Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: http://ees.elsevier.com

A multivariate analysis on spatiotemporal evolution of Covid-19 in Brazil

Marcio Luis Ferreira Nascimento Ph.D. [a,b,*]

[a] *Nano Group @ UFBA, Department of Chemical Engineering, Polytechnic School, Federal University of Bahia, Rua Aristides Novis 2, Federação, 40210 - 630, Salvador, BA, Brazil*

[b] *PEI - Graduate Program in Industrial Engineering, Department of Chemical Engineering, Polytechnic School, Federal University of Bahia, Rua Aristides Novis 2, Federação, 40210 - 630, Salvador, BA, Brazil*

ARTICLE INFO

Article history:
Received 18 June 2020
Received in revised form 25 August 2020
Accepted 29 August 2020
Available online xxx
Handling editor: Dr. J Wu

Keywords:
Pandemic
COVID-19
Coronavirus
K-means clustering
Factor analysis
Spatiotemporal analysis

ABSTRACT

This data-driven work aims to analyze and classify the spatiotemporal distribution of all Brazilian states considering data so diverse as the number of Covid-19 cases, deaths, confirmed cases per 100 k inhabitants, mortality per 100 k inhabitants and case fatality rates as health indicators. We also considered population, area and population density as geographic indicators. Finally, GDP and HDI were taken into account as economic and social criteria. For this task data were collected from April 3rd until August 8th, 2020, corresponding to epidemiological weeks 14–32, reaching three million cases and a hundred thousand deaths. With this data it was possible to classify Brazilian states using multivariate methods into possible groups by means of non-hierarchical (*k*-means) cluster as well as factor analysis. It was possible to group all states plus the Federal District into five clusters, taking into account these 10 variables over the first five months of the epidemic. Group changes between states were observed over time and clusters, and between three and four factors were found. However, even with great difference on health indicators during days, the number of clusters remains fixed. Also, São Paulo and Rio de Janeiro states were ranked at top list taking into account all epidemiological weeks. Correlations were observed between variables, such as the number of Covid cases and deaths with GDP for most of epidemiological weeks. Some clusters were more critical due to specific variables, including cities that are main hotspots. These multivariate findings would provide a comprehensive description of the ongoing Covid-19 epidemic and may help to guide subsequent studies to understand and control virus transmission.

1. Introduction

Covid-19 is a severe acute respiratory infection that turned pandemic and emerged as one of major global health, economic, geographic and social challenges in this century (Anonymous, 2020). The outbreak was declared a public health emergency by the World Health Organization (WHO) on January 30th, 2020, and contributions to data analysis for new policymaker strategies in order to support epidemiological decisions were done since then (Cipitelli et al., 2020; Wu et al., 2020).

In Brazil, the first Covid-19 case was confirmed on February 26th, 2020, by a traveler returning to São Paulo from Northern Italy, and on March 17 the first death from the disease was reported in epidemiological week (EW) 12 (Oliveira et al., 2020). São Paulo is the largest city in South America and the main air hub. Its Metropolitan Region has a population of 23 million inhabitants (Souza et al., 2020). The large-scale outbreak of Covid-19 in Brazil presents an important opportunity to carry out studies which may be relevant for understanding patterns of transmission relevant for the ensuing pandemic. On May 23rd, 2020 (EW 21), Brazil already ranked 3rd worldwide in terms of the number of confirmed cases (347,398) and deaths (22,013), and these are probably substantial

Peer review under responsibility of KeAi Communications Co., Ltd.

* Institute of Humanities, Arts and Sciences, Federal University of Bahia, Rua Barão de Jeremoabo s/n, Classroom Pavilion V, Ondina University Campus,40170-115 Salvador, BA, Brazil

E-mail address: mlfn@ufba.br (M.L.F. Nascimento)

https://doi.org/10.1016/j.idm.2020.08.012
2468-0427/

underestimates (SUS, 2020). Few weeks later, Brazil ranked 2nd worldwide in terms of Covid-19 infections (828,810) and deaths (41,828) (EW 24).

Most State governors have imposed quarantines and even a few lockdowns to prevent the spread of the virus across all five administrative regions of Brazil; however, there has been little leadership from Federal Government (Anonymous, 2020), that presented a *laissez faire* reaction[1] to Covid-19 (Conde, 2020). The most commonly reported clinical symptoms were dry cough, fever, dyspnoea, fatigue, ageusia, anosmia, or some combinations of these (Souza et al., 2020). There is a great risk of the virus reaching smaller cities with inadequate provision of intensive care beds and mechanical ventilators, no medications and testing kits, little physical distancing, hygiene recommendations and no personal protective equipment (PPE) (Anonymous, 2020).

Also, recent publications have drawn attention to the possible benefit of chloroquine diphosphate and hydroxychloroquine, both supported by the Brazilian regulatory agency and the Brazilian Ministry of Health in a compassionate manner given the severity of the disease. However, high dosages were associated with lethality (Borba et al., 2020).

Due to insufficient scientific knowledge, the fast pace of Covid-19 spread, and its capacity to cause deaths in vulnerable countries have generated uncertainties on the best strategies for confronting the pandemic (Werneck & Carvalho, 2020). This paper aims to consider a large amount of data to find similarities by means of non-hierarchical *k*-means clustering and factor analysis between 26 Brazilian states and one federal district with ten public parameters: the number of confirmed Covid-19 cases, the number of deaths, confirmed cases per 100 k inhabitants, mortality per 100 k inhabitants, case fatality rates (*i.e.*, the proportion of deaths compared to the total number of confirmed Covid-19 cases), population, area, the population density, Gross Domestic Product (GDP) and Human Development Index (HDI). These data, collected from April 3rd to August 8th, 2020, came from State Health Secretariats and the Brazilian Ministry of Health (SUS, *Sistema Único de Saúde* or Unified Health System - see Coronavirus Brasil, 2020), as well as the Brazilian Institute of Geography and Statistics (IBGE) and the Institute of Applied Economic Research (IPEA).

Briefly speaking, clustering is a computational technique that groups a given set of data points into a fixed number of clusters in such a way that points within the cluster are similar and those from two different clusters are dissimilar. *K*-means is a particular mathematical technique of processing data for further classification, as proposed by MacQueen (1967). Factor analysis (FA) is a singular case of transforming original data into a new coordinate system with fewer variables and in order of their importance in terms of the variation in the data (Spearman, 1904).

Clustering has applications in many fields as military aircraft (Santos et al., 2019), medicine (Ahlqvist et al., 2018) or genealogy (Kaplanis, 2018). Currently, FA is used in several fields of knowledge, such as economics, finance, marketing, actuarial science, accounting, logistics, strategy, medicine, psychology and biostatistics, among others (Favero & Belfiore, 2019, p. 383). The main goal of this paper is to present a multivariate spatiotemporal evolution applied to Covid-19 in Brazil. Using the data available from April 3rd to August 8th, 2020, on the ten parameters mentioned above, 26 Brazilian states plus the federal district were classified: 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).

The paper is organized in the following way. In Section 2 is presented a brief theory on clustering with a simple example, data parametrization and the fundamentals of factor analysis. Section 3 describes briefly the methodology. The numerical results with discussion and conclusions are presented in Sections 4 and 5 respectively.

2. Brief theory

In this section both non-hierarchical cluster and factor analysis techniques are briefly described.

2.1. Non-hierarchical cluster

Non-hierarchical clustering is an exploratory method in multivariate analysis that aims to identify groups that have similar characteristics (Hair et al., 2019, p. 189). The *k*-means is one of the most popular methods of data mining and cluster analysis because it reduces possible human subjectivity. It is an unsupervised clustering method that partitions data into a specific number of *k*-clusters for compressing or summarizing original values. Its precise mathematical algorithm determines that each group is nucleated by a *medoid* $\bar{m}_k$, the central point of the cluster (Everitt et al., 2011, p. 86; Kopec, 2019). According to MacQueen (1967), the *k*-means concept represents a generalization of the ordinary sample mean, where the medoid can be viewed as a special case of a mean, or average, of a cluster. The partitioning criterion is a distance measure, as exemplified below.

Conceptually, *k*-means is actually quite simple: in each iteration, every datum is associated with the cluster that it is nearest to in terms of the cluster's center. That center changes as new data are associated with the cluster until a convergence occurs. In simple terms, let us partition the numbers between 1 and 9 into two clusters A and B ($k = 2$). The basic procedure would be to sort two numbers from this data (called seeds or medoids) and to calculate the numerical distances between these seeds from original data. Taking for example **1** and **9** as medoids, their respective distances are:

[1] "So what? I'm sorry. What do you want me to do about it?" These were the words of Brazil's president, a right-wing populist leader that has constantly belittled the epidemic, on Tuesday 28 April, 2020 (EW 18), in agreement with Conde (2020).

- For $\bar{m}_1 = 1$ as initial medoid: $\sqrt{(1-1)^2} = 0$; $\sqrt{(2-1)^2} = 1$; $\sqrt{(3-1)^2} = 2$; $\sqrt{(4-1)^2} = 3$; $\sqrt{(5-1)^2} = 4$; $\sqrt{(6-1)^2} = 5$; $\sqrt{(7-1)^2} = 6$; $\sqrt{(8-1)^2} = 7$; $\sqrt{(9-1)^2} = 8$;
- For $\bar{m}_2 = 9$ as initial medoid: $\sqrt{(1-9)^2} = 8$; $\sqrt{(2-9)^2} = 7$; $\sqrt{(3-9)^2} = 6$; $\sqrt{(4-9)^2} = 5$; $\sqrt{(5-9)^2} = 4$; $\sqrt{(6-9)^2} = 3$; $\sqrt{(7-9)^2} = 2$; $\sqrt{(8-9)^2} = 1$; $\sqrt{(9-9)^2} = 0$;

The minimum value between the respective distances would define the first clusters, named as *A* and *B*: 1 to 5 (*A*) and 6 to 9 (*B*). The next medoids would be a simple average between 1 and 5 (that is, $\bar{m}_1 = 3$) and between 6 and 9 (that is, $\bar{m}_2 = 7.5$). Calculating new distances from the original data considering these new medoids would give the same clusters *A* and *B*, giving thus the result (the system converged).

It is important to note that each dimension of original data X_i needs to be comparable in magnitude. This process of making different data types comparable is known as *standardization* named *Z*-score (Everitt et al., 2011, p. 67; Hair et al., 2019, p. 101). Therefore, as every datum must be equivalent to other matrix raw data **X**, with *n* variables (or columns) and *p* labels (or rows), for each observation *i*, the value of a new standardized variable Z_i is obtained by subtracting the corresponding original variable value X_i from its mean $\bar{Z}$ and, after that, the resulting value is divided by its standard deviation σ_Z (Everitt et al., 2011, p. 67; Favero & Belfiore, 2019, p. 153).

Dissimilarities are found by means of calculated distances so that all *n* variables are equally important in determining these distances. This can be done by coding the variables so that the average $\bar{Z}$ values are all zero and the respective variances σ_Z^2 are all one. This standardization has the effect of minimizing cluster differences, because if groups are separated well by the original variable X_i, then the original variance σ_X^2 should be large (Everitt et al., 2011, p. 67).

This procedure simplifies and reduces the multidimensional dataset, promoting a reasonable measure of Euclidean distances $d_{i,j}$ from different Z_i elements (Everitt et al., 2011, p. 71; Hair et al., 2019, p. 208).

2.2. Factor analysis

In a historical context, FA development is partly due to Spearman's (1904) pioneer work over an analysis of the human mind. FA does not work if the original variables are uncorrelated. Best results are achieved when the original variables are very highly correlated.

As one of the simplest of the multivariate techniques (Hair et al., 2011, p. 121), it analyses a group of data described by many variables and extracts the most important information by means of new orthogonal *F* components. As this procedure reduces the number of original variables, the first factor is the direction throughout the data that explains the highest variability.

The Kaiser (1960) criterion was used for choosing the number of factors, in which only the factors that correspond to eigenvalues greater than 1 were considered. After finding corresponding *F* factors, the rotated factors *F′* were obtained using the Varimax method proposed by Kaiser (1958), which maximizes the loadings of each variable in a certain factor *F*. From this result it was possible to establish a ranking by generating performance indexes from the *F′* factors.

3. Methodology

Non-hierarchical *k*-means is a type of clustering algorithm that groups data points into certain predefined clusters, based on each point's relative distance to their medoids. In every round of *k*-means, the distance between every data point and every medoid is calculated through Euclidean distance as briefly described below.

The key issue with the *k*-means algorithm is selecting how to assign the initial conditions (the medoids). In the most basic form of most algorithms, the initial medoids are chosen randomly within the range of the data until convergence occurs. Another difficulty is deciding how many clusters to divide data (the "*k*" in *k*-means). In the classical algorithm, this number is determined by the user's experience, but the right number of clusters is unknown. The Elbow (or Scree) Data Chart would help with such a decision.

After this, PyCharm™, software based on Python language, developed by the Czech company JetBrains™, was used to elaborate clustering. PyCharm™ is an integrated development environment used in computer programming with the following steps (Kopec, 2019):

i) Initialize all data with "*k*" empty clusters;
ii) Standardize all data;
iii) Randomly choose *k* medoids associated with each cluster;
iv) Allocate each datum from the input set to the closest medoid based on the distance measurement;
v) Recalculate each medoid so its cluster center (or average) is associated with;
vi) Repeat fourth and fifth steps until a maximum number of iterations is reached or the medoids change no longer.

The Python code used in this work was applied by Santos et al. (2019) on military aircraft.

Factor analysis was done by means of the IBM SPSS Statistics 25 considering original variables. In order to verify the overall FA adequacy, the Kaiser-Meyer-Olkin statistic (KMO) and Bartlett's test of sphericity were performed (Favero & Belfiore, 2019, p. 387). The KMO statistic represents the proportion of variance considered common to all dataset variables under analysis. This statistic

varies from zero to one and, while values closer to one indicate that the variables share a very high proportion of variance (high correlations), values closer to zero may indicate that FA will not be adequate (Kaiser, 1970). The Bartlett's test consists in comparing a correlation matrix to an identity matrix of the same dimension. If there are no differences between them at a certain significance level, it is possible to consider that the FA will not be adequate (Bartlett, 1954).

4. Results and discussion

A geospatial analysis was conducted to understand the distribution of Covid-19 cases and other related variables in Brazil using multivariate methods. As examples of *k*-means applications, Ahlqvist et al. (2018) presented a refined classification on diabetes, extending from two main forms to five groups. They considered a data-driven cluster analysis based on six variables from 8980 patients of the Swedish All New Diabetes in Scania cohort (Ahlqvist et al., 2018). Another recent application of this procedure on Covid-19 was done by Zarikas et al. (2020), observing the clustering of 30 countries with respect to active cases, active cases per population and active cases per population and per area.

Since the aim of this work was to classify some public data over time, as presented in Supplementary Table S1*a-t*) presents data on the number of confirmed Covid-19 cases, the number of deaths, confirmed cases per 100 k inhabitants, mortality per 100 k inhabitants and case fatality rates. Supplementary Table S2 presents data on the population, area, the population density, GDP and HDI. These public data were from the Brazilian Ministry of Health (SUS - Coronavirus Brasil, 2020), IBGE and IPEA.

The *k*-means clustering algorithm was used to characterize best possible features to cluster similar Brazilian states with similar Covid-19 health indicators, population density, etc. together. As can be seen in Supplementary Table S1*a-t*), the first variables (X_1 to X_5) are related to health indicators. From Supplementary Table S2, X_6, X_7 and X_8 are geographic indicators, X_9 is economic (GDP) and X_{10} is social criteria.

The present analysis takes into account for every date, from April 3rd to August 8th, 2020, the first four variables X_1 to X_5 (Supplementary Table S1*a-t*) combined with X_6 to X_{10} (Supplementary Table S2). These dates correspond to epidemiological weeks 14–32. According to Supplementary Tables S1 and S2, the challenges that are looming are huge and made worse by the Brazilian social situation (Anonymous, 2020), which imposes precarious health and living conditions, especially for people who live on the poor outskirts of large urban centers (Oliveira et al., 2020).

Considering standardized data from raw data presented in Supplementary Tables S1*a-t*) and S2, Fig. 1*a-t*) present the decision on the number of clusters considered by means of the scree or elbow data chart. It is derived by plotting the total within sum against the number of factors in their order of extraction. The shape of the resulting curve is used to evaluate the cut-off point, that follows a tangent method (the intercept between two lines as indicated in Fig. 1*a-t*), in agreement with Hair et al. (2019, p. 813). The optimal number of k^*-clusters is indicated in each figure, related to the calculated total within sum (*WS*) and the tangent method. In the elbow method, one starts with $k = 2$ and increases k by 1 in each step and calculates the sum of Euclidean distances for each step. From theory (MacQueen, 1967), while increasing k, the sum of Euclidean distances will decrease dramatically in the beginning and will reach a plateau after increasing k further. The same result ($k^* = 5$) was found for all 20 dates, from April 3rd to August 8th, 2020. Therefore, we decided to use $k^* = 5$ and let the algorithm to group all Brazilian states in five clusters.

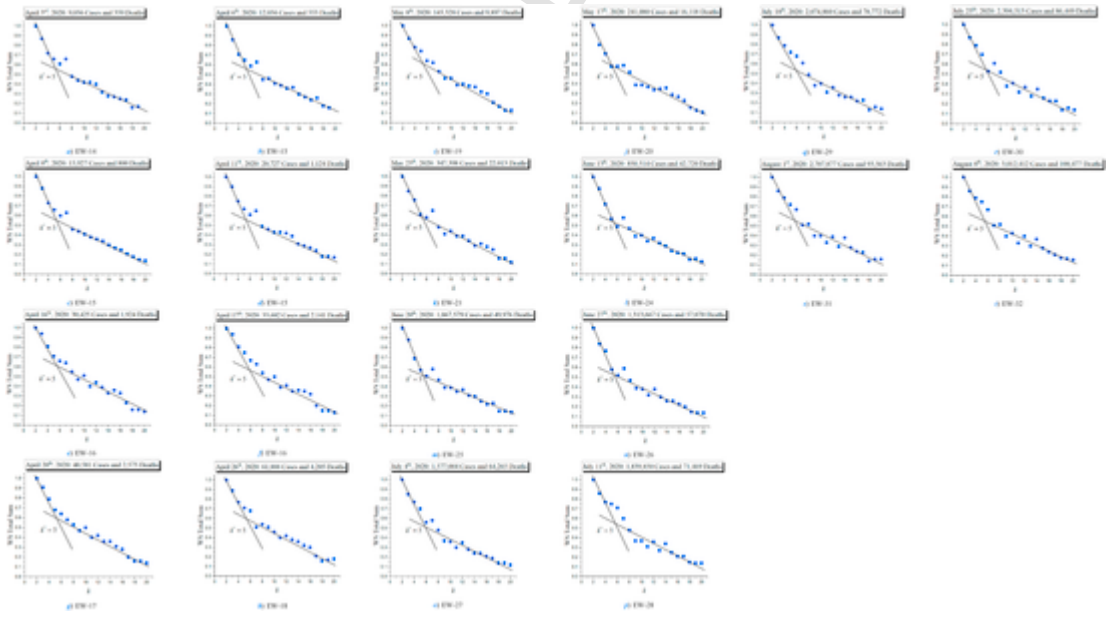


Fig. 1. *a-t*). Elbow data charts with the optimal number of clusters in all cases, $k^* = 5$, obtained from the tangent method (the intercept between lines). Covid-19 cases were taken from April 3rd up to August 8th, 2020. The corresponding epidemiological weeks (EW) are indicated.

Using standardized data from raw data presented in Supplementary Tables S1*a-t*) and S2, five medoids related to five respective clusters were chosen, in agreement with corresponding elbow data charts (Fig. 1*a-t*). Every group was defined according to their highest distances. For example, São Paulo presented the highest distance in all cases analyzed and was included in Group 5. Thus, each group was defined for every epidemiological bulletin, according to the first, second, third, fourth and fifth distances, in decreasing order.

Clustering studies on Covid-19 disease similar to this work are quite a few but is possible to cite the work of Machado and Lopes (Machado & Lopes, 2020), that observed the emergence of patterns and highlighted similarities and differences between some countries, as well as Zarikas et al. (2020).

Due to high correlations observed between variables as presented below, a factor analysis (FA) was done. For each epidemiological week it was necessary to verify the overall adequacy of the FA extraction considering the Kaiser-Meyer-Olkin statistic (KMO) and Bartlett's test of sphericity.

Supplementary Table S3 presents results corresponding to all epidemiological weeks. Almost all data presented a KMO above 0.5. The resulting Bartlett's test $\chi^2_{Bartlett}$ for the lowest value presented in Supplementary Table S3 is higher than the critical value $\chi^2_c = 65.656$ considering 45 degrees of freedom. It is thus possible to reject the null hypothesis that the correlation matrix of all epidemiological data are statistically equal to the identity matrix. Instead of analyzing if $\chi^2_{Bartlett} > \chi^2_c$, it was also possible to verify that the significance level of $\chi^2_{Bartlett}$ was less than 0.05. In fact, the *p*-value found was 4.964×10^{-39} for the lowest case, well below 5%. In resume, all cases were considered adequate.

A structural reduction of the data in order to create orthogonal factors was elaborated and observation rankings by using the same factors generated were defined. The weighted rank-sum criterion was considered in this work. In resume, for each observation, the values of all factors obtained, following Kaiser (1960) criterion (that takes into account eigenvalues greater than 1) were weighted by the respective proportions of shared variance added, with the subsequent ranking of the observations based on the results obtained (Favero & Belfiore, 2019, p. 408).

For all cases studied there were three to four eigenvalues higher than 1, and thus it was possible to extract between three and four factors, as presented in Supplementary Table S3. Taking into account all these factors, they covered between 78.9 and 91.8% of all cumulative variance related to such parameters. From such results was possible to rank all Brazilian states in each epidemiological week, after considering Varimax rotation (Favero & Belfiore, 2019, p. 397). In resume, from the rotated factors extracted it was possible to define a Brazilian state performance ranking. The top five are shown in Supplementary Table S3, and São Paulo and Rio de Janeiro are at the top two considering all epidemiological weeks.

4.1. Non-hierarchical data analysis of epidemiological week 32 in Brazil

Taking as example the moment which Brazil reached three million cases and a hundred thousand deaths (EW 32), as presented in combined data from Supplementary Tables S1*t* and S2 (as shown in Fig. 2*t*), it can be seen that São Paulo was the state with the highest number of Covid-19 cases, number of deaths, population and GDP. Roraima had the highest number of cases per 100 k inhabitants, the highest mortality per 100 k inhabitants, the lowest population and GDP. Rio de Janeiro presented the highest case fatality rate. Distrito Federal presented the highest population density, lowest area and highest HDI. Amazonas is the largest state with the lowest population density. From the same tables, Acre presented the lowest number of Covid-19 cases. Minas Gerais showed the lowest number of confirmed cases per 100 k inhabitants and mortalities per 100 k inhabitants. The lowest number of deaths was in Tocantins, the lowest case fatality rate was in Santa Catarina and the lowest HDI was in Alagoas.

Following this example, the first and largest cluster (Group 1) from combined Supplementary Tables S1*t* and S2 (EW 32, see Fig. 2*t*) was made up of (AC, AL, GO, MA, MS, PB, PI, RN, RO, SE, TO), and is related to some of the lowest values presented, as the number of Covid-19 cases, number of deaths and HDI. These states are located in the Mid-West, North and Northeast regions.

In contrast, São Paulo (SP), Minas Gerais (MG), Bahia (BA), Paraná (PR), Santa Catarina (SC) and Rio Grande do Sul (RS) form Group 5, due to some of the highest values presented in Supplementary Tables S1*t* and S2. SP is the richest state with the highest population, a high number of Covid-19 cases, relatives high HDI and population density as well as the highest number of deaths (followed by Bahia). MG has the second largest population. PR and RS are just below the richest states: MG, RJ and SP. SC presents the third highest HDI, just below DF and SP, in that order.

The next clusters found were:

Group 2: (AP, ES, MT, RR). This group is represented by states with lowest population and GDP, as well as the highest number of cases and mortality per 100 k inhabitants, all from Roraima. Amapá also presented the second highest number of cases per 100 k inhabitants.

Group 3: (DF). It is the smaller group, formed by the Federal District itself, that presented the highest population density and HDI but the lowest area. It also presented the third number of cases per 100 k inhabitants.

Group 4: (AM, CE, PA, PE, RJ). These are five states with highest number of case fatality rates (RJ, followed by PE and CE). RJ is also the third populous state (just after SP and MG) with the second highest population density and GDP. Only PA has a lower total area than AM, the biggest Brazilian state with the lowest density area. CE, RJ and PA presented the third, fourth and fifth number of Covid-19 cases. CE, RJ, AM and PE showed the second, third, fourth and fifth number of deaths, just after SP. RJ, CE, PE and PA presented the second, third, fourth and fifth position on mortality per 100 k inhabitants, just after RR. RJ, PE and CE were in the first three positions of case fatality rates.

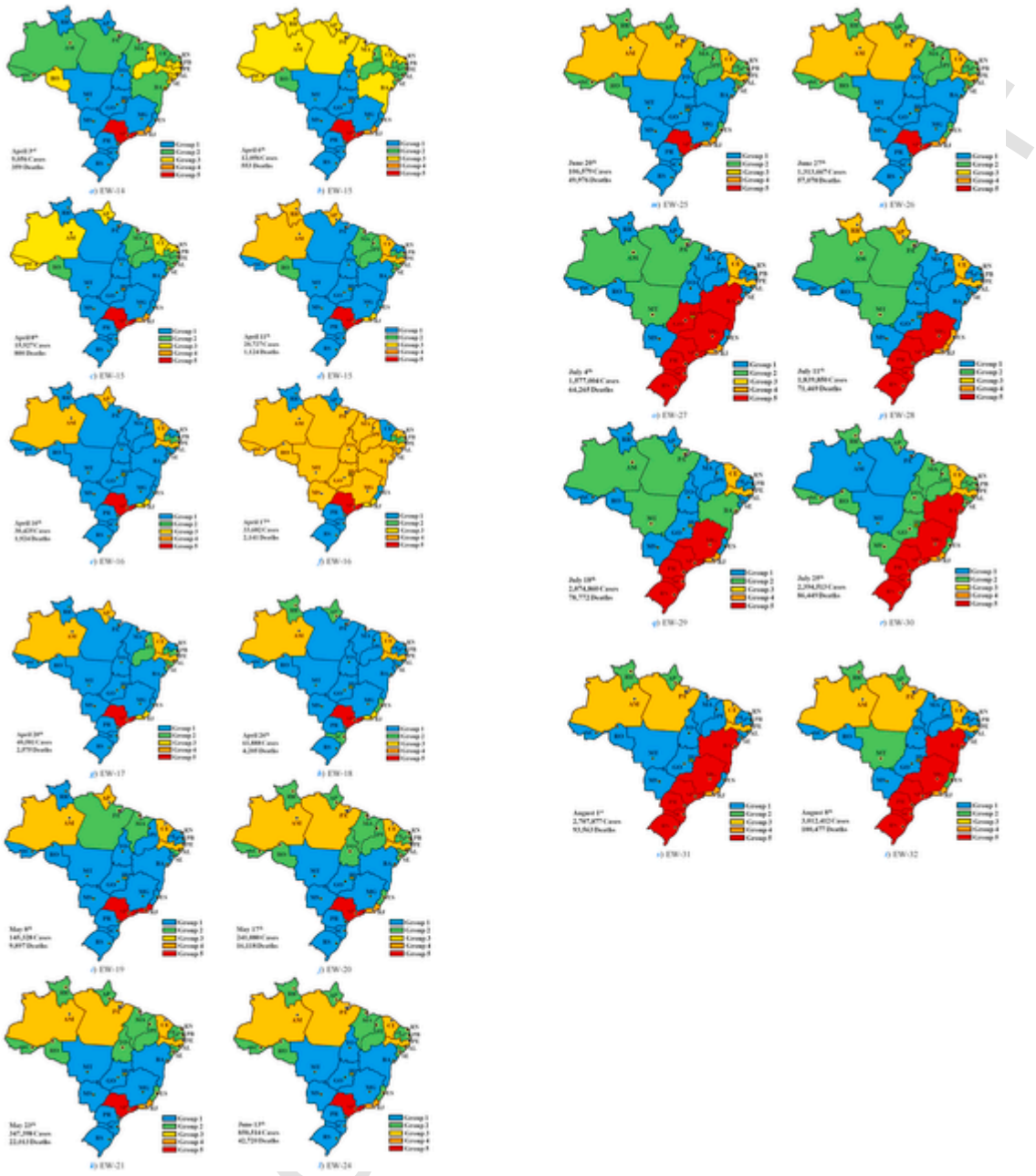


Fig. 2. *a-t*). Spatiotemporal evolution of Covid-19 in Brazilian states, that were grouped in five *k*-means clusters, from April 3rd up to August 8th, 2020. The corresponding epidemiological weeks (EW) are indicated, from 14 to 32.

Brazil is the world's second biggest Covid-19 hotspot after the United States during the period analyzed (EW 32). As there are concerns and early signs that infections are moving inland into smaller cities with inadequate provision of intensive care beds and ventilators (Anonymous, 2020), the present results show that special care should be taken in such state groups. Fig. 2*t* summarizes Brazilian states grouped in five clusters. It is important to note that Group 5 represents South and Southeast (plus Bahia, at Northeast), that presented the highest number of cases and deaths.

According to Hair et al. (2019, p. 813), distance measures focus on magnitude values and portray as similar the unities that are closer, even including their different patterns across the variables. In contrast, correlation measures focus on the patterns across the variables and do not consider the magnitude of the differences between Brazilian states. A correlational analysis focuses on patterns rather than the more traditional distance measure and requires a different interpretation of the results. However, such analyses are still relevant.

Thus, following Pearson's correlation coefficient (1896), the results show that number of confirmed Covid-19 cases and the number of Covid-19 deaths are *highly* correlated as expected, with a coefficient correlation $r(X_1,X_2) = 0.926$ considering all data available in Supplementary Tables S1*t* and S2. Another high correlation was observed between GDP with population size, the sum of confirmed Covid-19 cases and the sum of deaths. The $r(X_i,X_j)$ coefficients were $r(X_9,X_6) = 0.959$, $r(X_9,X_1) = 0.926$ and $r(X_9,X_2) = 0.889$, respectively. These particular correlations were verified for most of epidemiological weeks and is partially in agreement with Souza et al. (2020), that observed a positive association between higher per-capita income and Covid-19 diagnosis. Also observed was a high correlation between population with the number of confirmed Covid-19 cases and the number of deaths, with $r(X_1,X_6) = 0.931$ and $r(X_2,X_6) = 0.886$, respectively.

In contrast, the same procedure was carried with correlation coefficients near zero. According to Supplementary Tables S1*t* and S2, there was no correlation between area with all other parameters. The same was observed between population density with confirmed Covid-19 cases per 100 k inhabitants, resulting in $r(X_3,X_8) = 0.074$. Also, no correlation was found between mortality per 100 k inhabitants with the sum of confirmed Covid-19 cases and HDI, with $r(X_4,X_1) = 0.053$. This result is opposite to others considering different EW. No correlation was observed between GDP and area, resulting in $r(X_7,X_9) = -0.032$. Finally, no correlation was observed between HDI and confirmed Covid-19 cases per 100 k inhabitants as well as case fatality rate, with $r(X_3,X_{10}) = -0.085$ and $r(X_5,X_{10}) = 0.035$, respectively.

These correlation values partly explain the fourth (AM, CE, PA, PE, RJ) and fifth (BA, MG, PR, RS, SC, SP) clusters, which showed the highest number of confirmed Covid-19 cases, number of deaths, population and GDP. One way of improving this analysis is to obtain more data, for example of hospital bed occupancy rates, the number of ventilators available and so on.

4.2. Factor analysis of epidemiological week 32 in Brazil

The factor analysis of EW 32 is briefly presented below. The corresponding KMO value, as shown in Supplementary Table S3, is less than 0.5. This value suggests that the overall adequacy of FA would be unacceptable. However, in agreement with Favero and Belfiore (2019, p. 1170) one should always favor Bartlett's test of sphericity over the KMO statistic to take a decision about the factor analysis's overall adequacy. The corresponding *p*-value (related to Bartlett's test) considering 45 degrees of freedom was 4.964×10^{-39}, well below the significance level of 5%. Thus, is possible to conclude that the FA was adequate for this particular week.

As there are four eigenvalues higher than 1 related to this particular EW (4.063, 1.641, 1.637 and 1.564, respectively), all related to the Kaiser (1960) or latent root criterion, four rotated factors were considered, covering 89.033% of all variance related to all variables.

Fig. 3 shows the loading plots constructed from the rotated factor loadings (Pearson correlations) related to EW 32. Fig. 3*a* presents the mapping distribution of the first two components, and Fig. 3*b* presents the mapping distribution of components 2 and 3. Most data were characterized by the first axes. The first component represents 40.626% of variability, followed by the second component, representing 16.406% of variability, and by the third component, that represents 16.366% of variability. The cumulative percentage of all four components is shown in Supplementary Table S3, representing 89.033% of variability.

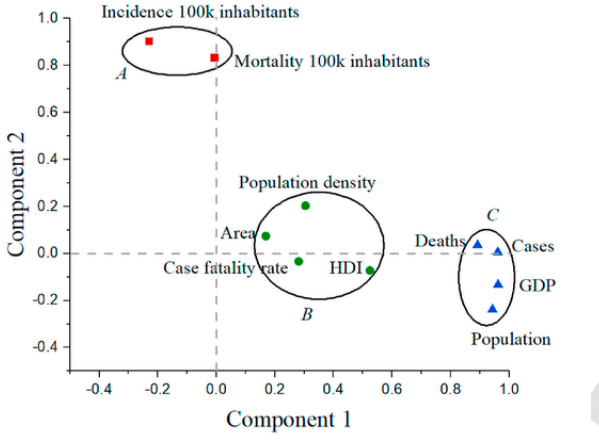
By analyzing the loading plot of Fig. 3*a*, the behavior of the Pearson correlations becomes clearer. While the variables number of Covid-19 cases (Z_1), number of deaths (Z_2), population (Z_6) and GDP (Z_9) show high correlation with the first factor (abscissa or *X*-axis), the variables confirmed cases or incidence per 100 k inhabitants (Z_3) and mortality per 100 k inhabitants (Z_4) shown strong correlation with the second factor (ordinate, or *Y*-axis).

From the rotated factor loadings, the rotated factor expressions can be expressed in terms of standardized Z_i variables as:

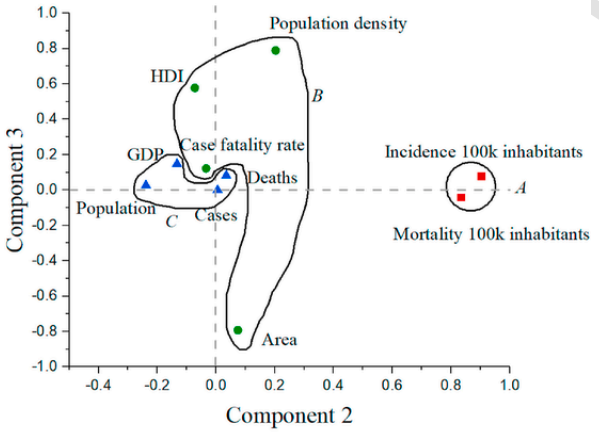
$$\begin{aligned}
F'_1 &= +0.285Z_1 + 0.205Z_2 + 0.080Z_3 + 0.012Z_4 - 0.089Z_5 + 0.245Z_6 + 0.168Z_7 - 0.008Z_8 + 0.265Z_9 + 0.142Z_{10},\\
F'_2 &= +0.104Z_1 + 0.071Z_2 + 0.606Z_3 + 0.485Z_4 - 0.118Z_5 - 0.063Z_6 + 0.130Z_7 + 0.093Z_8 + 0.014Z_9 + 0.018Z_{10},\\
F'_3 &= -0.116Z_1 - 0.046Z_2 + 0.010Z_3 - 0.069Z_4 + 0.075Z_5 - 0.078Z_6 - 0.552Z_7 + 0.476Z_8 - 0.010Z_9 + 0.314Z_{10},\\
F'_4 &= -0.079Z_1 + 0.152Z_2 - 0.307Z_3 + 0.284Z_4 + 0.644Z_5 - 0.041Z_6 - 0.097Z_7 + 0.080Z_8 - 0.123Z_9 - 0.305Z_{10}
\end{aligned} \tag{1}$$

All Z_i variables contribute to increase the rotated factors with different intensities due to their different coefficients. The coefficients (or scores) of first factor (F'_1) in Eq. (1) are related in order to Z_1, Z_9, Z_6, Z_2, Z_7, Z_{10}, Z_3 and Z_4 and to their respective eigenvectors. That is to say, F'_1 will be high when Z_1, Z_2, Z_3, Z_4, Z_6, Z_7, Z_9 and Z_{10} are high and Z_8 and Z_5 are low. On the other hand, F'_1 will be low Z_1, Z_2, Z_3, Z_4, Z_6, Z_7, Z_9 and Z_{10} are low but Z_8 and Z_5 are high. Due to their lower scores, standard variables Z_3, Z_4, Z_5 and Z_8 show weaker correlations with the first factor, in agreement with Fig. 3*a*. The same interpretation can be done for the other rotated factors. Also, based on the above factor expressions and on the standardized variables, it is possible to calculate the values corresponding to each factor for each observation.

Also, from Fig. 3 is possible to group variables in three groups, *A*, *B*, *C*: (confirmed cases or incidence per 100 k inhabitants, mortality per 100 k inhabitants), (case fatality rates, population density, area, HDI) and (number of Covid-19 cases, number of deaths, population, GDP), respectively. First group (*A*) contains only health indicators and is strongly related to second factor, F'_2; all other groups presented health, geographic, economic and social indicators, but the third group (*C*) is strongly related to the first factor, F'_1.



a)



b)

Fig. 3. Loading plots with rotated loadings of EW 32 for components 1 and 2 (*a*) and components 2 and 3 (*b*). The first two components represent 57.032% of all variability, and components 2 and 3 represent 32.772% of all variability. From these figures is possible to group variables in three clusters, *A*, *B*, *C*: (confirmed cases or incidence per 100 k inhabitants, mortality per 100 k inhabitants), (case fatality rates, population density, area, HDI) and (number of Covid-19 cases, number of deaths, population, GDP), respectively.

Finally, from data presented in Supplementary Tables S1*t* and S2 it was possible to propose a performance ranking of the Brazilian states in this particular week, as shown by Eq. (2):

$$ranking(\mathrm{EW32}) = 0.40626F'_1 + 0.16406F'_2 + 0.16366F'_3 + 0.15635F'_4 \quad (2)$$

Supplementary Table S3 shows the top five Brazilian states performance resulting from the ranking created after the sum of the values obtained from the factors weighted by the respective proportions of shared variance. This ranking points to SP as the first Brazilian state, followed by RJ, DF, CE and PE. This rank presents similarities to cluster analysis results, showing states from Group 5 (SP) followed by Group 4 (RJ, CE, PE) and Group 3 (DF).

4.3. Some general remarks

The analysis of the spatiotemporal evolution of the confirmed cases and deaths as presented in Fig. 2 has considerable interest from the point of view of delivering good information to physicians, health organizations, policymakers, politicians, economy/finance experts and to the general public. The Covid-19 pandemic is exposing Brazilian Health System structural weaknesses and bottlenecks, in particular the lack, or unequal distribution of hospitals, health workers and medications. Besides Brazil has is the world's largest public and universal health system, is possible to note the absence of medium and high complexity care infrastructure and limited capacity to produce and perform diagnostic tests, as well as PPE (Oliveira et al., 2020).

Clearly, *k*-means and FA were useful for data reduction. The results presented in Fig. 2 lead to the emergence of patterns both highlighting the main groups and the similarities or dissimilarities between them. As presented in Supplementary Table S2, the States

of Rio de Janeiro and São Paulo have 63 million inhabitants, which represents 30% of the Brazilian population. According to de Souza et al. (2020), although the first confirmed cases of Covid-19 and deaths from the disease were reported in those two states, it remains unclear whether they were the gateways for the disease to enter the country.

By EW 16, and according Supplementary Table S1*e*, all Brazilian states had reported Covid-19-related deaths. From Fig. 2*a-t*) is possible to observe that Group 5, represented by the São Paulo state, did not change its position in most of epidemiological weeks, and ranked as first in all such weeks by means of FA.

From non-hierarchical clustering, three states changed from one group to another (DF, PR, RS), eight ones changed three times (MG, MS, MT, PB, PE, RJ, SC, TO), fourteen changed four times (AC, AL, AM, AP, CE, ES, GO, MA, PA, PI, RN, RO, RR, SE) and just one state changed five times (BA). These results are summarized in Fig. 4, and reflects data presented in Supplementary Tables S1 and S2. Considering different epidemiological weeks, one might expect that characteristics of same groups would change with one exception: on May 17th and May 23rd there was observed the same cluster results.

The variability in such results should also include health care coverage, the occurrence of lockdowns in some regions and periods, or the contradictory recommendations issued by government authorities (Anonymous, 2020); (Werneck & Carvalho, 2020). For example, in early April, only 53% of Brazilian city dwellers stayed home (Dyer, 2020). This number changes in the next weeks, reflecting changes presented in Fig. 2.

According to Souza et al. (2020), the Southern states of Paraná, Santa Catarina and Rio Grande do Sul have three important determinants of mortality: older populations than those of other regions of Brazil; highest historical severe acute respiratory syndrome (SARS) incidence; and a fragile health care network, albeit more structured than other states. In addition, their proximity to the states of Rio de Janeiro and São Paulo represents and additional complicating factor, due to travel facilities between Southern and Southeast regions. This partially explains the variability of these states between groups, particularly during the last epidemiological weeks observed (from EW 27 to 32). In the Northeastern and Northern regions, social vulnerability is a chronic problem, partially reflected by geographical, economic and social data presented in Supplementary Table S2. For example, the Amazonas state, one of the epicenters of the pandemic in Northern region in the first epidemiological weeks, had only 271 intensive care units, or 6.5 beds/100,000 population (Rache et al., 2020).

The Covid-19 challenges are even greater in Brazil, since little is known about the characteristics of the disease transmission in a context of poverty, with huge social inequalities. The pandemic has reached the Brazilian population in a scenario of extreme vulnerability, with severe budget cuts in social policies and high unemployment rates (Werneck & Carvalho, 2020; Souza et al., 2020). There are many communities exposed to precarious housing and sanitation conditions, without systematic access to running water and inadequate provision of intensive health cares (Miller et al., 2020; Werneck & Carvalho, 2020). It is crucial to minimize the social, economic and psychological harms to the most vulnerable groups through the adoption of appropriate social and health measures (Werneck & Carvalho, 2020).

Covid-19 tests are very important, even during the no-symptom phase, according to Song et al. (2020). They observed the Covid-19 transmission in four Chinese families, totalizing 24 persons, with 22 infected. Among those, 20/22 had mild symptoms, including almost all children, and two presented moderate to severe clinical manifestations. The virus incubation period varied from two to thirteen days, but a period of virus shedding over a month was observed in six cases. Song et al. (2020) showed how Covid-19 spreads within families and observed that adults were more likely to be symptomatic compared to children.

While the incidence curves for European countries have consistently flattened and declined since the implementation of non-pharmaceutical interventions (NPI), Brazil's daily incidence curve has continued to increase. The most prevalent comorbidities due to Covid-19 in Brazil were cardiovascular disease and diabetes in middle- or older-aged individuals (Souza et al., 2020).

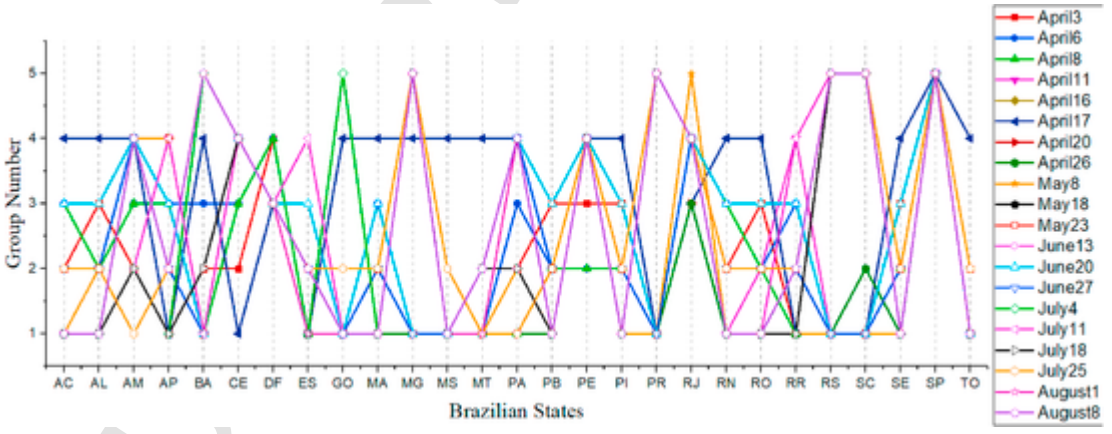


Fig. 4. Spatiotemporal evolution of all five groups, from April 3rd up to August 8th, 2020. These dates correspond to epidemiological weeks (EW) 14–32. It is possible to note that three states changed groups twice (DF, PR, RS), eight ones changed three times (MG, MS, MT, PB, PE, RJ, SC, TO), fourteen changed four times (AC, AL, AM, AP, CE, ES, GO, MA, PA, PI, RN, RO, RR, SE) and just one state changed five times (BA). Only SP remained fixed.

The present results can be applied for any country. The mathematical procedure can be considered adequate, *i.e.* with high degree of accuracy, despite being fallibilistic, following a *Big Data* sense (Alger, 2019, p. 456; Santos et al., 2019) because all the analyses were related to data available - *i.e.* more data would (probably) have given other results.

5. Conclusions

Due to the continental dimensions of Brazil and its internal social, economic, geographic and cultural inequalities, the impact of the disease might be heterogeneous. However, according to the non-hierarchical *k*-means algorithm it was possible to group Brazilian states into five clusters based on Covid-19 health indicators, as well as geographic, economic and social criteria.

In this work was adopted different forms for organizing and visualizing the data, representing the spread dynamics across different states. Besides the number of Covid-19 cases and deaths, the dynamic characteristics of this multivariate analysis play an important role that is not evident in standard representations.

The present data-driven study showed that, despite spatiotemporal differences of Covid-19 among Brazilian states might reflect social, economic, cultural, and structural inequalities, they were grouped into five clusters only. Just one had a fixed state, the fifth group that was represented by São Paulo in all weeks analyzed. Another relevant finding was that the number of clusters did not change over twenty days, even considering great difference on health indicators. Some clusters were more critical due to specific variables, including cities that are main hotspots in SP, RJ, CE, PE, AM, MA and PA during early April, for example.

A performance ranking of the Brazilian states based on factor analysis was proposed. São Paulo and Rio de Janeiro were at top of this list considering all epidemiological weeks. It is essential to note that the creation of performance rankings is a static procedure, because the inclusion of new observations or variables may alter the factor scores.

Considering the epidemiological week 32, the results show that the number of confirmed Covid-19 cases and the number of deaths were highly correlated. Some correlation values partly explain the existence of some clusters, as some that show the highest GDP with population densities, high number of confirmed Covid-19 cases and number of deaths.

It was also possible to group variables in three clusters, *A*, *B*, *C*, using FA. The first group contains only health indicators linked to first component, and other groups presented health, geographic, economic and social indicators related to specific factor components.

It is well-know that socioeconomic differences are associated with access to healthcare and should be taken into account when designing targeted interventions. As data-driven analyses are urgently needed to help tackle health inequities during the ongoing epidemic in Brazil, *k*-means and FA offer interesting fields of investigation. The combined information from a few variables could be useful to a variety of different policymaker decisions and new mitigation strategies with the aim of avoiding the occurrence of severe cases and deaths from the disease. These mathematical approaches would shed light on traditional and classical analysis, offering new possibilities, such as minimizing some subjectivities in health care issues.

Acknowledgements

The author thanks E. D. F. S. L. Santos for the development of Python code and A. P. Ricieri for his example on the partition of the numbers between 1 and 9 into two clusters. This work was supported by the National Council for Scientific and Technological Development (CNPq) [304705/2015–2, 404004/2016–4 and 305331/2018–3].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.idm.2020.08.012.

References

Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., … Groop, L. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: A data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol, 6, 361–369.

Alger, B.E. (2019). Defense of the scientific hypothesis: From reproducibility crisis to Big data. New York: Oxford University Press.

Anonymous (2020). COVID-19 in Brazil: "So what?". Lancet, 395, 1461. 1461.

Bartlett, M.S. (1954). A note on the multiplying factors for various χ^2 approximations. Journal of the Royal Statistical Society: Series B, 16, 296–298.

Borba, M.G.S., Val, F.F.A., Sampaio, V.S., Alexandre, M.A.A., Melo, G.C., Brito, M., … Lacerda, M.V.G. (2020). Effect of high vs low doses of chloroquine diphosphate as adjunctive therapy for patients hospitalized with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection A randomized clinical trial. JAMA Network Open, 3, 1–14.

Cipitelli, M.C., Valentin, E., da Cruz, N.V.G., Nogueira, T.L.S., Amaro, E., Silva, R., … Santos, C.G.M. (2020). SARS-CoV-2 diagnostic diary: From rumors to the first case. Early reports of molecular tests from the military research and diagnostic institute of Rio de Janeiro [Submitted. Mem Inst Oswaldo Cruz E-pub. doi:10.1590/0074-02760200200. 30 Apr 2020.

Conde, M. (2020). Brazil in the time of coronavirus. Geopolítica(s), 11, 239–249.

Dyer, O. (2020). Covid-19: Brazil's president rallies supporters against social distancing. British Medical Journal, 369, 1. 1.

Everitt, B.S., Landau, S., Morven Leese, M., & Stahl, D. (2011). Cluster Analysis. West Sussex: John Wiley & Sons, Ltd.

Favero, L.P., & Belfiore, P. (2019). Data science for business and decision making. Cambridge: Academic Press.

Hair, J.F., Jr., Black, W.C., Babin, B.J., & Anderson, R.E. (2019). Multivariate data analysis. Hampshire: Cengage.

IBGE - Brazilian Institute of Geography and statisticswww.IBGE.gov.br23 May 2020

IPEA - Institute of Applied Economic Research www.IPEA.gov.br23 May 2020

Kaiser, H.F. (1958). The Varimax criterion for analytic rotation in factor Analysis. Psychometrika, 23, 187–200.

Kaiser, H.F. (1960). The application of electronic computers to factor Analysis. Educational and Psychological Measurement, 20, 141–151.

Kaiser, H.F. (1970). A second generation little jiffy. Psychometrika, 35, 401–415.

Kaplanis, J., Gordon, A., Shor, T., WeissbrodBO, Geiger, D., Wahl, M., … Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. Science, 360, 171–175.

Kopec, D. (2019). Classic computer science problems in Python. Manning Publications Co.. Shelter Island 206 pp..
Machado, J. A. Tenreiro, & Lopes, A. M. (2020). Rare and extreme events: the case of COVID-19 pandemic. Nonlinear Dyn., 100, 2953–2972. doi:10.1007/s11071-020-05680-w.
MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281–297.
Miller, M.J., Loaiza, J.R., Takyar, A., & Gilman, R.H. (2020). COVID-19 in Latin America: Novel transmission dynamics for a global pandemic? PLoS Neglected Tropical Diseases, 14, e0008265.
Oliveira, W.K., Duarte, E., de França, GVA, & Garcia, L.P. (2020). Epidemiol. Serv. Saúde, 29, 1–8.
Souza, W.M., Buss, L.F., Candido, D.S., Carrera, J.-P., Li, S., Zarebski, A.E., … Faria, N.R. (2020). Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nat. Hum. Behav., 4, 856–865.
Souza, C.D.F., Paiva, J.P.S., Leal, T.C., Silva, L.F., & Santos, L.G. (2020). Spatiotemporal evolution of case fatality rates of COVID-19 in Brazil, 2020. Jornal Brasileiro de Pneumologia, 46, 1–3.
Pearson, K. (1896). VII. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society of London, 187, 253–318.
Rache, B., Rocha, R., Nunes, L., Spinola, P., Malik, A.M., & Massuda, A. (2020 Mar). Necessidades de Infraestrutura do SUS em Preparo à COVID-19: Leitos de UTI, Respiradores e Ocupação Hospitalar (Nota Técnica n. 3). Instituto de Estudos para Políticas de Saúde.
Santos, E.D.F.S.L., Pimentel, C.A.R., & Ricieri, A.P. (2019). Aerotaxonomy. J. Prod. Auto., 2, 41–60.
Song, R., Han, B., Song, M., Wang, L., Conlon, C.P., Dong, T., … Li, X. (2020). Clinical and epidemiological features of COVID-19 family clusters in Beijing, China. Journal of Infection, 81, e26–e30.
Spearman, C. (1904). General intelligence," objectively determined and measured. American Journal of Psychology, 15, 201–292.
SUS - coronavirus Brasilhttps://covid.saude.gov.br202023 May 2020
Werneck, G.L., & Carvalho, M.S. (2020). The COVID-19 pandemic in Brazil: Chronicle of a health crisis foretold. Cadernos de Saúde Pública, 36, 1–4.
World Health Organizationwww.WHO.int202023 May 2020
Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., … Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579, 265–269.
Zarikas, V., Poulopoulos, S.G., Gareiou, Z., & Zervas, E. (2020). Clustering analysis of countries using the COVID-19 cases dataset. Data in Brief, 31, 105787.