Identifying Similar Groups of Countries According to the Impact of Corona Virus (COVID-19) by a Two-Layer Clustering Method

Juan Riofrío\(^{(2)}\), Carlos Muñoz-Moncayo, Isidro R. Amaro, and Israel Pineda

Yachay Tech University, Hda. San José S/n Y Proyecto Yachay, Uruqu‘í 100119, Ecuador
{juan.riofrio,carlos.munoz,iamaro,ipineda}@yachaytech.edu.ec

Abstract. This paper presents a new clustering algorithm to identify groups of countries. First, a layer of several clustering methods is applied to the original dataset. Then, after performing dimensionality reduction techniques like t-SNE or SOM on the resulting data, a second clustering layer (K-Means) is applied to identify the final clusters. This method is applied to a dataset from 163 countries, considering the following variables population, area, Gross Domestic Product (GDP), Gross Domestic Product adjusted for Purchase Power Parity (GDP-PPP), and COVID-19 related data (Confirmed, Recovered, and Deaths). The implementation with SOM dimensionality reduction outperformed the one with t-SNE for the considered dataset. We expect that using this information, countries can have an insight on which measures against COVID-19 replicate or avoid, based on the results in countries from the same cluster.

Keywords: Clustering · COVID-19 · K-Means · Self-Organizing Map · Two-layer clustering

1 Introduction

Since an abnormal amount of transmittable pneumonia cases, similar to SARS-CoV-1, was reported in Wuhan on December 2019, authorities of China began fighting what seemed to be an epidemic outbreak. Indeed, this disease was identified as Novel Coronavirus (COVID-19), caused by the extremely contagious virus SARS-CoV-2. Some of its common symptoms are fever, dry cough, and dyspnoea; in severe cases it can even cause pneumonia, sepsis, and septic shock, see e.g., [4]. On March 11, 2020, the WHO characterized COVID-19 as a pandemic; since then, unwavering efforts from countries worldwide have been made to control the spreading and develop a vaccine to prevent it. Science-based decisions taken by politicians around the world is a decisive factor in what course the human race will take against such a threat against its well-being. Therefore, to analyze how the virus spreads and its global impact is essential.
Data Science, with its wide and versatile fields of study such as Big Data, Data Mining, and Machine Learning, is a well-suited discipline to analyze large datasets as the provided by the COVID-19 tracking worldwide. Clustering approaches have been proposed and shown to be of great significance for decision takers, see [10]. In this work, we present a novel method, based in a two-layer clustering, applied to data concerning the global impact of COVID-19. In the first layer of clustering, we use several well-known methods, such as Affinity Propagation, Agglomerative Clustering, BIRCH, K-Means, and Mini Batch K-Means. In the second layer, we use a combination of t-SNE/SOM and K-Means, as illustrated in Fig. 1. The variables considered in this work are the population, area, the number of confirmed cases, recovered cases, and deaths related to COVID-19 in 163 countries.

The remaining of this paper is structured as follows: Sect. 2 describes the dataset used and overviews the clustering methods considered in this work as well as the proposed experimental methodology. Section 3 gathers the results and discussion. Finally, the concluding remarks are drawn in Sect. 4.

2 Materials and Methods

In this section, we give an insight into the different stages of the algorithm. The first stage is fetching and preprocessing of data regarding COVID-19. Then, we review some clustering, comparison, and dimensionality reduction methods. Figure 1 presents the methodology and all the steps executed during the research.

2.1 Dataset

In the face of the Novel Corona Virus outbreak, the United Nations Office for the Coordination of Humanitarian Affairs (OCHA), in collaboration with the Johns Hopkins University Center for Systems Science and Engineering, gathered data from 266 countries related to the health crisis. We fetched this data from the Humanitarian Data Exchange, a platform managed by the OCHA [16]. For technical reasons, we reduced the number of observations to 163. The dataset
contains the total number of confirmed cases, deaths, and recoveries of COVID-19 of each country. It has been updated each day from January 22 of 2020, to June 29 of 2020. In this study, we consider the data corresponding to June 23, 2020.

To develop a broader analysis, and since recent works, e.g., [3,6], suggest that a correlation between population density and impact of COVID-19 may exist, we considered the area and population of each country. This data was retrieved from [13], and its last update was done on April 18, 2020. Finally, as stated in the previous section, it can be expected that the economic status of a country influences the way COVID-19 affects it and vice versa. Therefore, we took into account the Gross Domestic Product (GDP) and the Gross Domestic Product adjusted for Purchase Power Parity (GDP-PPP). Data corresponding to GDP and GDP-PPP were fetched from the International Monetary Fund report [19] and [20], and last updated in April 2019 and April 2020, respectively.

2.2 Data Clustering

When searching for patterns within a voluminous data set, it is key to classify the observations in distinguishable groups. Data clustering is a widely used method to do so. Its main characteristic is that classification groups or clusters are not explicitly provided a priori. They depend on the data instead. The process of performing clustering depends on several factors, such as the number of clusters, the desired approach, and the nature of the data.

We want elements of a cluster to be similar and elements from different clusters to be dissimilar. A widely used approach is to choose some observations adequately, often referred to as exemplar of each cluster, and then assigning each observation to the cluster corresponding to the closest exemplar observation. To do this, it is required to quantify the distance between observations, we do so using a function denoted similarity measure.

| Country     | Population | Area (Km²) | GDP (Billions) | GDP-PPP (Billions) | Confirmed | Recovered | Deaths |
|-------------|------------|------------|----------------|--------------------|-----------|-----------|--------|
| Afghanistan | 38742911   | 652860     | 20.682         | 74.79              | 29481     | 9260      | 618    |
| Albania     | 2878420    | 27400      | 17.21          | 38.08              | 2047      | 1195      | 45     |
| Algeria     | 43685618   | 2381740    | 193.0          | 638.36             | 12076     | 8674      | 861    |
| ...         | ...        | ...        | ...            | ...                | ...       | ...       | ...    |

As shown in Table 1, after preprocessing, all of the variables considered in our first dataset, which we will refer as Data 1, are numerical. Therefore, standard similarity measures between entries such as Euclidean, Minkowski or Mahalanobis distance may be used [15]. With the help of the Machine Learning package for Python, Scikit-learn [11], we applied different clustering methods on Data 1. Let us briefly describe them, assuming we are dealing with n observations and
Identifying Similar Groups of Countries According to the Impact

$p$ variables, thus a data matrix $Y_{n \times p} = (y_{ij})$; we denote the similarity between observations $y_i$ and $y_j$ by $s_{ij} := s(y_i, y_j)$.

**Affinity Propagation.** In this method, the similarity is often computed by $s_{ij} = -\|y_i - y_j\|^2$, although it can be set according to different necessities. The main characteristic of Affinity Propagation is that all the observations are candidates to be exemplars from the beginning [5]. The process followed to define the exemplars. Therefore, the clusters rely on sending messages between observations, and the following iterative process does it:

1. For all $i, j$, the responsibility $r_{ij}$, is updated. Each $r_{ij}$ is a real-valued indicator of how optimal is the observation $x_j$ to be set as an exemplar for the observation $x_i$, compared to other candidate observations to be exemplars for $x_i$. It is computed by

$$r_{ij} = s_{ij} - \max_{j' \neq j} \{a_{ij'} + s_{ij'}\},$$

where $a_{ij}$ is the availability; it denotes how ideal it would be for $x_i$ to choose $x_j$ as its exemplar, considering how optimal is $x_j$ to be the exemplar for other observations.

2. For all $i, j$ the availability is updated by

$$a_{ij} = \min \left(0, r_{jj} + \sum_{i' \in \{ij\}} \max(0, r_{i'j}) \right), \quad i \neq j, \quad (2)$$

$$a_{jj} = \sum_{i' \neq j} \max(0, r_{i'j}). \quad (3)$$

This process is repeated until the members of each cluster remain fixed for several iterations or a limit number of iterations is exceeded. The final exemplars are those $x_j$ that satisfy $r_{jj} + a_{jj} > 0$.

**Agglomerative Clustering.** This is a hierarchical method, in which to begin, every observation is set as a cluster and if not cut before, the whole dataset becomes a unique cluster at the end. Since the last is a trivial result, it is often required that the final number of clusters is given a priori. First, a distance, metric, or dissimilarity measure between observations needs to be defined. In our case, it is the Manhattan distance:

$$d_{ij} := \|x_i - x_j\|_1 = \sum_{k=1}^{p} |x_{ik} - x_{jk}|. \quad (4)$$

Then a linkage criteria or distance between clusters is defined, we use the complete linkage:

$$Lnk(A, B) := \max\{d_{ij} : x_i \in A, x_j \in B\}. \quad (5)$$
The agglomeration of clusters is done in an iterative process in which, at each step, two clusters are merged into one if the linkage between them is minimized. This is repeated until the desired number of clusters is reached. As can be expected, the complexity of this algorithm is considerable, $O(n^3)$ for time and $O(n^2)$ for memory.

**BIRCH.** This method works by building a Clustering Feature Tree (CFT) for the data, which is compacted into a set of Clustering Feature Nodes (CFN). The CFN has subclusters called Clustering Feature subclusters (CFS), and CFS can have CFN as children if they are not in terminal CFN (leaves). Each CFS carries the number of samples in a subcluster, the total sum of samples, the squared euclidean sum of samples, and the centroids. This is the information needed for clustering, and it is done to not handle the whole dataset, which is especially useful when dealing with voluminous datasets.

The algorithm provided by Scikit [11] requires two hyperparameters: A threshold to control the distance between a subcluster and a potential member observation, and a branching factor that limits the number of subclusters in a CFN. It works in the following way:

1. An observation is placed in the first CFN or root and then merged with one of its subclusters such that the radius is minimized and the threshold and branching constraints. If there are child nodes to this subcluster, this process is repeated until a leaf is reached; then, the properties of all crossed subclusters is updated.
2. If the radius of the subcluster exceeds the threshold and has a greater amount of subclusters than the branching factor, the observation is stored in a separate space. The two subclusters with greater distance between them are considered, and their subclusters are divided into two groups based on the distance between them.
3. If the separate node has a parent subcluster and the branching factor is exceeded, the parent is divided. If there is still no space, this process is repeated until the root is reached.

**K-Means.** This is the main method considered in this work; through it, we want to partition our dataset into k clusters, while minimizing the variance within each cluster or cluster’s inertia, i.e., we want to minimize

$$
\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2,
$$

(6)

Where each $C_i$ is a cluster and $\mu_i$ its corresponding mean. The clustering is performed in the following way:

1. A centroid for each cluster is selected (k in total); in the simplest case these are observations.
2. Clusters are constructed by assigning each observation to its nearest centroid.
3. For each cluster $C_i$, the mean $\mu_i$ is computed and assigned as the new centroid.
4. For each cluster, the distance $d(\mu_i, \tilde{\mu}_i)$ between the new centroid $\mu_i$ and the old centroid is computed $\tilde{\mu}_i$.

Steps 2), 3) and 4) are repeated until

$$\sum_{i=1}^{k} d(\mu_i, \tilde{\mu}_i) < \gamma,$$

(7)

For a threshold $\gamma$, fixed a priori. For congruence with 6, it is common to choose $d$ as the squared euclidean distance. Even though the k-means algorithm has complexity $O(n^2k)$ [8], some implementations aiming to maximize distance between initial centroids show a faster performance.

**Mini Batch K-Means.** This is a variation of the K-means algorithm aimed to make calculations in less time. It works in the following way:

1. k clusters and centroids are set as in k means.
2. A random subset of the data is taken, and each point is assigned to the nearest centroid.
3. For each point $x$ from the mini-batch $M$, its assigned centroid $c$ is updated using a convex combination of itself and the point:

$$c = (1 - \rho)c + \rho x.$$  

(8)

4. Each observation is assigned to its nearest centroid.

Steps 2, 3, and 4 are repeated until the centroids do not have significant changes. This is done by letting $\rho$ go to zero as the number of iterations increases. This method is faster than K-Means with a loss in the quality of the results.

**Self-Organizing Map (SOM).** A Self-Organizing Map (SOM) [9] is a type of artificial neural network that is trained through unsupervised learning and belongs to the category of competency learning networks. These types of artificial neural networks use a neighborhood function to preserve the topological properties of the input space. It is used as a grouping and visualization tool in exploratory data analysis. The primary goal of a SOM is to transform a high-dimensional input space into a smaller, discrete output space (dimensionality reduction) while preserving relationships in the data, but not actual distances [2]. SOM has three characteristic steps [7]:

1. Competition: For each sample, the neurons in the network compete with each other. The neuron that best represents the sample (most similar) based on a discriminant function is declared the winner.
2. Cooperation: The neighboring neurons of the winning neuron determine the neighborhood. These excited neurons learn (not equally) from the same entry.
3. Adaptation: The weight vectors of the winner neuron and its neighboring neurons are updated taking into account the learning rate and the distance to the winner neuron. Through this process neighborhood nodes become more similar to the entry sample.

**T-Distributed Stochastic Neighbor Embedding (t-SNE).** This is a nonlinear dimensionality reduction method. It aims to represent $h$ dimensional points $x_i$ with $l$ dimensional points $y_i$ such that $l < h$ and the probability that $y_i$ and $y_j$ are close implies that $x_i$ and $x_j$ are close is maximized. To review the process to do so, let us assume that we have a set of $n$ $h$-dimensional vectors $\{x_1, ..., x_n\}$. Then, for $i \neq j$, we compute

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/2\sigma_i^2)},$$

(9)

and define $p_{ii} = 0$. The similarity between $x_i$ and $x_j$ will be given the conditional probability $p_{ij}$ that $x_j$ is a neighbor of $x_i$, if neighborhood is proportional to the probability density under a Gaussian centered at $x_i$, see [18]. We also define

$$p_{ij} = \frac{p_{ij} + p_{ji}}{2n}.$$  

(10)

The goal of t-SNE is to obtain a function $x_i \mapsto y_i \in \mathbb{R}^l$ preserving the similarities $p_{ij}$. For this purpose and in an analogous way, the similarity between $y_i$ and $y_j$ is computed:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_i - y_k\|^2)^{-1}},$$

(11)

and $q_{ii} = 0$. The values of each $y_i$ are obtained by minimizing the Kullback–Leibler divergence of the distribution $P$ from the distribution $Q$, i.e.,

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$  

(12)

This minimization process is done by gradient descent.

**Rand Index.** This is a method to measure the similarity between clustering methods, let us describe it in its most simple form. Let $X = \{x_1, ..., x_n\}$ and two partitions of $X$, $B = \{B_1, ..., B_r\}$, $C = \{C_1, ..., C_s\}$, we set:

- $a$, the number of pairs of elements from $X$ that belong to $B_i$ and $C_j$ for some $i, j$.
- $b$, the number of pairs of elements from $X$ that belong to different elements of $B$ and $C$.
- $c$, the number of pairs of elements from $X$ that belong to $B_i$ for some $i$ and belong to different elements of $C$. 
Identifying Similar Groups of Countries According to the Impact

- \( d \), the number of pairs of elements from \( X \) that belong to \( C_i \) for some \( i \) and belong to different elements of \( B \).

The Rand index is computed by

\[
R = \frac{a + b}{a + b + c + d}. \tag{13}
\]

Here, \( a + b \) can be seen as the agreements between \( B \) and \( C \) and \( c + d \) can be seen as the number of disagreements between \( B \) and \( C \). \( R \) can take values between 0 and 1, with 0, meaning a complete dissimilarity between clusters and 1, indicating that the clusters are the same, see [14].

3 Results and Discussion

This section presents the main results derived from the research. Figure 1 shows the experimental approach consisting of two clustering steps applied to the data. The second clustering step is performed in the data obtained from the first clustering.

3.1 First Clustering

We performed this clustering step on the dataset obtained after joining and cleansing of the raw data. Five different clustering methods are applied to this dataset: affinity propagation, agglomerative, birch, K-means, and minibatch K-means. The number of desired clusters, in which the data will be clustered, is passed as an argument to all the methods except for affinity propagation, which identify the number of clusters on its own.

After applying all the methods, we obtained a new dataset that contains the countries and the group they belong to in each method. Table 3 presents the structure of the new dataset, and Table 2 shows the representative country for each cluster in all the clustering methods. Table 4 compares all the clustering methods using Adjusted Rand Index as a comparison measure.

3.2 Second Clustering

In this part of the research, we used two different methods, t-distributed Stochastic Neighbor Embedding (t-SNE) and Self-Organizing Map (SOM). This clustering step is applied to the data obtained from the previous step. We obtained a categorical dataset with five columns indicating the cluster to which each country belongs in every clustering method; before working with this dataset it is important to encode the categorical variables. For this purpose, we used one hot coding technique; in this technique, a single categorical column with \( d \) distinct values is transformed to \( d \) binary columns, each observation indicating if the sample belongs(1) or not(0) to a certain category, see e.g., [12]. The categorical
Table 2. Representative countries for each clustering method

| Cluster | Affinity | Agglomerative | Birch  | K-Means | MB K-means |
|---------|----------|---------------|--------|---------|------------|
| 0       | Bangladesh | Philippines | Russia | Chile   | Ethiopia   |
| 1       | Brazil    | Congo (Brazzaville) | Armenia | India   | China      |
| 2       | China     | Brazil        | Yemen   | Brazil   | Hungary    |
| 3       | France    | Burma         | Philippines | Turkey | Bhutan    |
| 4       | Honduras  | Saudi Arabia  | Turkey  | US      | Italy      |
| 5       | India     | Turkey        | Iraq    | Lithuania | Saudi Arabia |
| 6       | Indonesia | Sri Lanka     | Zambia  | Uganda   | Brazil     |
| 7       | Lithuania | Uganda        | Brazil  | Russia   | Indonesia  |
| 8       | Mali      | Russia        | Burma   | China    | India      |
| 9       | Japan     | US            | China   | Indonesia | Georgia   |
| 10      | Peru      | Japan         | India   | Philippines | Congo (Kinshasa) |
| 11      | Philippines | Bangladesh | Indonesia | Mozambique | Zambia     |
| 12      | Russia    | India         | US      | Italy    | Bangladesh |
| 13      | Spain     | Indonesia     | Japan   | Hungary  | Spain      |
| 14      | Turkey    | United Kingdom | Portugal | Japan    | Madagascar |
| 15      | US        | China         | United Kingdom | Bangladesh | Russia     |

Table 3. First rows of data 2

| Country      | Affinity | Agglomerative | Birch | K-Means | MB K-means |
|--------------|----------|---------------|-------|---------|------------|
| Afghanistan  | 10       | 4             | 5     | 6       | 5          |
| Albania      | 7        | 1             | 1     | 5       | 9          |
| Algeria      | 13       | 7             | 5     | 6       | 13         |
| ...          | ...      | ...           | ...   | ...     | ...        |

Table 4. Comparison of first clustering methods using Adjusted Rand Index

|               | Affinity | Agglomerative | BIRCH   | K-means | MB K-means |
|---------------|----------|---------------|---------|---------|------------|
| Affinity      | 1.0      | 0.59          | 0.81    | 0.96    | 0.7        |
| Agglomerative | 1.0      | 0.61          | 0.57    | 0.39    |            |
| BIRCH         | 1.0      | 0.84          |         |         | 0.59       |
| K-means       | 1.0      |               |         |         | 0.72       |
| MB K-means    |          |               |         |         | 1.0        |

dataset of shape $163 \times 5$ is transformed into a binary dataset of shape $163 \times 80$ as every column of the categorical dataset has 16 different groups.

To test the proposed technique, we also applied the second clustering to the original dataset used in the first clustering to see the differences in the obtained results.
**T-SNE Results.** First, we applied t-SNE to both datasets with a dimensionality reduction purpose. After obtaining the new two dimensions representing each dataset, a K-means clustering algorithm is applied to cluster the data in 10 different groups. Figure 2 presents the results obtained for the original dataset, while Fig. 3 presents the results obtained with the proposed technique of applying some clustering methods first and transforming the dataset (two-layer clustering).

![Clusters after applying t-SNE method to Data 1 (1 layer t-SNE+ K-means)](image)

**Fig. 2.** Clusters after applying t-SNE method to Data 1 (1 layer t-SNE+ K-means)

It is easier to visually distinguish some groups with the proposed method. Some samples are already grouped after applying t-SNE, making it easier for K-means to distinguish each cluster. The separation of the samples into groups is better, or at least the clusters are easier to identify with the proposed methodology, as shown in Fig. 3 in comparison with Fig. 2. To test this, we compute the mean variance of the distance from the centroid of the cluster to each sample within the cluster and the average distance among all the centroids in each approach. Table 5 presents the results of the comparison. A smaller variance is preferred, while a greater distance among centroids is desired.

When using t-SNE, the approach presented in this research has shown a good performance as it minimizes the variance of the distance from each sample to their corresponding centroid, yet the centroids are closer to each other in this approach. However, comparing one layer clustering t-SNE+K-means with two layers clustering t-SNE+K-means, we obtain an Adjusted Rand Index of 0.6154921500297419. This suggests that even though some observations are classified in similar clusters for both methods, there is still a great proportion of observations classified in a different way, which is congruent with the compactness of some clusters with our proposed method, as shown in Fig. 3.

**SOM Results:** We applied this method to both datasets, again with a dimensionality reduction purpose. After obtaining the new two dimensions that represent each dataset, we performed a K-means clustering algorithm to cluster the data in 10 different groups. Figure 4 presents the results obtained for the original dataset, while Fig. 5 presents the results obtained with the proposed technique of applying some clustering methods first and transforming the dataset.
This time the comparison of both approaches is easier to spot. When applied to the original dataset, SOM yields to a 2D map that does not have clusters as distinguishable as the one generated with the proposed technique. To test this, we compute the mean variance of the distance from the centroid of the cluster to each sample within the cluster, as well as the average distance among all the centroids in each approach. Table 6 presents the results of the comparison. It is important to notice that Fig. 5 has fewer plotted points. The reason for this is that many samples have the same coordinates. A single point in the 2D map may represent several samples, making it easier for K-means to cluster the data obtained with SOM as one point may cluster several countries with less disperse samples on the map. Fewer points are desirable as it shows the samples are already grouped by similarity.
When using SOM, the approach presented in this research minimized the variance of the distance from each sample to their corresponding centroid, yet the centroids are slightly closer to each other in this approach. Comparing between clusters from 1 layer clustering SOM+K-means and 2 layer clustering SOM+K-means, we get an Adjusted Rand Index of 0.2395; as can be seen in Figs. 4 and 5 this is an indicator that both methods perform clustering in a completely different way.

### 3.3 Final Results

We obtained the best results using the proposed method and SOM (Fig. 5), in the sense that samples were tightly grouped together, resulting in the smallest mean variance of distance sample-centroid of the four studied methods. Table 7 shows the clusters formed with this approach. The countries are expected to take actions similar to other countries within their groups to fight coronavirus infection and spreading.
Table 7. Groups formed with the two-layer SOM clustering method

| Group | Countries                                                                 |
|-------|---------------------------------------------------------------------------|
| 1     | Algeria, Argentina, Congo (Kinshasa), Germany, India, Iran, Iraq, Japan,   |
|       | Mexico, Russia, Spain, Turkey, US, Uganda, Ukraine                         |
| 2     | Albania, Armenia, Bosnia and Herzegovina, Botswana, Central African        |
|       | Republic, Congo (Brazzaville), Costa Rica, Croatia, Denmark, Eritrea,      |
|       | Finland, Gabon, Georgia, Guinea-Bissau, Ireland, Jamaica, Kuwait, Lesotho, |
|       | Liberia, Lithuania, Mauritania, Moldova, Mongolia, Namibia, New Zealand,   |
|       | Norway, Oman, Panama, Qatar, Singapore, Slovakia, Slovenia, Uruguay       |
| 3     | Australia, Cameroon, Cote d’Ivoire, Ghana, Madagascar, Mozambique, Nepal,  |
|       | Yemen                                                                      |
| 4     | Austria, Azerbaijan, Belarus, Belgium, Benin, Bolivia, Bulgaria, Burundi,   |
|       | Czechia, Dominican Republic, El Salvador, Greece, Guinea, Haiti, Honduras, |
|       | Hungary, Israel, Jordan, Lebanon, Libya, Nicaragua, Papua New Guinea,      |
|       | Paraguay, Portugal, Rwanda, Serbia, Sierra Leone, South Sudan, Sweden,     |
|       | Switzerland, Tajikistan, Togo, Tunisia, United Arab Emirates               |
| 5     | Antigua and Barbuda, Bahrain, Barbados, Bhutan, Comoros, Cyprus, Djibouti,|
|       | Dominica, Equatorial Guinea, Estonia, Fiji, Grenada, Guyana, Iceland,     |
|       | Latvia, Luxembourg, Maldives, Malta, Mauritius, Montenegro, San Marino,    |
|       | Seychelles, Suriname, Timor-Leste, Trinidad and Tobago                     |
| 6     | Burkina Faso, Cambodia, Chad, Chile, Ecuador, Guatemala, Kazakhstan,       |
|       | Malawi, Mali, Netherlands, Romania, Senegal, Somalia, Sri Lanka, Zambia,   |
|       | Zimbabwe                                                                   |
| 7     | Bangladesh, Colombia, Egypt, Ethiopia, France, Indonesia, Philippines,     |
|       | South Korea, Thailand, United Kingdom, Vietnam                             |
| 8     | Afghanistan, Angola, Brazil, Canada, Malaysia, Morocco, Nigeria, Pakistan,|
|       | Peru, Poland, Saudi Arabia, Uzbekistan                                     |
| 9     | China, Niger, Taiwan                                                       |
| 10    | Burma, Italy, Kenya, South Africa, Tanzania                                |

4 Conclusions and Recommendations

This work presented a novel clustering method. This methodology consists in transforming the data using several clustering methods followed by a new clustering with dimensionality reduction using t-SNE and SOM over the new data to group the data and later apply K-Means to identify the clusters. According to the minimum mean variance of distance sample-centroid, the best results were found using SOM with the proposed method of two-layer clustering. It also showed good performance when considering the mean distance among centroids, even though it was not the best one in this category. This method showed a reduced amount of plotted points in the graph, as it groups the countries
based on the first clustering similarities. After applying the dimensionality reduction, several countries have the same $x$ and $y$ coordinates, making it easier for K-Means algorithm to make groups. This technique is applied to COVID-19 data collected from 163 countries to group countries into clusters according to their similarities. With this information, countries can apply similar effective measures that worked in a particular country in the same cluster or, on the contrary, avoid applying measures that have shown poor performance in a given country within the same cluster.

The proposed method can be modified for future work to add different clustering methods in the first or second clustering layers. It is important to consider methods that work well with the selected dataset. To further study the influence of the variables in the clusters, and the contrast between the latter, a MANOVA Biplot could be performed [1]. The dataset can be expanded to consider more variables that may have significantly influenced the impact of COVID-19. For example, considering the public policies applied in the countries to stop the advance of the disease, as this information is directly related to the spreading of COVID-19. Another work that can be performed in the future is to compare the efficiency of the proposed method with other clustering algorithms.

References

1. Amaro, I.R., Vicente-Villardón, J.L., Galindo-Villardón, M.P.: Manova biplot for two way multivariate general linear models. Interciencia 29(1), 26 (2004)
2. Asan, U., Ercan, S.: An introduction to self-organizing maps. Computational Intelligence Systems in Industrial Engineering, pp. 295–315. Springer (2012)
3. Carozzi, F., Provenzano, S., Roth, S.: Urban Density and COVID-19. Discussion Paper Series, IZA Institute of Labor Economics. http://ftp.iza.org/dp13440.pdf
4. European Centre for Disease Prevention and Control: Q & A on COVID-19. European Union. (2020). https://www.ecdc.europa.eu/en/covid-19
5. Frey, B.J., Delbert, D.: Clustering by passing messages between data points. Science. Am. Assoc. Adv. Sci. 315(5814), 972–976 (2007)
6. Hamidi, S., Sabouri, S., Ewing, R.: Does density aggravate the COVID-19 pandemic? J. Am. Plann. Assoc. 86(4), 495–509 (2020). https://doi.org/10.1080/01944363.2020.1777891
7. Haykin, S.: Neural Networks: a Comprehensive Foundation. Prentice-Hall, Inc., Upper Saddle River (2007)
8. Inaba, M., Katoh, N., Imai, H.: Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. In: Proceedings of 10th ACM Symposium on Computational Geometry, pp. 332–339 (1994) https://doi.org/10.1145/177424.178042
9. Kohonen, T.: The self-organizing map. In: Proceedings of the IEEE, vol. 78, no. 9, pp. 1464–1480. IEEE (1990)
10. Sanjay, K.: Monitoring Novel Corona Virus (COVID–19) Infections in India by Cluster Analysis. Annals Data Sci. 7(3), 417–442 (2020)
11. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. JMLR 12, 2825–2830 (2011)
12. Potdar, K., Pardawala, T.S., Pai, C.D.: A comparative study of categorical variable encoding techniques for neural network classifiers. Int. J. Comput. Appl. 175(4), 7–9 (2017)
13. Prabhu, T.N.: Population by Country - 2020. Dataset. https://www.kaggle.com/tanuprabhu/population-by-country-2020/metadata
14. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association. American Statistical Association 66(336), 846–850 (1971) arXiv:1704.01036. https://doi.org/10.2307/2284239 JSTOR 2284239
15. Rencher, A.C.: Methods of Multivariate Analysis. vol. 492, Wiley, New York (2003)
16. The Humanitarian Data Exchange, service provided by the OCHA. https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases
17. The pandas development team: pandas-dev/pandas: Pandas. Zenodo. Version 1.0.5. (2020). https://doi.org/10.5281/zenodo.3509134
18. van der Maaten, L.J.P., Hinton, G.E.: Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008)
19. World Population Review: GDP Ranked by Country 2020. https://worldpopulationreview.com/countries/countries-by-gdp
20. World Economic Outlook Database. International Monetary Fund. April, 2020. https://www.imf.org/external/pubs/ft/weo/2020/01/weodata/index.aspx