Avrupa Bilim ve Teknoloji Dergisi
Sayı 41, S. 295-306, Kasım 2022

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 41, pp. 295-306, November 2022

Research Article

Determination of Worldwide Country Clusters by Selecting the Best Machine Learning Algorithm via MULTIMOORA for Covid-19 Pandemic

Sevgi Abdalla[1*], Ozlem Alpu[2]

[1*] Eskisehir Osmangazi University, Faculty of Science, Department of Statistics, Eskisehir, Türkiye, (ORCID: 0000-0003-4177-5868), sayhan@ogu.edu.tr
[2] Eskisehir Osmangazi University, Faculty of Science, Department of Statistics, Eskisehir, Türkiye (ORCID: 0000-0002-2302-2953), oalpu@ogu.edu.tr

(İlk Geliş Tarihi 30 Mayıs 2022 ve Kabul Tarihi 21 Ağustos 2022)

(**DOI:** 10.31590/ejosat.1123516)

ATIF/REFERENCE: Abdalla, S., Alpu, O. (2022). Determination of Worldwide Country Clusters by Selecting the Best Machine Learning Algorithm via MULTIMOORA for Covid-19 Pandemic. *Avrupa Bilim ve Teknoloji Dergisi*, (41), 295-306.

Abstract

In this study, to present an integrated approach to clustering analysis based on multi-objective decision making, it is aimed to determine the best clustering algorithm among 11 different clustering algorithms by evaluating all 27 internal validity criteria simultaneously with MULTIMOORA method. In the study, initially, the best clustering algorithm was determined according to the optimal number of clusters for two COVID-19 datasets. Then, it focuses on determining the relationship of the country clusters with the classes determined according to the human development index. In the result of the analyses, countries affected by the COVID-19 pandemic have clustered via the CLARA and SOM algorithms according to their proximity calculated from the Euclidean distance. Three optimal number of clusters were determined for both datasets. The incidence rate variable is the more dominant factor than case fatality rate in the real difference between clusters. Another remarkable finding is that while countries with economic power and a high level of human development are expected to be less affected by the pandemic before the vaccination, the level of being affected by the pandemic increases in terms of both variables as the level of human development increases.

Keywords: Machine learning, Clustering, Internal validation criteria, MULTIMOORA, COVID-19.

MULTIMOORA ile En İyi Makine Öğrenimi Algoritmasını Seçimi ve Covid-19 Pandemisi için Dünya Çapında Ülke Kümelerinin Belirlenmesi

Öz

Bu çalışmada, çok amaçlı karar vermeye dayalı kümeleme analizine entegre bir yaklaşım sunmak amacıyla, 27 iç geçerlilik kriterinin tamamı MULTIMOORA yöntemi ile eş zamanlı olarak değerlendirilerek 11 farklı kümeleme algoritması arasından en iyi kümeleme algoritmasının belirlenmesi amaçlanmıştır. Çalışmada öncelikle iki veri kümesi için en uygun küme sayısı ve bu küme sayısına bağlı olarak en iyi kümeleme algoritması belirlenmiştir. Daha sonra, belirlenen ülke kümelerinin insani gelişmişlik sınıflarıyla ilişkisinin belirlenmesine odaklanılmıştır. Yapılan analizler sonucunda COVID-19 salgınından etkilenen ülkeler, Öklid uzaklığı aracılığıyla hesaplanan yakınlıklarına göre CLARA ve SOM algoritmaları ile kümelenmiştir. Her iki veri kümesi için de en uygun küme sayısı olarak üç küme belirlenmiştir. Vaka-ölüm oranına kıyasla insidans oranının kümeler arasındaki gerçek farkta daha baskın faktör olduğu bulunmuştur. Bir diğer dikkat çekici bulgu ise, ekonomik gücü ve insani gelişmişlik düzeyi yüksek ülkelerin, aşılama öncesinde pandemiden daha az etkilenmesi beklenirken, insani gelişmişlik düzeyi yüksek olan ülkelerin pandemiden etkilenme düzeyinin her değişken bakımından da yüksek olmasıdır.

Anahtar Kelimeler: Makine Öğrenimi, Kümeleme, İç geçerlilik kriterleri, MULTIMOORA, COVID-19,

[*] Sorumlu Yazar: sayhan@ogu.edu.tr

1. Introduction

Since December 2019, the world has been faced with the COVID -19 pandemic, which is perhaps one of the biggest health crises of the last century. Researchers around the world are making efforts to compare and understand the spread of the COVID-19 pandemic by country (Khafaeie &Rahim, 2020; Wu et al., 2020; Chu, 2021; Hasell et al., 2020; Li et al., 2021; Kuster & Overgaard, 2021; McKenzie & Adams, 2020; Yuan et al., 2021; Harapan et al., 2020).

Additionally, the COVID-19 pandemic has shown that the ability of countries, societies, and individuals to respond quickly to such crises is dramatically low and not evenly distributed. The unequal conditions of countries are related to their economic and socio-cultural power, besides many other reasons. In some studies, the effects of the economic and socio-cultural power of countries' ability to cope with the COVID-19 pandemic have been revealed (Li et al., 2021; Siddik, 2020; Sharma et al., 2021; Liu et al., 2020; Shahbazi &Khazaei, 2020; Marziali et al., 2021; Gokmen et al., 2021; Rocha et al., 2021; Hezam, 2021; Asem et al., 2021).

As a result, while the impact of the epidemic on the economy is felt both in countries with well-developed economies and in developing countries, it has been revealed that pre-existing inequalities have increased in many areas, especially in the fields of health, education, and employment. In addition to these inequalities, clustering countries based on the incidence and case fatality rates announced for COVID-19 is of vital importance to create a geographical risk assessment. It is obvious that the results of cluster analysis are of undeniable importance for researchers and commentators of different disciplines from all walks of society in determining strategies to suppress the epidemic. In clustering analysis, the use of case fatality rates (CFR) and incidence rates (IR) as key indicators of disease characteristics is important for the comparison of these indicators between countries, determining national and international priorities, and recognition of health system performance. It is seen that the pre-vaccination pandemic had a greater effect on countries with a large economy and high human development index, with the degree of impact varying from country to country. In this respect, the following question is suggested: 'Can the levels of human development of countries be a measure of success/failure in the combat against the COVID-19 pandemic?' Based on this question, four main objectives were determined in the study. These objectives can be listed as follows:

(1) To categorize countries in terms of similarities according to their level of being impacted by the COVID-19 pandemic before vaccination,

(2) To reveal if there is a relationship between the categories of countries in terms of human development index and levels of being impacted by the COVID-19 epidemic; and to determine the extent of its relationship.

Furthermore, additional objectives have been defined to determine the course of the post-vaccination pandemic and whether there is a difference between the categories of countries pre- and post-vaccination. For these purposes, CFR and IR variables as well as the ratio of fully vaccinated people relative to population (VAC-R) are included in the evaluation of the post-vaccination process. Thus, additional objectives are defined as follows when the post-vaccination pandemic data are considered in the study.

(3) To reveal similarities/dissimilarities between the groups of countries pre- and post- vaccination and,

(4) To reveal similarities/dissimilarities between the level of human development and the groups of countries post-vaccination.

However, the classification of countries with similar structures in terms of incidence, case fatality rate, and people fully vaccinated rate in the same clusters, and thus the performance of clustering results (although these changes depending on time), is possible by determining the optimal clustering algorithm, the optimal number of clusters and the appropriate criteria. For this reason, it is not possible to mention the optimal solution for the clustering problem that is analyzed with a single randomly selected and applied algorithm and validation criterion.

For all these objectives, 11 clustering algorithms, which are considered successful in the literature, were used. Among the algorithms considered, determining the optimal clustering algorithm and the number of clusters is included in the class of multi-objective decision-making (MODM) problems. There are many MODM techniques developed for the solution of such problems in the literature. The MULTIMOORA method has been used due to the numerical values of the selection criteria determined in the solution of this problem and the effectiveness of the algorithm. Thus, to cluster countries according to their level of impact from the COVID-19 pandemic, various algorithms have been compared according to different criteria and the optimal clustering algorithm, and, accordingly, the number of clusters has been determined.

Finally, in this study, the relationship between the clustering results and the categories according to the human development index of the countries has been revealed. Therefore, this study contributes to the literature both regarding the proposed solution approach and in terms of demonstrating the relationship of clustered countries with the level of human development.

2. Material and Method

2.1. Clustering Methods

The algorithms used in the study are included in the following sub-sections.

2.2.1. K-means Algorithm

The k-means algorithm is one of the partitioning cluster algorithms that require predetermining the number of clusters to be created. Thus, it is aimed to minimize the sum of squares within the group for the specified number of clusters (Hartigan and Wong, 1979). Although it is an advantage that this algorithm is easy to implement and can be applied in large data sets, it has the disadvantage that if the initial number of clusters cannot be determined, the clusters obtained depend on the selection method of the first cluster centroids. Also, it is sensitive to outliers, cannot be used in categorical data, and is not suitable for finding non-convex clusters (Berkhin, 2002).

The k-means algorithm is a poor choice for clustering unless special conditions (i.e., spherical cluster, no outliers, and properly initializing, etc.) are met on the data (Estivill-Castro and Yang, 2000). It can also converge to the local maximum at low quality

(Bradley et al., 1997; Fraley and Raftery,1998). It is also stated that since it is sensitive to transformations such as scaling and is statistically biased, it can converge to incorrect parameter values (Estivill-Castro and Yang, 2000).

2.2.2. Partitioning Around Medoids (PAM)

In this algorithm, the number of clusters must be predetermined by the researcher, and k initial cluster centroids are required to start the algorithm. PAM is considered as more robust because it allows the use of the L_2 norm as well as other measures of dissimilarity. The algorithm gives successful results in small datasets. However, due to its computational complexity, its performance is poor on massive datasets. Another disadvantage is that the researcher has to determine the number of clusters (Han et al., 2012).

2.1.3. Clustering Large Applications (CLARA)

Since the PAM algorithm cannot show the desired sensitivity in large data sets, the use of the CLARA algorithm developed by (Kaufman and Rousseeuw, 2005) in large data sets seems to be the biggest advantage. However, choosing the right sample size provides good results from the algorithm. Since the algorithm determines the cluster centers as k-medoids amongst these samples, if an observation with k-medoids is not selected for sampling, the algorithm cannot make a correct clustering. Because the algorithm determines cluster centers as k-medoids from these samples, the algorithm cannot make an accurate clustering if the sampling of a medoid that is a k-medoid is not selected. Since CLARA finds the best k-medoids in the sample from the dataset, then it cannot perform a good clustering if any of the best sampled medoids are away from the best k-medoids (Han et al., 2012)

2.1.4. Divisive Analysis (DIANA)

Diana algorithm belongs to the class of hierarchical clustering algorithms. It begins with the clustering process in a single cluster included all observations, and repeatedly classifies until each cluster includes only one observation (Kaufman and Rousseeuw, 2005). It can be applied to datasets where distance measures, which are similarity measures, are used. Additionally, DIANA is not affected by the initial selection of the cluster center; in other words, it always provides a unique cluster. Determining how to classify from a large cluster into smaller clusters is a drawback of divisive methods. Examining the probabilities of how many different ways a dataset will be partitioned takes computational time. In short, since the algorithm runs heuristic methods for partitioning, it can lead to erroneous clustering. In addition, due to the computational workload, DIANA does not optimize by backtracking the partitioning within the algorithm. Because of the high computational load, it is not preferred to be used in large datasets.

2.1.5. Fuzzy Analysis (FANNY)

This algorithm implements fuzzy clustering, where each cluster contains observations with partial membership. Therefore, a vector is defined that indicates the partial membership of each observation in each cluster. The algorithm estimates the membership function to minimize the objective function. Observations are grouped into clusters with high membership. This algorithm requires the use of datasets with at least interval-scaled variables or a dissimilarity matrix. The algorithm does not allow the occurrence of one cluster; it provides fuzzy sets to the researcher to assign observations to more than one cluster. In addition, due to the numerical reasons in the initial step and the difficulty of interpreting large k values, the algorithm allows the utmost number of clusters to be half (n/2) of the maximal number of observations (Kaufman and Rousseeuw, 2005).

According to existing fuzzy clustering algorithms, the advantages of the algorithm are that it uses the dissimilarity matrix, is more robust for the spherical cluster assumption, and yields a new graphical representation (Silhouette plot) to the user (Itoh, 2013).

Fuzzy clustering presents more detailed insight into the structure of the dataset than hard clustering. However, a large number of observations and clusters, as well as the number of these outputs (the amount of detailed information), can sometimes become a drawback for using the algorithm. The lack of representative observations, the complexity of fuzzy clustering algorithms, and the long computation time can be considered as other disadvantages. However, it is considered attractive by researchers because of the fuzziness property, which allows using the uncertainties of real-life data (Kaufman and Rousseeuw, 2005).

2.1.6. Self-organizing Map (SOM)

This algorithm includes an unsupervised learning based on artificial neural networks for dimension reduction and data clustering. The SOM algorithm, also known as Kohonen Networks, has been used as a classification technique in a variety of areas with great success (Kiang, 2001). Besides being used for clustering purposes, it is a data visualization tool (Flexer, 2001).

One of the most important reasons why neural network-based techniques such as the SOM method are preferred to statistical modelling techniques is that they do not necessitate any assumptions regarding the data distribution. In addition, unlike some statistical clustering methods, the SOM method does not require assumptions about the initial number of clusters, the variables distributions, and the independence between the variables. The SOM method has started to be used frequently in cluster analysis studies due to its usefulness and flexibility against statistical assumptions. Furthermore, SOM presents a map with a lower-dimensional representation of higher-dimensional data and displays clusters on this map (Dunham, 2003)

SOM is also considered an effective algorithm when working with high dimensional data.

The one drawback of the SOM method is that it cannot provide a measure of the validity of cluster analysis results. In addition, since SOM cannot provide the features/variables that will enable clusters to be distinguished from each other, it would be more meaningful to use it together with a rule set such as C5.0.

2.1.7. Self-organizing Tree Algorithm (SOTA)

SOTA is a hybrid neural network algorithm developed by (Dopazo, and Carazo, 1997) combining the advantages of SOM and hierarchical clustering. The algorithm has a structure which is based on a divisive hierarchical binary tree. It uses a fast algorithm and is, therefore, suitable for clustering datasets with a large number of observations. It is considered an advantage to combine hierarchical clustering with SOM. The algorithm provides a mapping by reducing complex data sets. SOTA allows a built-in assessment of the reliability of any cluster in the entire hierarchy in the analysis. Producing dendrograms that explain the clustering at different hierarchical levels allows the management

of the resolution. Additionally, it is robust to outliers. The fact that SOTA has a hierarchical tree structure and that clusters are obtained proportionally to the heterogeneity of the data are considered as two important advantages (Herrero et al., 2001).

2.1.8. Agglomerative Nesting (AGNES)

AGNES, one of the hierarchical clustering algorithms, has a bottom-to-top clustering structure. Each observation is considered as a cluster at the initial stage, and the most similar pairs among them are clustered. This process continues until there is no more observation to the cluster.

This algorithm has the advantages of not requiring predetermining the number of clusters and being easy to use. The sensitivity of the distance matrix to outliers, the splitting of large clusters, and the inability to cope with convex and different size clusters can be considered as the disadvantages of the algorithm. In addition, the computational time may be longer depending on the size of the dataset. Determining the correct number of clusters with a dendrogram can be difficult.

2.1.9. Hierarchical Clustering Algorithm

This algorithm is preferred if the number of clusters is not predetermined or decided. In the algorithm, the visual perceptibility of the clusters is high because the observations or variables that are close to each other are grouped in terms of proximity or distance measures. It gives much better results in data sets containing values that are characterized as outliers, extremes, and noisy values, compared to partitional clustering methods. It can be applied to all types of data and gives more accurate results than partitional clustering methods. In general, the disadvantage of hierarchical clustering algorithms is that the number of operations to be performed increases as the number of observations or variables increases. Therefore, it is a very time-consuming algorithm in large data sets. However, this disadvantage can be reduced by computers with more powerful processing capabilities. Another drawback of the method is that hierarchical clustering algorithms do not have a precise criterion to determine at which stage of the solution to stop, and an observation cannot be reassigned to another cluster after it is included in a cluster (Aydın and Seven, 2015).

2.1.10. EM Algorithm

The EM algorithm uses the model-based soft clustering method (Dempster et al., 1977) For this reason, the clusters formed are generally not in a disconnected form from each other. The EM algorithm prefers to use an estimation method instead of using distance measures to determine which cluster an observation will be in. The fact that the likelihood increases with each iteration, both steps (E and M) are easy to implement, and the solution of the maximization (M) step has a closed form are the advantages of the EM algorithm (Couvreur, 1997, Gupta, and Chen, 2011). Very slow convergence of the algorithm, convergence to the local optimal, and the use of forward-backward probabilities are the disadvantages of the EM algorithm (McLachlan et al., 2004).

2.1.11. Genie Algorithm

The algorithm Genie is a multi-objective clustering algorithm, as effective and simple as any other distance-based hierarchical clustering algorithm (Gagolewski et al., 2016). It only requires a measure of similarity between a pair of observations. Partitioning of a dataset can be carried out with various structure of data such as intensive, sparse, and string. It can also be used with various metrics such as L_2 norm, L_1 norm, cosine, and Levenshtein. It also allows the number of clusters to be predetermined. It always returns exactly the desired number of clusters. The algorithm provides coherent clustering results against small changes in the Gini threshold.

The time complexity of a clustering algorithm mostly depends on size of data which can increase the number of dissimilarity computations. Such a restriction makes it disadvantageous to use all classical linkage criteria for larger datasets, except for the single linkage criterion. However, it is known that the results of single linkage method are affected by outliers and therefore will not always produce an accurate clustering structure. To cope with these drawbacks, the Genie algorithm has been suggested.

According to the comparisons made by (McLachlan et al., 2004) this algorithm is shown to be quite practical and useful. While single linkage gives fast results, they state that it mostly outperforms such methods as Ward and average linkage in terms of quality of clustering. Furthermore, the algorithm can be easily parallelized and accelerated further. Additionally, there is no requirement to pre-compute the complete distance matrix to accomplish clustering.

2.3. Clustering Validation Measures

The performance scores of clustering algorithms are associated with validity measures, which represent the quality of a clustering algorithm that creates convenient clusters without any information about the cluster [Rendón et al., 2011, Pérez et al., 2020]. In the literature, cluster validity measures, mostly called validation indices are grouped as internal, external, and relative measures (Wani, and Riyaz, 2016). The internal type of validation measures employs the information by processing the observed unlabeled original data (Moshtaghi et al., 2019). In contrast, external validity indices use prior information such as pre-defined class-labelled data to extract information about the clustering structure [Rendón et al., 2011]. The indices achieve the scores by comparing a cluster to a given partition. Finally, relative validation indices consider the structure of clustering by comparing different conditions (e.g., parameter values, i.e., varying the number of clusters) for the same algorithm (Dalton et al., 2009).

In the literature, many different clustering validation measures exist (Milligan, 1981, Bolshakova, and Azuaje, 2003). Desgraupes (2012) brought together a total of 42 validation measures, internal, external, and relative, in the clusterCrit package he developed for the R software and gave detailed information about the mathematical structure of the measures.

The quality or performance of a clustering analysis and a clustering algorithm is often examined using measures of internal validation when not to know the number of clusters in advance, or the cluster labels are not known (Van Craenendonck and Blockeel, 2015) While some of these developed measures should be minimized (e.g., connectivity, Ball-Hall, etc.), some of them should be maximized (silhouette, etc.). Since some measures are based on the difference between two measures, then, the difference should be minimized or maximized (Desgraupes, 2012). However, choosing the right one among so many validation measures is a difficult decision for the researcher.

Since many studies use a single clustering technique or algorithm, such as k-means (Kurniawan et al., 2020), the choice of validity measures has not been given due attention. If more than one algorithm is used, it is seen that the internal validity measures of silhouette (Kucukefe, 2020, António et al., 2021], and within-cluster sum of squares (Virgantari, and Faridhan, 2020) are often taken into account. However, to the best of our knowledge, no clustering study or program has been encountered that recommends the best algorithm by considering all the criteria.

In this study, a common consensus-based solution was aimed at by choosing the best clustering algorithm for the COVID-19 dataset by considering all of the internal validity criteria. For this purpose, 27 internal validity measures included in the clusterCrit package were taken into account (see Desgraupes, 2012, p. 21) and values for all of these internal validity measures were obtained.

2.4 MULTIMOORA

The MOORA method built on the base of multi-objective methods was suggested by Brauers and Zavadskas for ranking the alternatives concerning their performances (Brauers et al., 2008). The remarkable feature of the method is to satisfy the robustness conditions (Brauers et al., 2008, Brauers and Zavadskas, 2011] which might be depicted as follows; (i) has to make allowances for all potential independent objectives, (ii) enables autonomy in consumers' preferences, (iii) handles the interrelations between objectives and alternatives in one structure at a time, not required paired comparisons, (iv) represent the performance scores by the cardinal numbers, (v) requires to be far from subjectivity, and (vi) processes the actual data.

The fact that the MOORA is the only one which satisfies all conditions mentioned above, hence, makes the method more prominent than the others. The method is a combination of two methods, such as the ratio system and reference point method (Brauers and Zavadskas, 2011). In addition, the MULTIMOORA method was developed by integrating the Full Multiplicative Form to these methods The solution process of the MULTIMOORA is given in the following:

2.4.1. The Ratio System

Step 1. Obtaining a decision matrix: The values (x_{ij}) in a decision matrix refer to performances of an alternative on the identified objectives. Here i and j represent the objectives and the alternatives, respectively.

Step 2. Normalize the values. To utilize the ratio system, the performance values are normalized by using the formula in the following.

$$x_{ij}^{*} = \sqrt{\frac{x_{ij}}{\sum_{j=1}^{m} x_{ij}^{2}}}.$$

Step 3. Optimize the ratio system by calculating the utility.

The utility is a value, that is, the difference between the utility of maximized and minimized objectives, represented as follows.

$$U(a_j) = \sum_{i=1}^{g} \left(x_{ij}^{*}\right) - \sum_{i=g+1}^{n} \left(x_{ij}^{*}\right)$$

Step 4. This step requires ranking of the utilities from the top to the bottom and maximum utility $U(a_i)$ shows the most preferable alternative.

2.4.2. The Reference Point Approach

This approach requires the normalized values of performance scores as in the ratio system method. Then, by using the Tchebycheff metric that is called min-max metric, the reference point is determined. Brauers and Ginevičius (2010) have highlighted that this metric is the most robust among all the alternative metrics of reference point theory. The reference point r_i, refers to the best utility value for the ith criterion, which has to be maximized or minimized.

$$\min(j)\left\{\max(i)\left|r_i - x_{ij}^{*}\right|\right\} \quad \text{where } r_i \begin{Bmatrix} \max x_{ij}^{*} \\ \min x_{ij}^{*} \end{Bmatrix}$$

2.4.3. The Full Multiplicative Form and MULTIMOORA

The MULTIMOORA method uses a multiplicative utility function that provides such advantages as simple application, less computational time, basic mathematical calculations, high stability, no extra assumptions, or parameters.

$$U_j \prod_{i=1}^{n} x_{ij}$$

with $j: 1,2,\ldots,m$ alternatives and $i: 1,2,\ldots,n$ objectives

U_j represents the overall utility of alternative j only if it has a one-sided objective. However, if the decisions require performing the mixed objectives, the utility function has been formulated as follows:

$$U_j^{*} = \frac{A_j}{B_j};\ A_j = \prod_{i=1}^{g} x_{ij} \text{ and } B = \prod_{i=g+1}^{n} x_{ij}.$$

A_j and B_j are the overall utility of the related objectives which are maximized and minimized. Finally, if the alternative has the best utility, the ranks are assigned by ordering those final utilities by utilizing dominance-based theory (Brauers and Zavadskas, 2011).

2.5. COVID-19 Datasets and Preprocessing

In this study, three datasets are used. One is the COVID-19 pre-vaccination dataset consisting of case fatality rate and incidence rate (per 1m) variables, provided from (OWD, 2022), and involving the collection of records from the beginning of the pandemic until the 14th of March 2021 inclusive. These two variables have been seen to play a significant role in explaining the pandemic and determining its effects in different areas (Karmakar et al., 2021; VoPham et al., 2020; Tosepu et al., 2020; Ahmad et al., 2020). The reason for taking this date as a reference is the fact that vaccination was started in economically strong countries and was carried out rapidly. Considering the possibility that vaccination may completely change the course of the pandemic based on countries and continents, the data regarding the pre-vaccination period has been analyzed to try to shed light onto the second phase of the study.

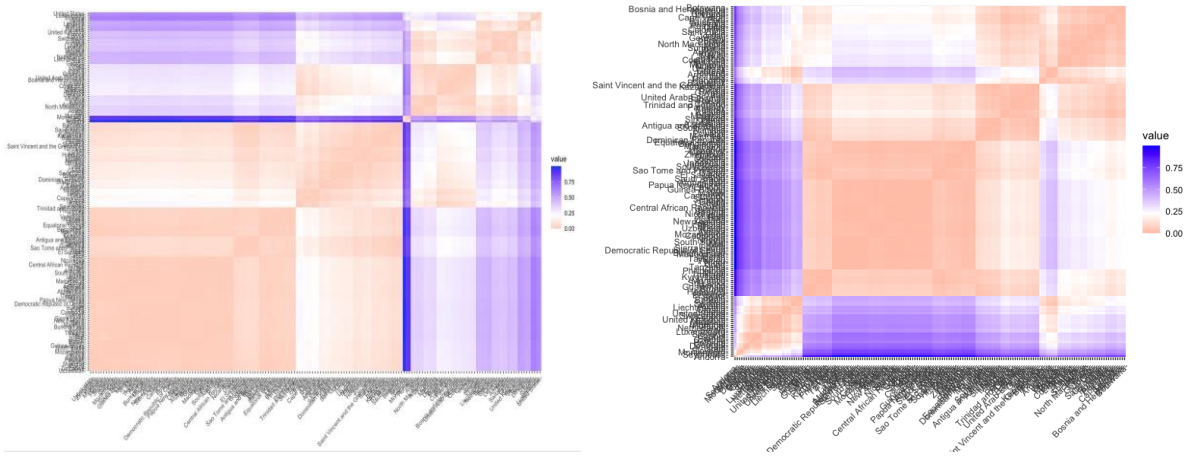
The second dataset in the study contains information about VAC-R which is added as an extra variable to the dataset (OWD, 2022). The dataset is called COVID post-vaccination dataset includes the records of countries for three variables such as case fatality rate, incidence rate, and people fully vaccinated rate from March 15th, 2021, to January 31st, 2022.

The third one is the Human Development Index dataset. This dataset is obtained from the 2020 Human Development Report (HDR, 2020). The dataset initially included 222 country records. However, since it will be a comparative study with country groups clustered according to their degree of impact from the pandemic with human development levels, country data common to both datasets have been analyzed. The final datasets include COVID-19 pandemic records for a total of 172 countries. The whole data analysis process was carried out in R-Studio version 1.4.1103 (RStudio Team, 2021) by using packages NbClust (Charrad et al., 2014), clusterCrit (Desgraupes, 2016), optCluster (Sekula et al., 2017), cluster (Maechler et al., 2021), genieclust (Gagolewski et al., 2016), and MESS (Wickham et al., 2021).

3. Results and Discussion

3.1. Determining The Optimal Numbers Of Clusters for The COVID-19 Datasets

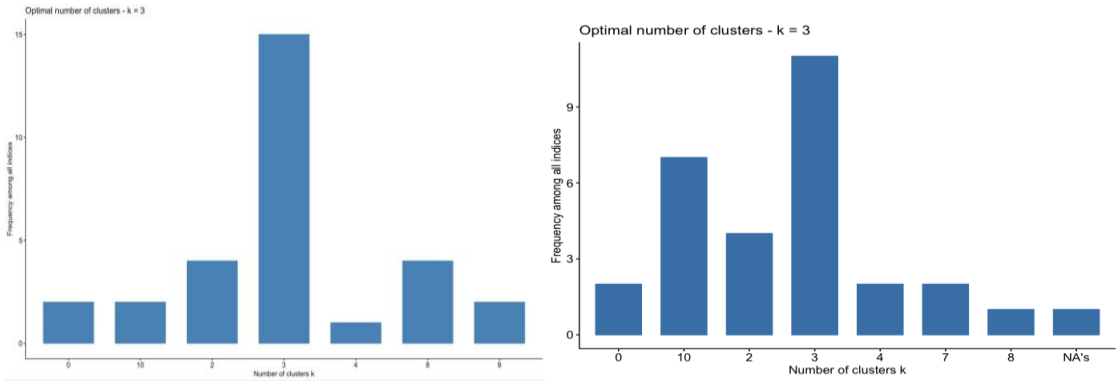
In the study, initially, the optimal number of clusters to which countries belong was determined for both datasets according to the level of impact from the COVID-19 pandemic. For this purpose, the clustering tendency of the datasets was determined by using the Hopkins statistics which shows that they are more likely to be clustered and have a strong clustering tendency. The values of the Hopkins statistics were found as 0.82 and 0.77 for the pre- and post-vaccination datasets, respectively.



(a) Pre-vaccination *(b) post-vaccination*

Figure 1. Ordered dissimilarity images (ODI) for the COVID-19 datasets

Another method that is called visual assessment tendency (VAT) introduced by (Bezdek & Hathaway, 2002), provides a reordered matrix of pairwise object dissimilarities via an intensity image. When the ordered dissimilarity images (ODI) shown in Figure 1 are examined, the objects represented by the pink-colored pixels represent more similar objects, while the blue represents objects that are completely distant from each other. According to Figure 1, it is seen that the countries in the pixels coded as three regions form homogeneous groups for both datasets and there is good clustering.



(a) pre-vaccination *(b) post-vaccination*

Figure 2. The optimal number of clusters suggested by the NbClust package for the COVID-19 datasets sayıları

Furthermore, hierarchical clustering analysis was conducted via the NbClust package. For the number of clusters from 2 to 10, analysis performances were obtained in terms of 30 criteria (Charrad et al., 2014). The number of clusters recommended by the largest number of criteria for each dataset according to the modal value (Figure 2) was adopted in the study. In other words,

three clusters were recommended as the optimal number of clusters in terms of a total of 15 and 11 criteria in both datasets for pre-vaccination and post-vaccination periods, respectively.

Finally, by combining the results of the ODI and NbClust package, the number of clusters was determined as three for each dataset as shown in Figure 1 and 2.

3.2. Determining the optimal clustering algorithm for both COVID-19 datasets

In the clustering analysis, it is observed that the Ward.D2 method for the COVID-19 datasets has better clustering performance in terms of 27 criteria compared to the single linkage and complete linkage methods as a result of various trials. The performances of the algorithms represent the dissimilarities of the clusters with the total within-cluster variance between countries according to each internal criterion. Based on the decision matrices which are constructed separately for each dataset, alternative clustering algorithms through the MULTIMOORA technique are ranked according to their clustering performances and ranking results of two analyses are demonstrated in Table 1.

As a result of the algorithm performance rankings obtained by considering 27 criteria together, the optimal clustering algorithm to be used to determine the clusters of countries in the COVID-19 pre-vaccination dataset is determined as CLARA. In addition, the optimal clustering algorithm for the COVID-19 post-vaccination data is determined by utilizing ta similar calculation procedure of the COVID-19 pre-vaccination dataset. According to the results given in Table 1, the SOM algorithm is selected as the optimal clustering algorithm for the post-vaccination dataset.

Table 1. Ranking results of the MULTIMOORA method to select the best clustering algorithm

Alternative Clustering Algorithms	**Ranking for the pre-vaccination dataset**	**Ranking for the post-vaccination dataset**
AGNES	6	7
CLARA	1	10
DIANA	7	2
FANNY	2	4
GCLUST	4	6
HCLUST	3	5
KMEANS	10	11
EM MODEL	11	3
PAM	5	8
SOM	9	1
SOTA	8	11

3.3. Clustering results according to the level of impact of the COVID-19 pandemic on countries

Countries affected by the COVID-19 pandemic have clustered via the CLARA and SOM algorithms according to their proximity calculated from the Euclidean distance in terms of CFR, IR (per 1m), and VAC-R per relative population variables.

3.3.1. Clustering results of the COVID-19 pre-vaccination dataset

According to the results given in Table 2, the cluster medoid for Cluster 1 is determined as Slovakia. Pandemic statistics of Slovakia are given as 2.53 for CFR and 61,887.78 (per 1m) for IR in the records. It is observed that the most affected countries in the world are given in Cluster 1. When Cluster 2 is examined, Libya is determined as the cluster medoid (IR: 21,101.29 and CFR: 1.65). In terms of IR, the IR of Cluster 2 is almost 3 times lower, when compared to the country statistics in the cluster medoid (61,877.78) of Cluster 1. The IR of the countries in Cluster 2, as well as the CFR, is determined to be low (1.65<2.53). In addition, when their position on the world map is observed, most of the countries in Cluster 2, which are represented in green color, are in the northern hemisphere, and that many countries in this cluster (Cluster 2) are economically strong. This situation offers a perspective to address and evaluate the COVID-19 pandemic from another viewpoint.

Finally, most of the countries in Asia and Africa are gathered in Cluster 3 and this cluster is represented in black in Figure 3 (a). Ethiopia is determined as the cluster medoid for Cluster 3 with an IR of 1,526.28 and a CFR of 1.45. Although they are close to Cluster 2 in terms of CFR, a great difference is observed in terms of the IR. It is concluded that the countries in Cluster 1 are the country group most affected by the pandemic.

3.3.2. Clustering results of the COVID-19 post-vaccination dataset

According to Table 3, the clustering of 172 countries differs slightly from those of the pre-vaccination period. Cluster 1 included a total of 30 countries, with variable averages of 0.77% for CFR, 270240.91 for IR, and 70.39% for VAC-R, respectively. The vaccination rates of the countries in this cluster are highest compared to other clusters of countries. At the same time, the number of cases compared to the population is observed to be at the highest level in Cluster 1. However, despite the high number of cases, the average CFR in this cluster is found to be considerably lower than the average of other clusters.

Table 2. List of countries in terms of their level of being impacted by the COVID-19 pandemic for the pre-vaccination period

Cluster No	Countries
Cluster 1 (Cluster size: 47) medoid: Slovakia (CFR:2.53; IR:61,887.78)	Andorra, Argentina, Armenia, Austria, Bahrain, Belgium, Bosnia and Herzegovina, Brazil, Chile, Colombia, Croatia, Cyprus, Czechia, Estonia, France, Georgia, Ireland, Israel, Italy, Jordan, Kuwait, Latvia, Lebanon, Liechtenstein, Lithuania, Luxembourg, Malta, Moldova, Montenegro, Netherlands, North Macedonia, Panama, Peru, Poland, Portugal, Qatar, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, United States
Cluster 2 (Cluster Size: 46) medoid: Libya (CFR:. 1.65; IR:21,101.29)	Albania, Azerbaijan, Bahamas, Barbados, Belarus, Belize, Bolivia, Botswana, Bulgaria, Canada, Cape Verde, Costa Rica, Denmark, Dominican Republic, Ecuador, Eswatini, Finland, Germany, Greece, Guyana, Honduras, Iceland, Iran, Iraq, Kazakhstan, Libya, Maldives, Mexico, Morocco, Namibia, Norway, Oman, Palestine, Paraguay, Russia, Saint Lucia, Saint Vincent and the Grenadines, Seychelles, South Africa, Suriname, Tanzania, Tunisia, Turkey, Ukraine, Uruguay
Cluster 3 (Cluster Size: 79) medoid: Ethiopia (CFR:1.45; IR:1,526.28)	Afghanistan, Algeria, Angola, Antigua and Barbuda, Australia, Bangladesh, Benin, Bhutan, Brunei, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, China, Comoros, Congo, Cuba, Democratic Republic of Congo, Djibouti, Egypt, El Salvador, Equatorial Guinea, Ethiopia, Gabón, Gambia, Ghana, Grenada, Guatemala, Guinea, Guinea-Bissau, Haiti, India, Indonesia, Jamaica, Japan, Kenya, Lesotho, Liberia, Madagascar, Malawi, Malaysia, Mali, Mauritania, Mauritius, Mongolia, Mozambique, Myanmar, Nepal, Niger, Nicaragua, Nigeria, Pakistan, Philippines, Rwanda, Sao Tome and Principe, Saudi Arabia, Senegal, Sierra Leone, Singapore, South Korea, South Sudan, Sri Lanka, Sudan, Syria, Tajikistan, Thailand, Togo, Trinidad and Tobago, Uganda, Uzbekistan, Venezuela, Vietnam, Yemen, Zambia, Zimbabwe

*CFR: Case fatality rate. **IR: Incidence rate

Table 3 List of countries in terms of their level of being impacted by the pandemic for the COVID-19 post-vaccination dataset

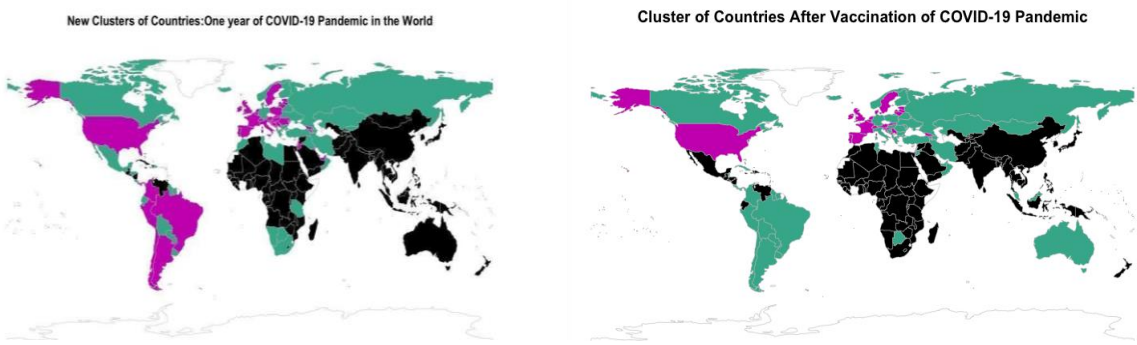
Cluster No	Countries
Cluster 1 (Cluster size: 30) mean values: (CFR:0.77; IR:270240.91. VAC-R: 70.39%)	Andorra, Austria, Bahrain, Belgium, Croatia, Cyprus, Czechia, Denmark, Estonia, France, Georgia, Ireland, Israel, Latvia, Liechtenstein, Lithuania, Luxembourg, Maldives, Montenegro, Netherlands, Portugal, Serbia, Seychelles, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom, United States
Cluster 2 (Cluster Size: 59) mean values: (CFR:1.61; IR:111777.54. VAC-R: 60.94%)	Albania, Antigua and Barbuda, Argentina, Armenia, Australia, Azerbaijan, Bahamas, Barbados, Belarus, Belize, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Canada, Cape Verde, Chile, Colombia, Costa Rica, Cuba, Finland, Germany, Greece, Grenada, Guyana, Hungary, Iceland, Iran, Italy, Jordan, Kazakhstan, Kuwait, Lebanon, Malaysia, Malta, Moldova, Mongolia, North Macedonia, Norway, Oman, Palestine, Panama, Paraguay, Peru, Poland, Qatar, Romania, Russia, Saint Lucia, Saint Vincent and the Grenadines, Singapore, Suriname, Trinidad and Tobago, Tunisia, Turkey, Ukraine, United Arab Emirates, Uruguay
Cluster 3 (Cluster Size: 83) mean values: (CFR:2.18; IR:15635.86 VAC-R: 33.90%)	Afghanistan, Algeria, Angola, Bangladesh, Benin, Bhutan, Brunei, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, China, Comoros, Congo, Democratic Republic of Congo, Djibouti, Dominican Republic, Ecuador, Egypt, El Salvador, Equatorial Guinea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guatemala, Guinea, Guinea-Bissau, Haiti, Honduras, India, Indonesia, Iraq, Jamaica, Japan, Kenya, Kyrgyzstan, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Mexico, Morocco, Mozambique, Myanmar, Namibia, Nepal, New Zealand, Nicaragua, Niger, Nigeria, Pakistan, Papua New Guinea, Philippines, Rwanda, Sao Tome and Principe, Saudi Arabia, Senegal, Sierra Leone, South Africa, South Korea, South Sudan, Sri Lanka, Sudan, Syria, Tajikistan, Tanzania, Thailand, Togo, Uganda, Uzbekistan, Venezuela, Vietnam, Yemen, Zambia, Zimbabwe

**CFR: Case fatality rate. **IR: Incidence rate ***VAC-R: people's fully vaccinated rate per relative population*

3.3.3. Comparative results of the clustering analysis on the world map

In this part of the study, a comparative evaluation of the clustering analysis was presented. The categories determined by the level of impact on the countries of the pandemic for COVID-19 datasets are shown on the map in Figure 3. Considering the distribution of the COVID-19 pandemic clusters in the world, Cluster 1 is represented in purple on the map. According to Figure 3(a), the pandemic has seriously affected a large part of the American continent and the European continent. It is noteworthy to point out that the countries of the European Union and USA, which hold world power economically, are also in this cluster. There were 47 countries in Cluster 1 in the pre-vaccination period while 30 countries were included in it in the post-vaccination period. When Figure 3(a) and 3(b) are compared, it is seen that all countries in cluster 1 in South America are in cluster 2 in the post-vaccination period. The reason for this is that the countries in South America are more behind the vaccination rates of developed countries in vaccination. In addition, it is observed that 10 countries in Europe have shifted from Cluster 1 to Cluster 2 in terms of vaccination and case fatality rates. It is known that only 30 countries have a very high vaccination rate, and the majority of these countries have a high level in terms of the economic and human development index

.



Pre-vaccination *b) post-vaccination*

Figure 3 New clusters of countries for COVID-19 pandemic in the world.

Cluster 2 for both datasets is represented in green on the world map. Considering the spread of the countries in Cluster 2 in the world, there are 46 countries in this cluster in the pre-vaccination period while 59 countries are included in it according to post-vaccination statistics. This is because, as previously stated, the vaccination statistics of many countries in the South American continent are higher than Cluster 3 and are included in Cluster 2.

Examining the spread of countries on the world map, the number of countries in Cluster 3 increased from 79 pre-vaccination to 83 in the post-vaccination period. Almost all countries in Central America, except for Cuba, are in Cluster 3. Cuba is in Cluster 3 before vaccination, while it is in Cluster 2 in the post-vaccination period. In addition, a similar interpretation can be made for the Australian continent. This suggests that Cuba and Australia have a high vaccination rate.

As a result, as the vaccination rate increased, the CFR decreased, and vaccination was found to make a significant difference to the clusters of countries.

3.3.4. Relationship between human development levels and clusters of COVID-19 pandemic

To examine the difference between the human development levels and the country clusters, the two-dimensional contingency tables given in Table 4 were formed. To calculate the association between these two ordinal categorical variables, the gamma (γ) coefficient of Goodman and Kruskal (Kvålseth, 2017) was preferred. The value of Goodman-Kruskal's Gamma statistic was calculated as G=0.8228 for the pre-vaccination contingency table and G=0.924 for the post-vaccination contingency table given in Table 4. Furthermore, according to the result of the hypothesis test ($Z = 14.326$. p-value $< 2.2e-16$) for the significance of this test statistic, the null hypothesis was rejected at the 0.01 significance level. The value of gamma statistic investigates that the strong similarity between the groups of countries based on human development and the country clusters of the COVID-19 pandemic is quite high and positively associated.

Besides, when the clusters created for the post-vaccination period are taken into account, it was found that the similarity between the categories of countries after vaccination and HDI (gamma:0.924) was greater than the pre-vaccination similarity (gamma:0.823). Therefore, as these results show, the countries with high vaccination rates are also countries with very high and high HD level

Table 4 Contingency tables of HD levels and COVID-19 pandemic clusters for both COVID-19 datasets

	Human Development Levels of Countries			
Before vaccination	**Very High**	**High**	**Medium**	**Low**
Cluster 1 (high)	38	9	0	0
Cluster 2 (medium)	17	20	8	1
Cluster 3 (low)	9	17	23	31
After vaccination	**Very High**	**High**	**Medium**	**Low**
Cluster 1 (high vac.)	28	2	0	0
Cluster 2 (medium vac.)	30	27	2	0
Cluster 3 (low vac.)	6	17	29	31

4. Conclusions and Recommendations

Since the declaration of the pandemic, many clustering studies have been conducted for the COVID-19 data presented in different sources in the world and Turkey and taken at various date intervals. Although the k-means algorithm is mostly preferred due to its general success in determining country clusters, it has been observed that the hierarchical clustering analysis algorithm is used in a few studies. To the best of our knowledge, a study on COVID-19 has not been encountered in which a detailed clustering process considering all the criteria was performed as in this study.

In this study, it has been proposed to use one of the MODM methods based on 27 validation measures for the selection of algorithms with which researchers have the most difficulty and indecisiveness in cluster analysis.

According to the results, the incidence rate variable is the most dominant factor in the real difference between clusters. Another remarkable finding is that while countries with economic power and a high level of human development are expected to be less affected by the pandemic before the vaccination, the level of being affected by the pandemic increases in terms of both variables as the level of human development increases.

According to the cluster statistics in question, it has been observed that the vaccination rate has no positive effect on the number of cases while significantly reducing the case fatality rate. Therefore, countries with high HDI levels have a high vaccination rate and consequently low case fatality rates. However, the effect of vaccination on the number of cases might not seem to be very significant.

South Asian countries, except for Central America, almost the entire African continent, and the Australian continent, lagged far behind in the vaccination process and therefore there was no significant decrease in case fatality rates. Compared to developed countries, there has been a delay in the delivery of vaccines to developing world countries, or it has been concluded that their culture and the way they handle the pandemic process have influenced societies' attitudes towards vaccination. In short, this study has revealed the geographical location of countries and clustering similarities pre- and post- vaccination in the pandemic, and then has provided the possible relationship of these clusters with the level of human development.

Consequently, being able to deal with diseases that occur worldwide requires global cooperation in identifying, controlling, and preventing these diseases. It is therefore vital to establish a geographic risk assessment and to group countries for this risk assessment. Thus, it can be ensured that diseases are prevented before they become pandemics thanks to the measures to be taken quickly and on time.

References

Ahmad, K., Erqou, S., Shah, N., Nazir, U., Morrison, A.R., Choudhary, G., Wu, W. C. (2020). Association of poor housing conditions with COVID-19 incidence and mortality across US counties. PloS One, 15(11), e0241327.

Asem, N., Ramadan, A., Hassany, M., Ghazi, R.M., Abdallah, M., Ibrahim, M., Gamal, E. M. Hassan, S., Kamal, N., & Zaid, H. (2021). Pattern and determinants of COVID-19 infection and mortality across countries: An ecological study. Heliyon, 7(7).

Aydın, N. & Seven, A. N. (2015). İl nüfus ve vatandaşlik müdürlüklerinin iş yoğunluğuna göre hibrid kümeleme ile sınıflandırılması. Journal of Management and Economics Research, 13 (2), 181-201.

Berkhin, P. Survey of Clustering Data Mining Techniques, Accrue Software Inc., San Jose, California, USA (2002).

Bezdek, J., & Hathaway, R.J. (2002). VAT: A tool for visual assessment of (cluster) tendency. Proceedings of the International Joint Conference on Neural Networks, 3, 2225 - 2230. https://doi.org/10.1109/IJCNN.2002.1007487.

Bolshakova, N. Azuaje, F.J. (2003). Cluster validation techniques for genome expression data, Signal Process. 83 825-833. https://doi.org/10.1016/S0165-1684(02)00475-9.

Bradley, P. S., Mangasarian, O. L. and Street, W. N. Clustering via Concave Minimization, in Advances in Neural Information Processing Systems 9, M. C. Mozer, M. I. Jordan, and T. Petsche (Eds.) (1997) 368- 374, MIT Press.

Brauers, K.W.M., Zavadskas, E.K., Turskis, Z., Vilutienė, T. (2008). Multi-objective contractor's ranking by applying the MOORA method. Journal of Business Economics and Management, 9(4) 245-255. https://doi.org/10.3846/1611-1699.2008.9.245-255

Brauers, W.K.M., & Zavadskas, E. K. (2011). MULTIMOORA optimization used to decide on a bank loan to buy property, Technological and Economic Development of Economy 17(1) 174-188. https://doi.org/10.3846/13928619.2011.560632.

Brauers, W.K.M. & Ginevičius R., (2010). The Economy of the Belgian Regions tested with MULTIMOORA, Journal of Business Economics and Management. 11(2), 173–209. http://doi.org/10.3846/jbem.2010.09.

Cebeci, Z. (2020). fcvalid: an r package for internal validation of probabilistic and possibilistic clustering. Sakarya University Journal of Computer and Information Sciences, 3(1). https://doi.org/10.35377/saucis.03.01.664560

Charrad, M. Ghazzali, N. Boiteau, & V. Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set. Journal of Statistical Software, 61 (6) 1–36. https://doi.org/10.18637/jss.v061.i06.

Chu, J. (2021). A statistical analysis of the novel coronavirus (COVID-19) in Italy and Spain. PLoS ONE, 16(3), e0249037. https://doi.org/10.1371/journal.pone.0249037.

Couvreur, C. The EM algorithm: a guided tour. In: Kárný M., Warwick K. (eds) Computer Intensive Methods in Control and Signal Processing. Birkhäuser, Boston, MA (1997). https://doi.org/10.1007/978-1-4612-1996-5_12.

Dalton, L. Ballarin, V., & Brun, M. (2009). Clustering algorithms: on learning, validation, performance, and applications to genomics, Current Genomics. 10 430-445. https://dx.doi.org/10.2174/138920209789177601.

Dempster, A.P., Paird, N.M. & Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. 39(1),1–38.

Desgraupes, B. (2012). ClusterCrit: Clustering Indices. Available online: https://cran.r-proje ct.org/web/packages/clusterCrit/.

Desgraupes, B. (2016). ClusterCrit: clustering indices R package version 1.2.8. https://cran.r-proje ct.org/web/packages/clusterCrit/.

Dopazo, J. Carazo, J.M. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree, J Mol Evol. 44(2) (1997) 226-33. http://dx.doi.org/10.1007/pl00006139.

Dunham, M.H. Data Mining Introductory and Advanced Topics, Prentice Hall, USA (2003).

Fraley, C. Raftery, A.E. How many clusters? Which clustering method? Answers via model-based cluster analysis, Computer Journal. 41(8) (1998) 578–588.

Flexer, A. On the use of self-organizing maps for clustering and visualization, Intelligent Data Analysis, 5(5) (2001) 373-384.

Gagolewski, M., Bartoszuk, M., & Cena, A. (2016). Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm. Inform Sci, 363, 8–23. http://dx.doi.org/10.1016/j.ins.2016.05.003.

Gokmen, Y., Baskici, C., & Ercil, Y. (2021). The impact of national culture on the increase of COVID-19: A cross-country analysis of European countries. International Journal of Intercultural Relations, 81, 1-8. https://doi.org/10.1016/j.ijintrel.2020.12.006.

Gupta, M. R. &Chen, Y. (2011). Theory and use of the EM algorithm, Foundations and Trends in Signal Processing. 4(3), 223-296. http://dx.doi.org/10.1561/2000000034.

Halkidi M., Batistakis Y., & Vazirgiannis M., On clustering validation techniques, Journal of Intelligent Information Systems. 17 (2001) 107–145. https://doi.org/10.1023/A:1012801612483.

Han, J. Kamber M., Pei, J. Data mining: Concepts and techniques, (3rd ed.). Morgan Kaufmann Publishers (2012).

Harapan, H., Itoh, N., Yufika, A. Winardi, W., Keam, S. Te, H., Megawati, Hayati, D. Z., Wagner, A.L., & Mudatsir, M. (2020). Coronavirus disease 2019 (COVID-19): A literature review. J Infect Public Health, 13(5), 667-673. doi: 10.1016/j.jiph.2020.03.019.

Hartigan, J.A & Wong, M.A., Algorithm AS 136: A k-means clustering algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics). 28 (1979) 100-108. http://dx.doi.org/10.2307/2346830.

Hasell, J., Mathieu, E., Beltekian, D., Macdonald, B., Giattino, C., Ortiz-Ospina, E., Roser, M., & Ritchie, H. (2020). A cross-country database of COVID-19 testing. Scientific Data, 7(1), 345. https://doi.org/10.1038/s41597-020-00688-8.

Herrero, J. Valencia A., Dopazo, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns, Bioinformatics. 17(2) (2001) 126-36. https://doi.org/10.1093/bioinformatics/17.2.126.

Hezam, I.M. (2021). COVID-19 Global Humanitarian Response Plan: An optimal distribution model for high-priority countries. ISA Transactions. https://doi.org/10.1016/j.isatra.2021.04.006.

HDR. (2020). Human Development Reports. http://hdr.undp.org/en/2020-report (google Scholar)

Itoh, H. Market area analysis of ports in Japan: an application of fuzzy clustering, in: The IAME2013 Annual Conference, Marseille, France. (2013) 1-21. hal-00918672

Karmakar, M. Lantz, P. M., & Tipirneni, R. (2021). Association of social and demographic factors with COVID-19 incidence and death rates in the US. JAMA network open, 4(1), e2036462.

Kaufman, L., Rousseeuw, P. J. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons (2005).

Khafaie, M.A., & Rahim, F., (2020). Cross-country comparison of case fatality rates of COVID-19/SARS-COV-2. Osong. Public Health Res Perspect, 11(2), 74-80. https://dx.doi.org/10.24171/j.phrp.2020.11.2.03.

Kiang M.Y., Extending the Kohonen self-organizing map networks for clustering analysis, Computational Statistics and Data Analysis. 38 (2001) 161–180. https://doi.org/10.1016/S0167-9473(01)00040-8.

Kurniawan, R. Sheikh Abdullah, S. N. H. Lestari, F. Nazri, M. Z. A. Mujahidin, A. and Adnan, N. (2020) Clustering and correlation methods for predicting coronavirus COVID-19 risk analysis in pandemic countries, 8th International Conference on Cyber and IT Service Management (CITSM). 1-5. htpps://doi.org/ 10.1109/CITSM50537.2020.9268920.

Kuster, A.C., & Overgaard, H.J. (2021). A novel comprehensive metric to assess effectiveness of COVID-19 testing: Inter-country comparison and association with geography,

government, and policy response. PLoS One, 16(3), e0248176. doi: 10.1371/journal.pone.0248176

Kucukefe, B. (2020). Clustering macroeconomic impact of COVID-19 in OECD countries and China, Ekonomi Politika Ve Finans Araştırmaları Dergisi. 5 (2020) 280–291. https://doi.org/10.30784/epfad.811289.

Kvålseth, T.O. (2017). An alternative measure of ordinal association as a value-validity correction of the Goodman–Kruskal gamma. Communications in Statistics - Theory and Methods, 46 (21), 10582-10593. http://doi.org/ 10.1080/03610926.2016.1239114

Li, M., Zhang, Z., Cao, W., Liu, Y., Du, B., Chen, C., Liu, Q., Uddin, M.N., Jiang, S., Chen, C., Zhang, Y., & Wang, X. (2021). Identifying novel factors associated with COVID-19 transmission and fatality using the machine learning approach. Sci Total Environ, 764, 142810. doi: 10.1016/j.scitotenv.2020.142810.

Liu, K., He, M., Zhuang, Z., He, D., & Li, H. (2020). Unexpected positive correlation between human development index and risk of infections and deaths of COVID-19 in Italy. One Health, 10, 100174. DOI: https://doi.org/10.1016/j.onehlt.2020.100174

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2021). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.2 — For new features, see the 'Changelog' file (in the package source). https://CRAN.R-project.org/package=cluster.

Marziali, E.M., Hogg, R.S., Oduwole, O.A. & Card, K.G. (2021). Predictors of COVID-19 testing rates: A cross-country comparison. International Journal of Infectious Diseases, 104, 370-372.

McKenzie, G., & Adams, B. (2020). A country comparison of place-based activity response to COVID-19 policies. Applied geography, 125, 102363.https://doi.org/10.1016/j.apgeog.2020.102363.

McLachlan, G.J. Krishnan, T. & Ng, S.K. (2004). The EM algorithm, Working Paper No. 2004, 24, Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics (CASE), Berlin http://hdl.handle.net/10419/22198.

Milligan, G.W. (1981). A monte carlo study of thirty internal criterion measures for cluster analysis, Psychometrika. 46(2), 187–199.

Moshtaghi M., Bezdek, J. CErfani, S.M., Leckie, C. & Bailey, J. (2019). Online cluster validity indices for performance monitoring of streaming data clustering, International Journal of Intelligent Systems. 34, 541 - 563. https://dx.doi.org/10.1002/int.22064.

OWD. (2022). COVID-19 Data, https://ourworldindata.org/coronavirus-testing#testing-for-covid-19-background-the-our-world-in-data-covid-19-testing-dataset.

Pérez, L.A., García-Vico, Á.M., González, P., & Carmona, C.J. (2020). Techniques for evaluating clustering data in R, The Clustering Package. https://cran.r-project.org/web/packages/Clustering/vignettes/Clustering.pdf

Estivill-Castro, V&Yang., J. (2000), Fast and Robust General Purpose Clustering Algorithms. In: Mizoguchi R., Slaney J. (eds) PRICAI 2000 Topics in Artificial Intelligence. PRICAI 2000. Lecture Notes in Computer Science. vol 1886. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44533-1_24.

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E.M. (2011). Internal versus external cluster validation indexes. International Journal of Computers and Communications, 5(1).

Rocha, R., Atun, R., Massuda, A., Rache, B., Spinola, P., Nunes, L., Lago, M., & Castro, M.C. (2021). Effect of socioeconomic inequalities and vulnerabilities on health-system preparedness and response to COVID-19 in Brazil: a comprehensive analysis. Lancet Glob Health, 9, e782–92.

RStudio Team. (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA http://www.rstudio.com/.

Sekula, M., Datta, S., & Datta, S. (2017). optCluster: An R package for determining the optimal clustering algorithm. Bioinformation, 13(3), 101-103. http://doi.org/:10.6026/97320630013101

Shahbazi, F., & Khazaei, S. (2020). Socio-economic inequality in global incidence and mortality rates from coronavirus disease 2019: an ecological study. New Microbe and New Infect, 38, 100762.

Sharma, A., Borah, S. B., & Moses, A.C. (2021). Responses to COVID-19: The role of governance, healthcare infrastructure, and learning from past pandemics. Journal of Business Research, 122, 597-607. https://doi.org/10.1016/j.jbusres.2020.09.011

Siddik, N. A. (2020). Economic stimulus for COVID-19 pandemic and its determinants: evidence from cross-country analysis. Heliyon, 6 (12). https://doi.org/10.1016/j.heliyon.2020.e05634.

Tosepu, R., Gunawan, J., Effendy, D. S., Lestari, H., Bahar, H., & Asfian, P. (2020). Correlation between weather and COVID-19 pandemic in Jakarta, Indonesia. Science of the Total Environment, 725, 138436.

Van Craenendonck, T., & Blockeel, H. (2015). Using Internal Validity Measures to Compare Clustering Algorithms. ICML 2015 AutoML Workshop.

Virgantari, & Faridhan, Y.E. K-means clustering of COVID-19 cases in Indonesia's provinces, in: Proceedings of the International Conference on Global Optimization and Its Applications Jakarta, Indonesia (2020).

VoPham, T., Weaver, M.D., Hart, J. E., Ton, M., White, E., Newcomb, P. A. (2020). Effect of social distancing on COVID-19 incidence and mortality in the US. MedRxiv: the preprint server for health sciences. https://doi.org/10.1101/2020.06.10.20127589

Yuan, J., Wu, Y., Jing, W., Liu, J., Du, M., Wang, Y., & Liu, M. (2021). Association between meteorological factors and daily new cases of COVID-19 in 188 countries: A time series analysis, Science of The Total Environment, 780. https://doi.org/10.1016/j.scitotenv.2021.146538.

Wani, M.A. & Riyaz, R. A (2016). new cluster validity index using maximum cluster spread based compactness measure, International Journal of Intelligent Computing and Cybernetics. 9(2) 179-204. https://doi.org/10.1108/IJICC-02-2016-0006.

Wickham, H., Hester, J., & Chang, W. (2021). devtools: Tools to make developing R packages Easier. R package version 2.4.2. https://CRAN.r-project.org/package=devtools

Wu, J.T., Leung, K., & Leung, G.M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modeling study. The Lancet, 395 (10225), 689-697.