 9 Prevalence curves pertaining to total number of COVID-19 cases in each Brazilian FU, grouped by clustering analysis

cluster consists of the Federal District, Espírito Santo, Goiás, Minas Gerais, Mato Grosso, Mato Grosso do Sul, Paraná, Rio de Janeiro, Roraima, Santa Catarina and São Paulo, while Rio Grande do Sul and Rondônia make up the fifth cluster. The sixth contains only the state of Amazonas.

4.3 Discussion on real application

In general, when analyzing the present clustering results, the seven FUs that grouped into the first two clusters of COVID-19 prevalence curves, i.e., those with the lowest numbers of cases per 100K inhabitants, six also formed the first cluster of incidence curves. Of these six states, four are located in the Northeast, while Pará is in the North and Rio de Janeiro lies in southeastern Brazil. Of note are the pairs and trios of states that clustered together both in terms of case prevalence and incidence: Minas Gerais and Acre, Pernambuco and Rio de Janeiro, the Federal District and Tocantins; regarding trios, the first was Espírito Santo, Santa Catarina and Rondônia, while the second was Alagoas, Pará and Maranhão.

An examination of the 10 FUs that formed the first and second clusters pertaining to mortality, seven were present in the first and second clusters formed by the cumulative mortality indicator. It is noteworthy that, with the exception of Amapá, all of the states in these two clusters are all located in northeastern Brazil. It is also possible to observe groups of two,

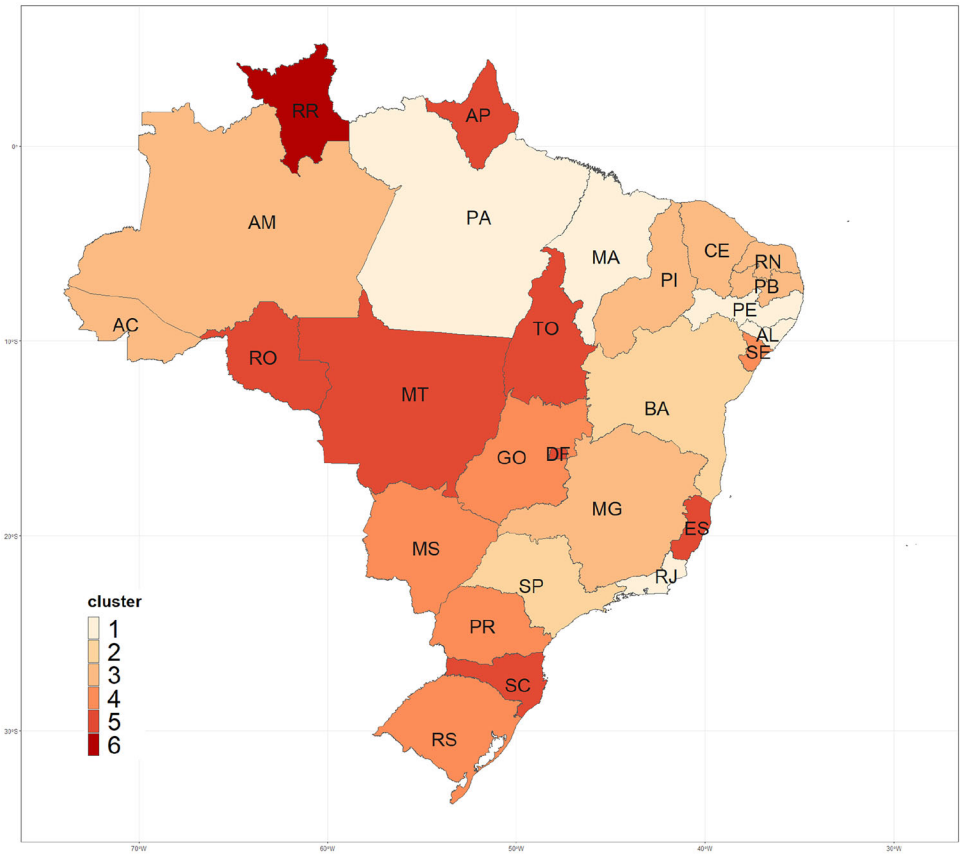


Fig. 10 Clustering of Brazilian FUs according to COVID-19 case prevalence

three and four FUs within the same groups of weekly deaths and cumulative mortality: Minas Gerais and Santa Catarina, Alagoas and Bahia; Espírito Santo, Roraima and São Paulo; and finally, Amapá, Paraíba, Pernambuco and Piauí.

Also of note is the group of nine Brazilian FUs (all in the Northeast, except Pará, in addition to Rio de Janeiro, Minas Gerais and São Paulo in the Southeast/South) that grouped together in lower intensity clusters of both cumulative cases and deaths per 100K inhabitants. The duo of Rondônia and Mato Grosso was also in another cluster of case prevalence, as well as in the last two clusters with the highest numbers of cumulative deaths per 100K inhabitants. It is also worth mentioning that the state of Rio de Janeiro, despite being present in the cluster with the lowest numbers of cases per 100K inhabitants, this federation unit also grouped into the cluster corresponding to the highest cumulative mortality recorded.

5 Final comments

The clustering method applied herein was found to reliably group Brazilian federal units considering the criteria established for analysis. In the context of simulation studies, the scenarios in which the simulated series exhibited within-group homogeneity and heterogeneity between groups, under equal or different mean distances, unanimous classification and distinction between all clusters were observed.

It is worth noting that initially a hierarchical strategy was used for clustering time series, employing DTW distance. This approach is functional for small samples, such as the 27 time series, one for each Brazilian federative unit. However, if we consider other geographic segments such as health regions, municipalities, census tracts, etc., the number of time series would be considerably higher, and the processing would be computationally costly. While the time complexity of a hierarchical clustering algorithm is generally quadratic, the k-means algorithm has linear time complexity. Therefore, for time-series clustering, k-means with DTW distance and DBA centroid becomes a more efficient clustering strategy than the hierarchical one.

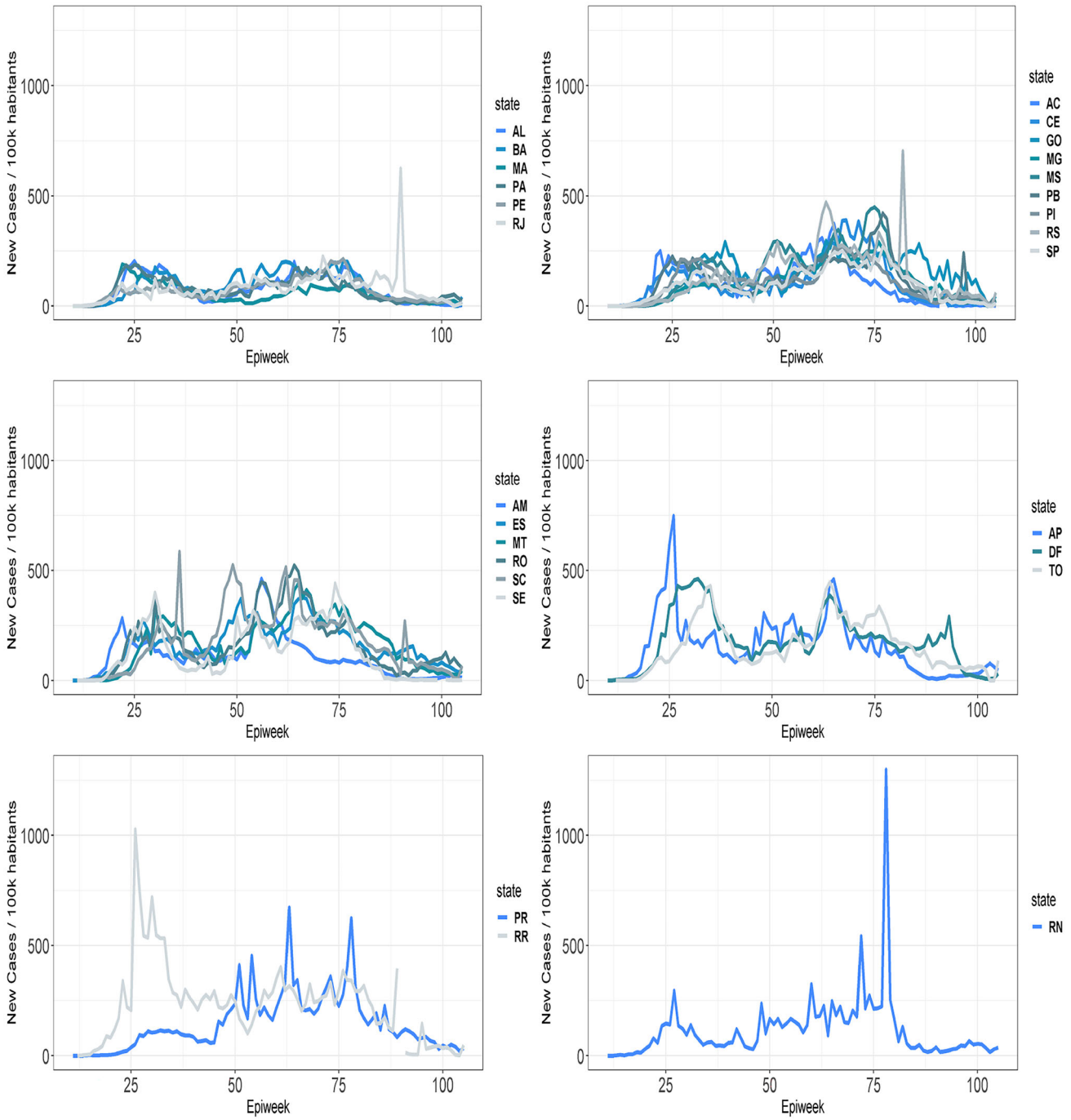


Fig. 11 Incidence curves corresponding to numbers of new weekly COVID-19 cases per 100K inhabitants in each Brazilian FU, grouped by clustering analysis

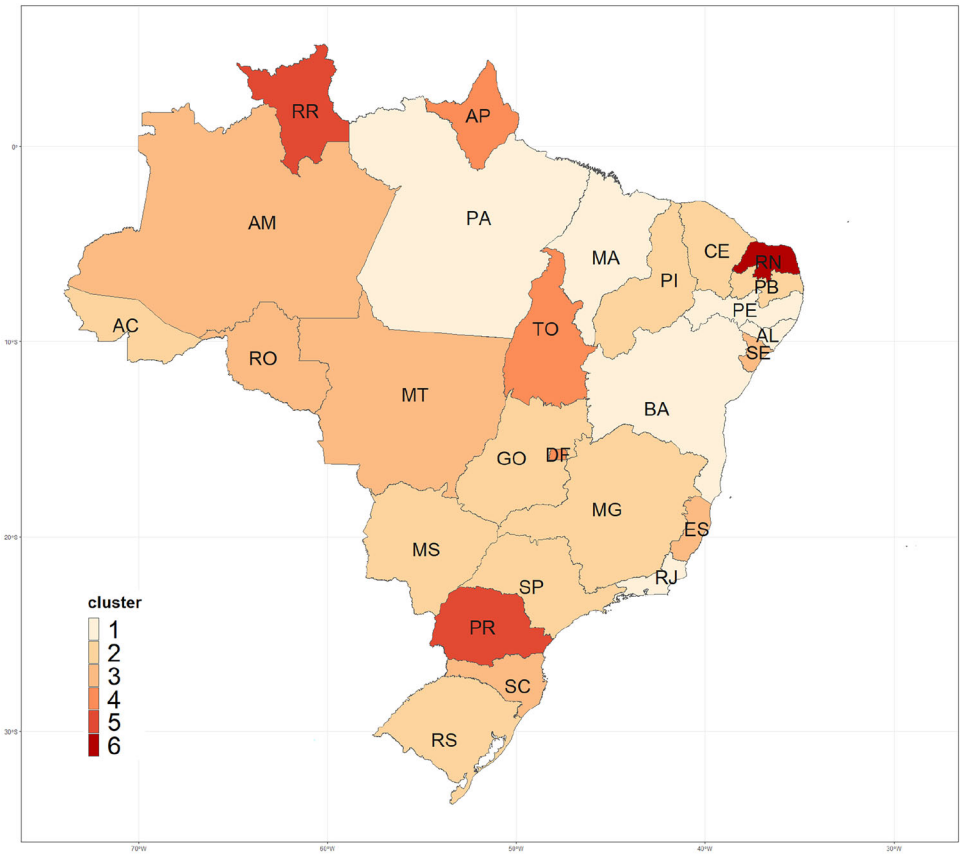
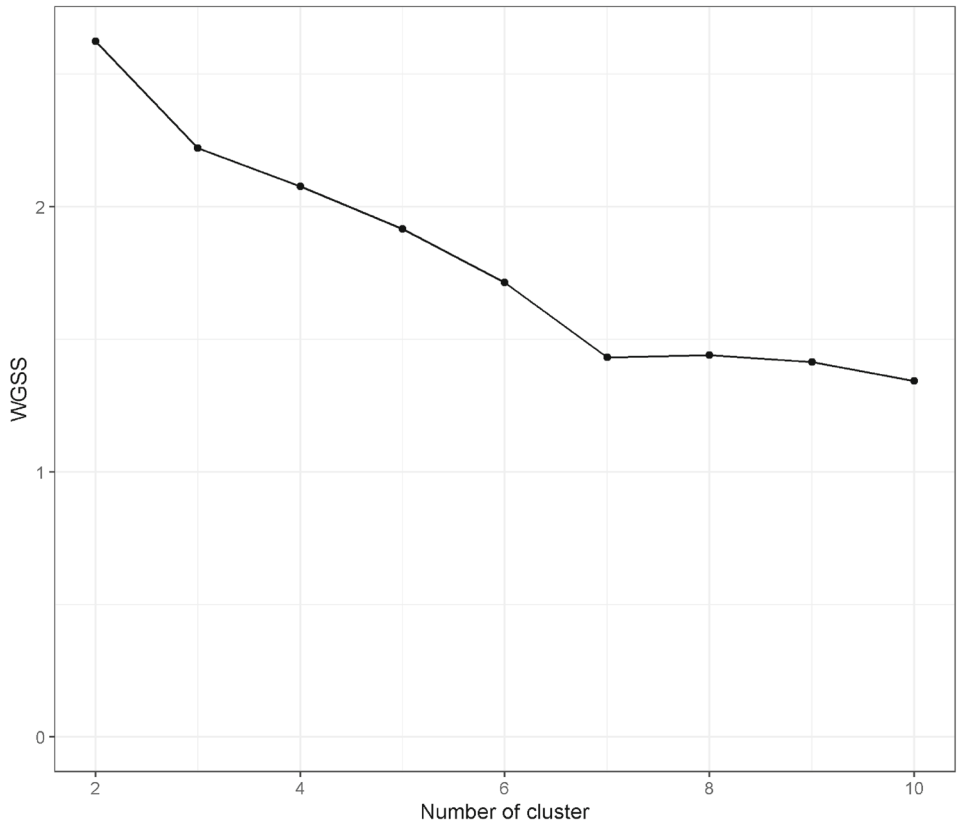
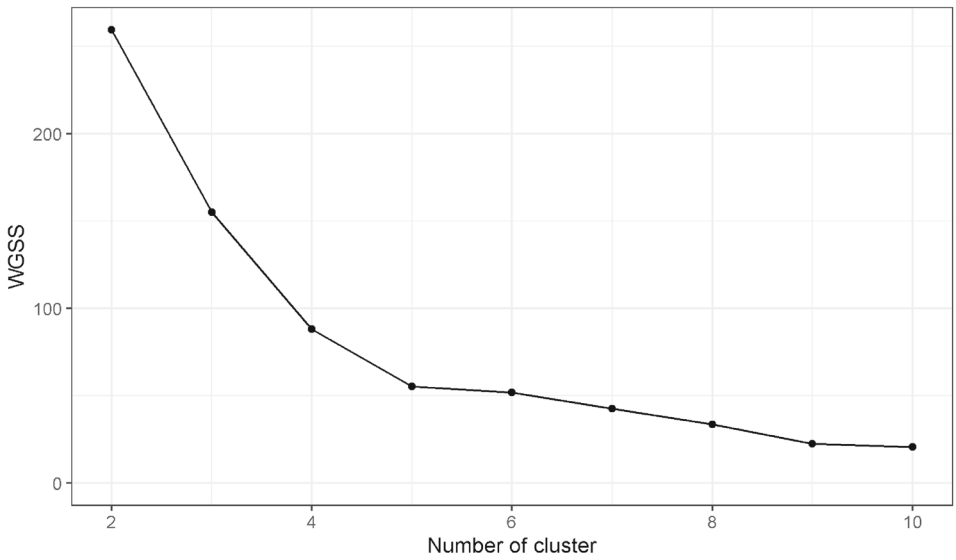


Fig. 12 Clustering of Brazilian FUs according to COVID-19 case incidence



(a)



(b)

Fig. 13 WCSS of clustering simulations of new and total deaths per 100k inhabitants in all Brazilian Federal Units. **a** New deaths per 100k inhabitants. **b** Total deaths per 100k inhabitants

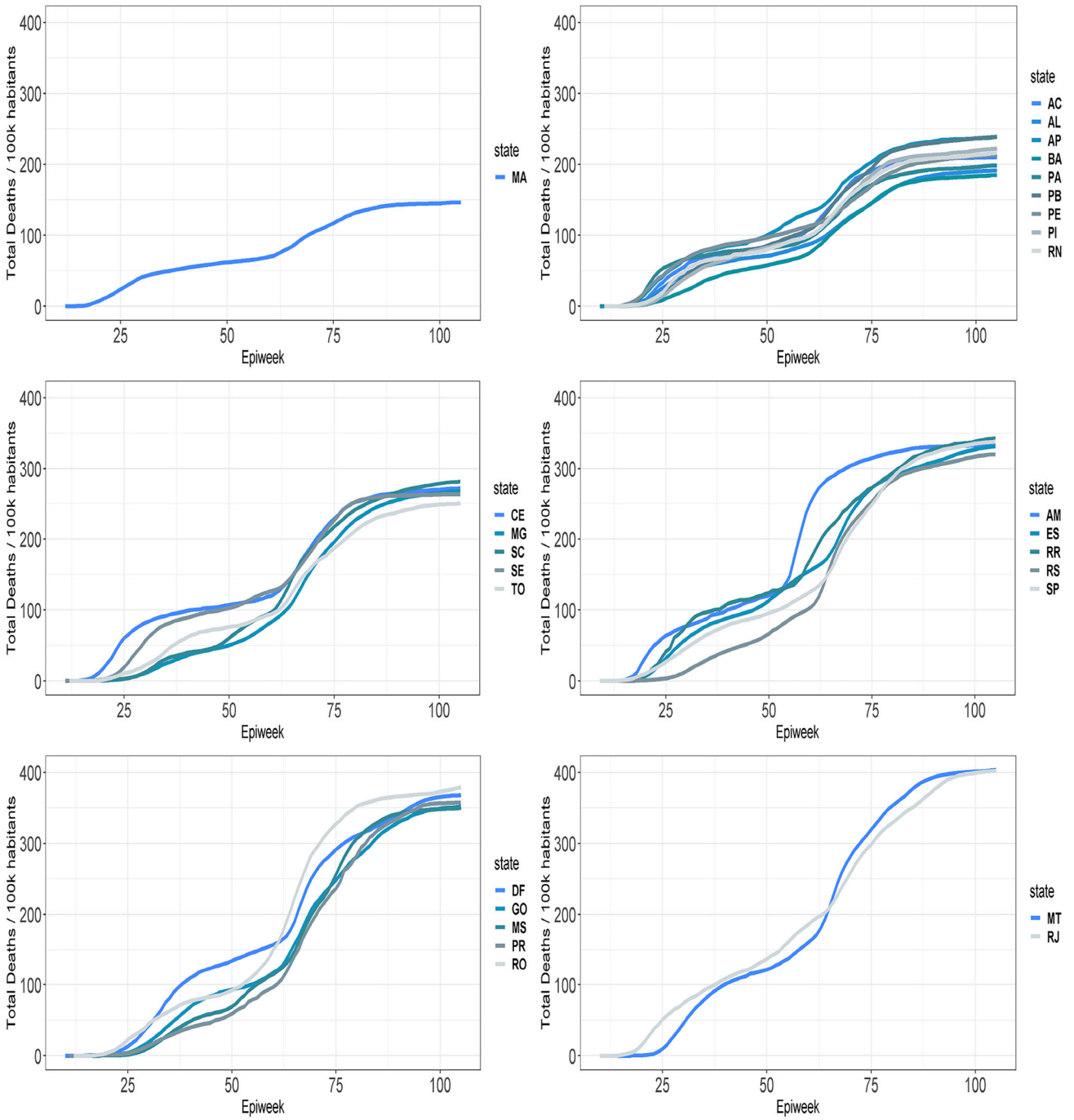


Fig. 14 Curves corresponding to numbers of cumulative COVID-19 deaths per 100K inhabitants in each Brazilian FU, grouped by clustering analysis

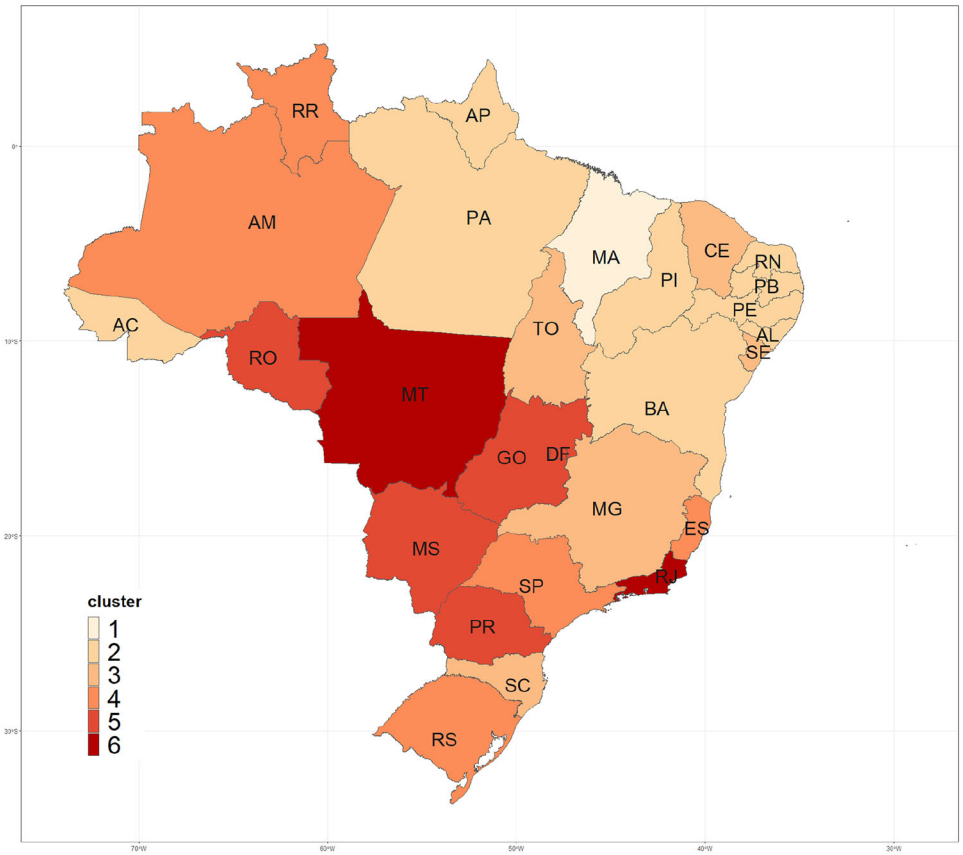


Fig. 15 Clustering of Brazilian FUs according to cumulative number of COVID-19 deaths

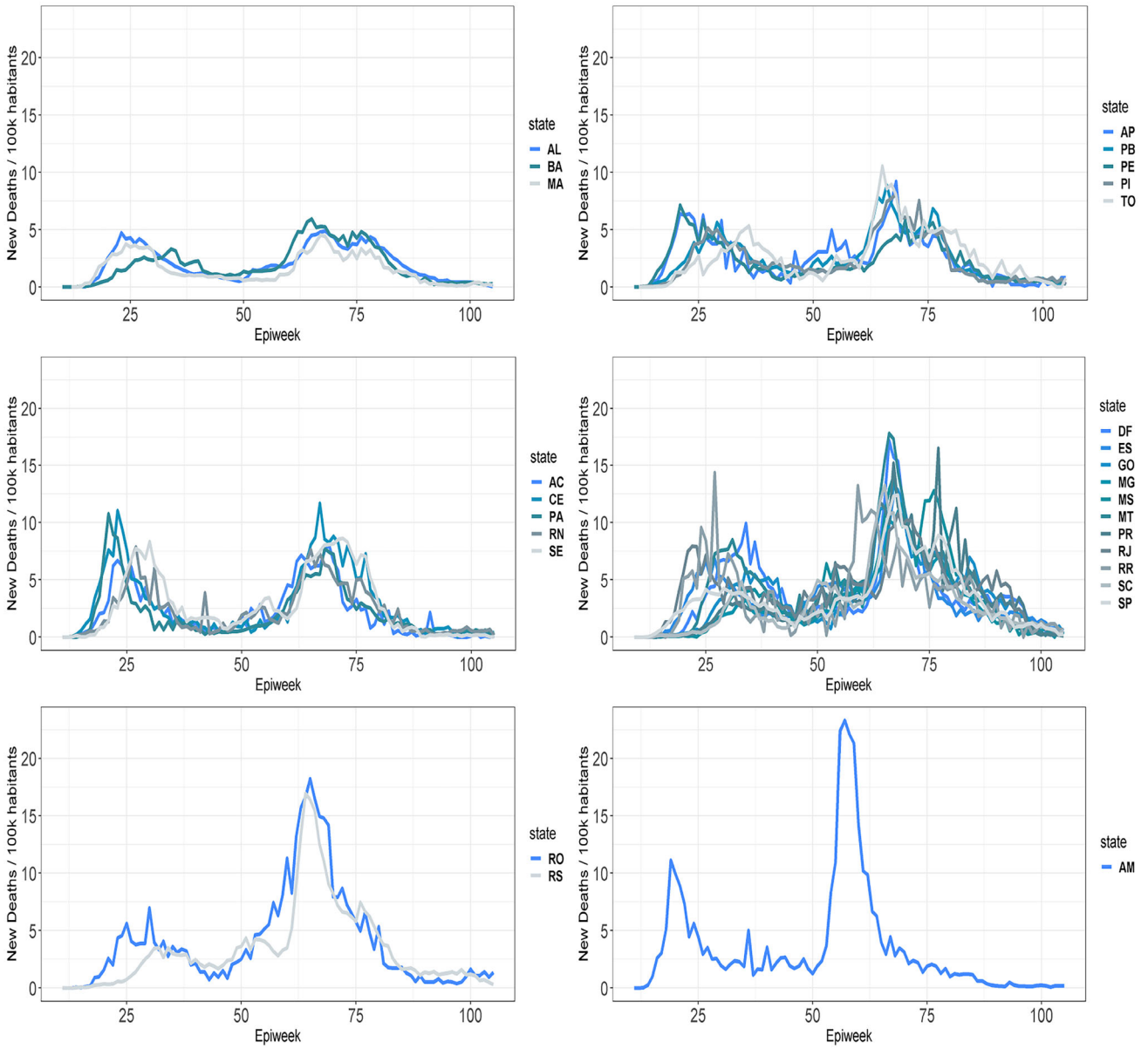


Fig. 16 Curves corresponding to numbers of new weekly COVID-19 deaths per 100K inhabitants in each Brazilian FU, grouped by clustering analysis

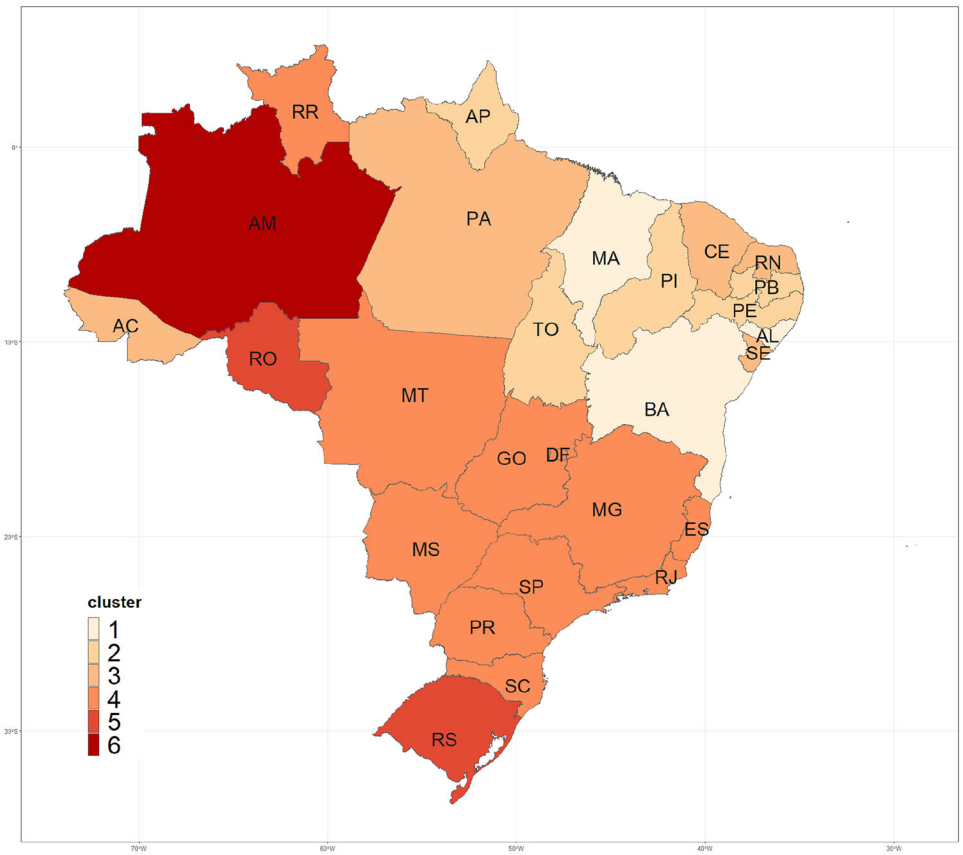


Fig. 17 Clustering of Brazilian FUs according to number of new weekly COVID-19 deaths

With regard to the strategies to group curves corresponding to COVID-19 cases and deaths among the Brazilian federation units, the combination of k-means grouping employing DTW and DBA proved effective for characterizing both cases and mortality. The method was found to be efficient in both COVID-19 case and mortality scenarios, capably incorporating similarities among the curves to accurately form homogeneous groups. Future perspectives include expanding this method's applicability to cover the country's network of municipalities, which presents a significant challenge due to the need for additional computational resources, considering that Brazil is divided into over 5,500 municipalities. The use of additional data sources constitutes an additional challenge that would enable the use of other markers of COVID-19 cases and deaths, such as the date of symptom onset or date of positive test result.

Acknowledgements This article is the result of part of the project entitled "Avaliação dos efeitos das desigualdades sociais na pandemia COVID-19 em país de baixa e média renda," (Assessment of the effects of social inequalities during the COVID-19 pandemic in low- and middle-income countries) carried out by the Center for Integration of Data and Knowledge for Health, linked to the Gonçalo Muniz Institute, Oswaldo Cruz Foundation (CIDACS/IGM-FIOCRUZ). This project was funded by Health Data Research United Kingdom (HDR-UK) and received support from the Bill and Melinda Gates Foundation.

Author Contributions All authors played essential roles throughout the development of this work, actively contributing from the initial planning phase of the study to the completion of the article's writing. Jonatas Silva do Espirito Santo and Jackson Santos da Conceição were responsible for creating the figures, in collaboration with Lilia Carolina Carneiro da Costa, Rosemeire Leovigildo Fiaccone, and Anderson Ara, who drafted the manuscript. The critical review of the article and final approval were conducted by Maria Yury Ichihara, Marcos Ennes Barreto, and all other authors.

Funding This project was funded by Health Data Research United Kingdom (HDR-UK) and received support from the Bill and Melinda Gates Foundation.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Brasil.: Brasil Confirma Primeiro Caso do Novo Coronavírus. https://www.gov.br/pt-br/noticias/saude-e-vigilancia-sanitaria/2020/02/brasil-confirma-primeiro-caso-do-novo-coronavirus. Accessed 30 Jan 2024

2. Moura, E.C., Cortez-Escalante, J., Cavalcante, F.V., Barreto, I.C.D.H.C., Sanchez, M.N., Santos, L.M.P.: Covid-19: temporal evolution and immunization in the three epidemiological waves, Brazil, 2020–2022. Revista de Saúde Pública **56**, 105 (2022)
3. Ichihara, M.Y., C Costa, L.C., Fiaccone, R.L., de Medeiros, A.G., Bellido, J., Souza, R.F.D.S., Rocha, C., Anjos, A.F.D., Sebastião, M., Pimenta, D., et al.: Measuring social inequalities in health in the covid-19 pandemic in a middle-income country: the ids-covid-19 index (2023)
4. Zeiser, F.A., Donida, B., da Costa, C.A., de Oliveira Ramos, G., Scherer, J.N., Barcellos, N.T., Alegretti, A.P., Ikeda, M.L.R., Müller, A.P.W.C., Bohn, H.C., et al.: First and second covid-19 waves in Brazil: a cross-sectional study of patients' characteristics related to hospitalization and in-hospital mortality. Lancet Region. Health Am. **6**, 8 (2022)
5. Cota, W.: Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level. SciELOPreprints (2020). https://doi.org/10.1590/scielopreprints.362
6. Oded Maimon, L.R.: Data Mining and Knowledge Discovery Handbook, 1st edn. Springer, Berlin (2005)
7. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Data Management Systems Series, Morgan Kaufmann Publishers (2001)
8. Lahreche, A., Boucheham, B.: A fast and accurate similarity measure for long time series classification based on local extrema and dynamic time warping. Expert Syst. Appl. **168**, 114374 (2021). https://doi.org/10.1016/j.eswa.2020.114374
9. Landmesser, J.: The use of the dynamic time warping (dtw) method to describe the covid-19 dynamics in Poland. Oeconomia Copernicana **12**(3), 539–556 (2021). https://doi.org/10.24136/oc.2021.018
10. Jeong, Y.-S., Jeong, M.K., Omitaomu, O.A.: Weighted dynamic time warping for time series classification. Pattern Recogn. (2011). https://doi.org/10.1016/j.patcog.2010.09.022
11. Sakoe, H., Chiba, S.: A dynamic programming approach to continuous speech recognition. Proc. Seventh Int. Congress Acoust. **3**, 65–69 (1971)
12. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. **26**(1), 43–49 (1978). https://doi.org/10.1109/TASSP.1978.1163055
13. Itakura, F.: Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoust. Speech Signal Process. **23**(1), 67–72 (1975). https://doi.org/10.1109/TASSP.1975.1162641
14. Ratanamahatana, C.A., Keogh, E.: Making time-series classification more accurate using learned constraints (2004)
15. Niennattrakul, V., Ratanamahatana, C.A.: Inaccuracies of shape averaging method using dynamic time warping for time series data. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) Computational Science—ICCS 2007, pp. 513–520. Springer, Berlin (2007). https://doi.org/10.1007/978-3-540-72584-8_68
16. Niennattrakul, V., Ratanamahatana, C.A.: On clustering multimedia time series data using k-means and dynamic time warping. In: 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07), pp. 733–738 (2007). https://doi.org/10.1109/MUE.2007.165
17. Petitjean, F., Ketterlin, A., Gançarski, P.: A global averaging method for dynamic time warping, with applications to clustering. Pattern Recogn. **44**(3), 678–693 (2011). https://doi.org/10.1016/j.patcog.2010.09.013
18. Petitjean, F., Forestier, G., Webb, G.I., Nicholson, A.E., Chen, Y., Keogh, E.: Dynamic time warping averaging of time series allows faster and more accurate classification. In: 2014 IEEE International Conference on Data Mining, pp. 470–479 (2014). https://doi.org/10.1109/ICDM.2014.27
19. Jang, M., Han, M.-S., Kim, J.-H., Yang, H.-S.: In: Mehrotra, K.G., Mohan, C., Oh, J.C., Varshney, P.K., Ali, M. (eds.) Dynamic Time Warping-Based K-Means Clustering for Accelerometer-Based Handwriting Recognition, pp. 21–26. Springer, Berlin (2011). https://doi.org/10.1007/978-3-642-21332-8_3
20. Anh, D.T., Thanh, L.H.: An efficient implementation of k-means clustering for time series data with dtw distance. Int. J. Bus. Intell. Data Min. **10**(3), 213–232 (2015). https://doi.org/10.1504/IJBIDM.2015.071311
21. Forestier, G., Petitjean, F., Dau, H.A., Webb, G.I., Keogh, E.: Generating synthetic time series to augment sparse datasets. In: 2017 IEEE International Conference on Data Mining (ICDM), pp. 865–870 (2017). https://doi.org/10.1109/ICDM.2017.106
22. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series (2017). arXiv preprint arXiv:1703.01541. https://doi.org/10.48550/arXiv.1703.01541
23. Leodolter, M., Plant, C., Brändle, N.: Incdtw: an r package for incremental calculation of dynamic time warping. J. Stat. Softw. Art. **99**(9), 1–23 (2021). https://doi.org/10.18637/jss.v099.i09
24. Javed, A., Rizzo, D.M., Lee, B.S., Gramling, R.: Somtimes: self organizing maps for time series clustering and its application to serious illness conversations. CoRR (2021) arXiv: 2108.11523. https://doi.org/10.48550/arXiv.2108.11523
25. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. **3**(1), 1–27 (1974). https://doi.org/10.1080/03610927408827101
26. da Silva, P.L.P.: Um estudo sobre o agrupamento de séries temporais e sua aplicação em curvas de carga residenciais. Master's thesis, Universidade Federal de Minas Gerais (2016). https://repositorio.ufmg.br/bitstream/1843/BUOS-APWMJD/1/versao_final_dissertacao_impressao_capa_dura_pedro_pazzini.pdf
27. R Core Team.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing (2021). https://www.R-project.org/
28. Giorgino, T.: Computing and visualizing dynamic time warping alignments in R: the dtw package. J. Stat. Softw. **31**(7), 1–24 (2009). https://doi.org/10.18637/jss.v031.i07
29. Sardá-Espinosa, A.: Time-series clustering in r using the dtwclust package. Roy J. (2019). https://doi.org/10.32614/RJ-2019-023
30. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H.: Welcome to the tidyverse. J. Open Source Softw. **4**(43), 1686 (2019). https://doi.org/10.21105/joss.01686
31. Box, G., Jenkins, G.M.: Time Series Analysis: Forecasting and Control. Holden-Day (1970)