Journal of Health and Social Sciences 2020; 5,4:587-594
The Italian Journal for Interdisciplinary Health and Social Development

ORIGINAL ARTICLE
IN COVID 19 AND STATISTICS

Clustering of countries in terms of deaths and cases of COVID-19

Ozge PASIN[1], Tugce PASIN[2]

Affiliations:

[1] PhD, Department of Biostatistics, Faculty of Medicine, Istanbul University, Istanbul, Turkey
[2] MD, Department of Physical Medicine and Rehabilitation, Istanbul Goztepe Training and Research Hospital, Istanbul, Turkey.

Corresponding author:

Ozge Pasin, PhD, Department of Biostatistics, Istanbul Faculty of Medicine, Istanbul University, Capa-Fatih Istanbul, Turkey. E-mail address: ozgepasin90@yahoo.com.tr.

Abstract

Introduction: The novel coronavirus 'Severe Acute Respiratory Syndrome Coronavirus Type 2' (SARS-CoV-2), responsible for the disease termed as 'Coronavirus disease 2019' (COVID-19 pandemic) first broke out in Wuhan, Hubei province, mainland China. With its rapid spread and reports revealing the crucial consequences of this spread, countries adopted strict measures to tackle the disease. The objective of this study is to cluster the various countries describing the course of the COVID-19 outbreak.
Methods: The data used was obtained from the Worldometers' website on October 22, 2020 in its most current form for 191 countries. The number of total cases and deaths were used. The numbers were calculated considering population sizes. The total deaths / 1million population and total cases/ 1million population were used for clustering. For clustering k-means clustering method and elbow method were used. Also the two-step clustering method was used for the clustering process.
Results: As a result, Armenia, Aruba, Bahrain, French Guiana, Israel, Kuwait, Montenegro, Oman, Qatar, San Marino were a single cluster apart from other countries for two-step clustering. Also in k-means clustering Aruba, Bahrain, French Guiana, Israel and Qatar was a single cluster apart from other countries for k-means clustering.
Conclusion: This study will be of great importance, when showing the differences among countries in terms of total deaths and cases in terms of population. Proper use of these data will help states take precautions regarding COVID-19.

KEY WORDS: Clustering; Coronavirus; COVID-19: k-means; statistics.

Riassunto

Introduzione: Il nuovo coronavirus 'Severe Acute Respiratory Syndrome Coronavirus Type 2' (SARS-CoV-2), responsabile della malattia denominata 'Coronavirus disease 2019' (COVID-19 pandemic) è scoppiata all'inizio a Wuhan, nella provincia dell'Hubei, in Cina. Con la sua rapida diffusione ed i report che rivelano le conseguenze cruciali di questa diffusione, i Paesi hanno adottato misure severe per contrastare la malattia. L'obiettivo di questo studio è di raggruppare i vari Paesi descrivendo l'andamento dell'epidemia da COVID-19.
Metodi: I dati usati sono stati ottenuti dal website Worldometer il 22 Ottobre 2020 nella sua più recente forma per 191 Paesi. Il numero di casi totali e di morti è stato usato. I numeri sono stati calcolati considerando le dimensioni della popolazione. Il rapporto tra morti totali ed 1 milione di abitanti per popolazione e numero di casi totali per 1 milione di abitanti per popolazione è stato utilizzato per il raggruppamento. Un metodo di raggruppamento a 2 fasi (metodi Elbow e k-means clustering) è stato usato per le analisi. Inoltre il metodo di raggruppamento a 2 step è stato usato per il processo di raggruppamento.
Risultati: Come risultato, Armenia, Aruba, Bahrain, French Guiana, Israel, Kuwait, Montenegro, Oman, Qatar, San Marino sono stati un cluster singolo ad eccezione degli altri Paesi per il clustering a 2 step. Inoltre nel raggruppamento k-means, Aruba, Bahrain, French Guiana, Israel e Qatar hanno rappresentano un singolo cluster ad eccezione degli altri Paesi per il clustering K-means.
Conclusione: Questo studio sarà di grande importanza, per mostrare le differenze tra Paesi in termini di morti totali e di casi in termini di popolazione. L'appropriato uso di questi dati aiuterà gli stati a prendere precauzioni rispetto al COVID-19.

TAKE-HOME MESSAGE

Cluster results will enable governments to implement suitable measures for subsequent similar situations and it is useful to a variety of different policies.

Competing interests - none declared.

Cite this article as: Pasin O, Pasin T. Clustering of countries in terms of deaths and cases of COVID-19. J Health Soc Sci. 2020;5(4):587-594

DOI 10.19204/2020/clst15

Received: 30/09/2020 Accepted: 25/10/2020 Published Online: 30/10/2020

INTRODUCTION

The novel coronavirus 'Severe Acute Respiratory Syndrome Coronavirus Type 2' (SARS-CoV-2), responsible for the disease termed as 'Coronavirus disease 2019' (COVID-19 pandemic) first broke out in Wuhan, Hubei province, mainland China. With its rapid spread and reports revealing the crucial consequences of this spread, countries adopted strict measures to tackle the disease. There is a lot of data in the literature on COVID-19 and they are constantly updated. Proper use of these data will help states take precautions regarding COVID-19. Clustering is a helpful tool for this purpose. Cluster results could be useful to a variety of different policy makers, such as physicians and managers of the health sector, finance experts, politicians and even to sociologists [1, 2].

The virus was identified in the first half of January 2020. The epidemiological features of the disease are still unknown, and the number of total cases and the number of total deaths significantly vary day by day. When the rapid spread and serious consequences of the disease were observed, precautions were taken and positive cases began to be recorded after the second half of January. Because the number of total cases and the number of total deaths are used for examining the course of the outbreak, modeling of these indicators is an important issue. Modeling results are valuable for determining the appropriate prevention strategies.

The objective of this study is to cluster the various countries describing the course of the COVID-19 outbreak. To present this, we focused on: the number of total cases diagnosed with the disease and the number of deaths. Numbers are calculated considering population sizes.

METHODS

Study instruments and measures

Cluster analysis

Clustering is the process of grouping similar objects. While the objects in a set are similar to each other in terms of the properties studied, the objects in different clusters are not similar in terms of the properties in question. Homogeneous groups are revealed with the help of cluster analysis. Thus, many objects will be divided into meaningful groups [3].

K-means method

K-means clustering method is the most widely used partitional clustering methods in the literature. It is a non-hierarchical clustering method with unsupervised learning. Since its application and interpretation is simple, it is seen that this method is frequently used in the literature. The reason why the method is called K-mean is that K indicates the number of clusters determined and to be formed at the beginning, and the averages show the center of cluster used to measure cluster similarity. Assume that sets of N samples are given in N dimensional space. The space is divided into K clusters as $\{C1, C2 \ldots C_k\}$.

The mean vector of cluster (C_k) is calculates as below.

$$M_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ik}$$

X_{ik} is the i-sample of C_k cluster. The difference of clusters are generally calculated with Euclidian distance measure [4].

Two-step clustering method

The two-step clustering algorithm combines both hierarchical and partitioning methods. It utilizes a two-step approach similar to the "balanced iterative reducing and clustering using hierarchies" (BIRCH) technique. The two-step method involves two steps including the pre-clustering and the clustering steps. The former scans the data record one by one and decides whether the current record can be added to one of the previously formed clusters or it starts a new cluster based on distance criterion. It uses clustering feature (CF) for clustering. In CF there are nodes and these nodes have a number of entries. In this step, it is investigated the nearest leaf entry in leaf nodes. If this leaf entry is

within the threshold distance that was determined initially, it is included into the nearest leaf entry. Otherwise, a new value is generated for the leaf node. In the clustering step sub-clusters are used, being obtained from the pre- clustering step as input, and then they are grouped in the desired number of clusters. Also in this method, there is no need to specify an input parameter like the number of clusters, because this method automatically estimates the ideal number of clusters by the help of the Bayesian (BIC) and Akaike (AIC) information criteria. The initial estimation of the optimal number of clusters is calculated easily with these indicators [5].

Study participants and sampling

The data used in the study was obtained from the Worldometers' website on October 22, 2020. Worldometers provides reliable data by updating data on COVID-19 on a daily basis. Raw cases and death numbers were not considered because taking these numbers into consideration would create bias, as such analyses were carried out considering the total number of cases and deaths per population, adjusting for the population numbers. Also, the figures used in this study are cumulative numbers . In other words, the data includes all the number of cases from the beginning of COVID-19 to the date of 22 October 2020. Thus, more accurate results will be obtained by considering the populations of the countries. 191 different countries were included in the study for cluster analysis. We excluded missing data from the analysis. Therefore, we analysis 191 country data [6].

Data analysis

A cluster analysis was used for grouping countries in terms of deaths and cases number. The total within-cluster sum of square was calculated for determining number of clusters. In the study, the Elbow method was used for defining the best cluster number. After deciding the optimal number of clusters, K -means algorithm was implemented to divide the data into clusters, thus determining cluster membership for each observation. Then the two-step clustering algorithm was used for the clustering. In this method, the optimal cluster numbers are defined by the method, not based on the user information. Statistical analysis in the study was made using SPSS 21.0 and R package program. The level of statistical significance was taken as 0.05 and $P < 0.05$ was considered statistically significant.

RESULTS

The number of clusters should be decided, before clustering analysis. In order to determine the number of clusters, the best cluster number was decided by the Elbow method using total within cluster sum of square values. The results obtained are given in Figure 1. When the figure is examined, it is concluded that the most suitable number of clusters will be four.
In the second stage of the study, after deciding on the number of clusters, the k-means clustering method was applied. There are 11,5% of the data in cluster 1, 20,4% in cluster 2, 65,4% in cluster 3 and 2,6% in the last cluster. It has been observed that most of the countries gather in the third cluster.
The results of the distances between clusters are given in Table 2. When the table is examined, it is determined that the clusters closest to each other are cluster 2 and cluster 3, while the most distant clusters are cluster 3 and cluster 4.
Table 3 shows the memberships of countries in four clusters. When the table is examined, it can be seen that Aruba, Bahrain, French Guiana, Israel and Qatar were a single cluster apart from other countries.
Descriptive statistics of total deaths and total cases per population in terms of sets are given in Table 4 as means and standard deviations. It has been observed that the average of total deaths and cases / 1million population of Cluster 4, was higher than other countries. In cluster 3, the average of total cases and death per population were low compared to other clusters. In cluster 1, the number of deaths was quite high compared to other clusters. In the countries in cluster 4, although the total cases / 1 million population average was higher, the death figures were relatively higher.

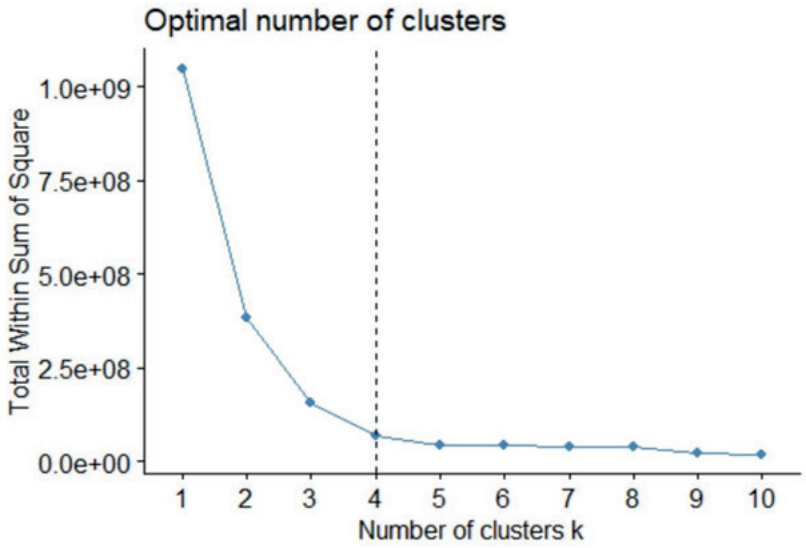

Figure 1. The optimal number of clusters according to the Elbow method.

Table 1. Distribution of clusters according to the k-means clustering.

		N	%
Cluster	1	22	11.5
	2	39	20.4
	3	125	65.4
	4	5	2.6
N		191	100

Table 2. Distances between Final Cluster Centers according to the k-means clustering.

Distances between Final Cluster Centers				
Cluster	1	2	3	4
1		12567.95	20449.58	17902.79
2	12567.95		7881.63	30467.29
3	20449.58	7881.63		38347.65
4	17902.79	30467.29	38347.65	

Table 3. Cluster memberships of countries according to the k-means clustering.

Cluster	**Countries**
Cluster 1	Argentina, Armenia, Belgium, Brazil, Chile, Colombia, Costa Rica, Czechia, Guadeloupe, Kuwait, Luxembourg, Maldives, Moldova, Montenegro, Oman, Panama, Peru, San Marino, Sint Maarten, Spain, Turks and Caicos, USA.
Cluster 2	Albania, Austria, Bahamas, Belarus, Belize, Bolivia, Croatia, Denmark, Dominican Republic, France, Honduras, Iceland, Iran, Iraq, Ireland, Kazakhstan, Kyrgyzstan, Lebanon, Libya, Liechtenstein, Malta, Martinique, Mayotte, Mexico, Monaco, Netherlands, North Macedonia, Palestine, Paraguay, Portugal, Russia, Saint Martin, Saudi Arabia, Slovenia, South Africa, Suriname, Switzerland, UK, Ukraine.
Cluster 3	Afghanistan, Algeria, Andorra, Angola, Antigua and Barbuda, Australia, Azerbaijan, Bangladesh, Barbados, Benin, Bermuda, Bosnia and Herzegovina, Botswana, British Virgin Islands, Brunei, Bulgaria, Burkina Faso, Burundi, Cabo Verde, Cameroon, Canada, CAR, Caribbean Netherlands, Cayman Islands, Chad, Channel Islands, China, Comoros, Congo, Cuba, Curaçao, Cyprus, Djibouti, DRC, Ecuador, Egypt, El Salvador, Equatorial Guinea, Estonia, Eswatini, Ethiopia, Fiji, Finland, French Polynesia, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Hong Kong, Hungary, India, Indonesia, Isle of Man, Italy, Ivory Coast, Jamaica, Japan, Jordan, Kenya, Latvia, Lesotho, Liberia, Lithuania, Madagascar, Malawi, Malaysia, Mali, Mauritania, Mauritius, Montserrat, Morocco, Mozambique, Myanmar, Namibia, Nepal, New Zealand, Nicaragua, Niger, Nigeria, Norway, Pakistan, Papua New Guinea, Philippines, Poland, Réunion, Romania, Rwanda, S. Korea, Sao Tome and Principe, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Somalia, South Sudan, Sri Lanka, Sudan, Sweden, Syria,Taiwan,Tajikistan,Tanzania,Thailand,Togo,Trinidad and Tobago, Tunisia, Turkey, UAE, Uganda, Uruguay, Uzbekistan, Venezuela, Vietnam, Western Sahara, Yemen, Zambia, Zimbabwe.
Cluster 4	Aruba, Bahrain, French Guiana, Israel, Qatar.

Table 4. Descriptive statistics of clusters according to the k-means clustering.

	Cluster Number of Case	Mean	Median	Std. Deviation	Minimum	Maximum
Total Cases/ 1Million Population	1	22260.31	22108.50	3539.34	17065.00	29163.00
	2	9694.30	9488.00	2767.38	5832.00	15342.00
	3	1814.21	928.00	1817.23	8.00	5704.00
	4	40161.60	40868.00	6173.49	33461.00	46374.00
Total Deaths/ 1Million Population	1	445.41	401.00	282.44	1.24	1025.00
	2	224.33	171.00	177.31	26.00	730.00
	3	68.15	21.00	130.14	2.00	815.00
	4	212.80	229.00	91.36	80.00	327.00

Table 5. Distribution of clusters according to the two-step clustering.

Cluster		N	%
Clusters	Cluster 1	160	83,8
	Cluster 2	21	11,0
	Cluster 3	10	5,2
	Total	191	100,0

Table 6. Cluster memberships of countries according to the two step clustering.

Cluster	**Countries**
Cluster 1	Afghanistan, Albania, Algeria, Angola, Antigua and Barbuda, Australia, Austria, Azerbaijan, Bahamas, Bangladesh, Barbados, Belarus, Belize, Benin, Bermuda, Bosnia and Herzegovina, Botswana, British Virgin Islands, Brunei, Bulgaria, Burkina Faso, Burundi, Cabo Verde, Cameroon, Canada, CAR, Caribbean Netherlands, Cayman Islands, Chad, Channel Islands, China, Comoros, Congo, Costa Rica, Croatia, Cuba, Curaçao, Cyprus, Czechia, Denmark, Djibouti, Dominican Republic, DRC, Egypt, El Salvador, Equatorial Guinea, Estonia, Eswatini, Ethiopia, Fiji, Finland, French Polynesia, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Guadeloupe, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Isle of Man, Ivory Coast, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Kyrgyzstan, Latvia, Lebanon, Lesotho, Liberia, Libya, Liechtenstein, Lithuania, Luxembourg, Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Martinique, Mauritania, Mauritius, Mayotte, Monaco, Montserrat, Morocco, Mozambique, Myanmar, Namibia, Nepal, New Zealand, Nicaragua, Niger, Nigeria, Norway, Pakistan, Palestine, Papua New Guinea, Paraguay, Philippines, Poland, Portugal, Réunion, Romania, Russia, Rwanda, S. Korea, Saint Martin, Sao Tome and Principe, Saudi Arabia, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Slovenia, Somalia, South Africa, South Sudan, Sri Lanka, Sudan, Suriname, Switzerland, Syria, Taiwan, Tajikistan, Tanzania, Thailand, Togo, Trinidad and Tobago, Tunisia, Turkey, Turks and Caicos, UAE, Uganda, Ukraine, Uruguay, Uzbekistan, Venezuela, Vietnam, Western Sahara, Yemen, Zambia, Zimbabwe.
Cluster 2	Andorra, Argentina, Belgium, Bolivia, Brazil, Chile, Colombia, Ecuador, France, Italy, Mexico, Moldova, Netherlands, North Macedonia, Panama, Peru, Sint Maarten, Spain, Sweden, UK, USA.
Cluster 3	Armenia, Aruba, Bahrain, French Guiana, Israil, Kuwait, Montenegro, Oman, Qatar, San Marino.

Table 7. Descriptive statistics of clusters according to the two step clustering.

	TwoStep Cluster Number	N	Mean	Median	Std. Deviation	Minimum	Maximum
Total Cases/ 1Million Population	1	160	4029.16	2357.50	4674.17	8.00	20823.00
	2	21	15635.61	17065.00	9377.11	493.00	29163.00
	3	10	32237.60	30546.00	9442.55	21666.00	46374.00
Total Deaths/ 1Million Population	1	160	76.85	40.50	88.64	2.00	377.00
	2	21	647.71	649.00	157.03	401.00	1025.00
	3	10	223.32	225.00	125.48	1.24	398.00

After k-means algorithm, we applied the two-step clustering method. According to this method, there are 83.8% of the data in cluster 1, 11% in cluster 2 and 5.2% in cluster 3. It has been observed that most of the countries gather in the first cluster.

Table 6 shows the memberships of countries according to the two-step clustering method. In the two-step clustering, three clusters were found. When the table was examined, it can be seen that Armenia, Aruba, Bahrain, French Guiana, Israel, Kuwait, Montenegro, Oman, Qatar, San Marino were a single cluster apart from other countries. Also in the k-means clustering these countries were a single cluster. The distribution of all other countries by clusters can be seen in detail in the table. The descriptive statistics of the clusters are given in Table 7. It has been observed that the mean of total cases / 1million population of Cluster 3, was higher than other countries (clusters) and the cluster 1 mean was the lowest. The mean of total death per population of cluster 1 was lower compared to other clusters and the cluster 2 was the highest.

DISCUSSION AND CONCLUSION

The epidemiological features of the COVID-19 are still unknown, and the number of total cases and the number of total deaths varies day by day [7]. When the rapid spread and serious consequences of the disease were observed, precautions were taken and positive cases began to be recorded after the second half of January. Because the number of total cases and the number of total deaths are used for examining the course of the outbreak, modeling these indicators is a major issue. The model results are valuable for determining the appropriate prevention strategies. Due to the worldwide interest in this area, COVID-19 related literature has increased significantly [8]. In this study, the rising number of cases and deaths has been utilized to group countries by considering population size. Thus, it has been revealed which countries are similar or different in terms of these features. In this sense, the policies implemented by countries in the same group for the pandemic are similar. When the results of the study were evaluated, it was determined that in Aruba, Bahrain, French Guiana, Israel, Qatar; the average of total cases per population was higher compared to other countries.

The limitation of this study is that more detailed clustering could be made by adding different socio-demographic characteristics of the countries. In this study we did not take the characteristics of the countries. The results of this study will be of great importance, when showing the differences among countries in terms of total deaths and cases / 1Million population. This information will enable governments to implement suitable measures for subsequent similar situations. Cluster results could be useful to a variety of different policies.

References

1. Celik S, Ankarali H, Pasin O. Modelling of Covid-19 Outbreak Indicators In China Between January and June. Disaster Med Public Health Prep. 2020; Sep 9:1–20.
2. Zarikas V, Poulopoulos SG, Gareiou Z, Zervas E. Clustering analysis of countries using the COVID-19 cases dataset. Data in Brief. 2020;31:1–8.
3. Pasin Ö, Ankarali H. Usage of Cluster Algorithms in Health Studies: An Application. Turkiye Klinikleri J Med Sci. 2016;36(1):40–52.
4. MacQueen J. Some methods for classification and analysis of multivariate observations. In Procedings of the Fifth Symposium on Math, Statistics and Probability. California: 1967.
5. Pasin Ö, Ankarali H. Comparison of EM and Two-Step Cluster Method for Mixed Data: An Application. Int J Med Sci Clin Invent. 2017;4(3):2768–2773.

6. COVID-19 coronavirus pandemic [Cited 2020 September 11]. Available from: https://www.worldometers.info/coronavirus/.
7. Chirico F, Nucera G, Magnavita N. Estimating case fatality ratio during COVID-19 epidemics: Pitfalls and alternatives. J Infect Dev Ctries. 2020;14(5):438–439. doi:10.3855/jidc.12787.
8. Chirico F, Teixeira da Silva JA, Magnavita N. "Questionable" peer review in the publishing pandemic during the time of COVID-19: implications for policy makers and stakeholders. Croatian Med J. 2020;61(3):300–301. doi: 10.3325/cmj.2020.61.300.