Identifying Global Patterns for COVID-19 using Cluster Analysis

UNEB GAZDER
Department of Civil Engineering
University of Bahrain
Sakhir 32038
BAHRAIN
https://orcid.org/0000-0002-9445-9570

Abstract: - Clustering is an effective technique for unsupervised classification of data. K-means clustering is considered one of the most effective approaches for partitioning the data. In this study, k-means clustering was applied using squared Euclidean distance metric for identifying temporal patterns in the global data of infected and death cases of COVID-19. It was found that there was a significant shift in infected cases on 13, 20 and 27 March 2020, the last being for decrease in cases. The increase points could be attributed to increase testing and surveillance and spread of viruses across borders. On the other hand, the decrease in infected cases could be attributed to the closure of schools, businesses, sporting events and travel activities. Death cases mostly follow the increasing trend shown by infected cases with an approximate lag of 7 days. However, a breakpoint with increase in death cases was observed on 4 April 2020 which could be attributed to falsified medicines and equipment. A reduction in death cases is observed recently with possible explanation being increased knowledge for treating the infected persons and managing the health care facilities.

Key-Words: - unsupervised learning; infectious disease; COVID-19; clusters; temporal patterns

Received: May 17, 2024. Revised: April 12, 2025. Accepted: May 9, 2025. Published: August 4, 2025.

1 Introduction

Clustering is an unsupervised learning algorithm which is fed with features belonging to unclassified objects. The algorithm identifies the patterns and assigns the appropriate classes to features [1]. K-means clustering is considered as the most efficient mechanism to sort large amounts of data without any need for priori-information [2]. In case of COVID-19, researchers have reported sudden significant changes in rates of infections and deaths for specific countries [3]. In spite this fact, other researchers have developed time series models for COVID-19 data without considering these breakpoints in the dataset. This has resulted in models with low accuracy and/or increased complexity [4].

Considering the above-mentioned gaps in the literature, this study aims to apply clustering analysis to identify the breakpoints in the COVID-19 pandemic which have resulted in a shift in its trend. The clusters can be used to update the time series models, developed thus far, and help to visualize the effects of different measures which have been taken globally to handle COVID-19. Following objectives have been set to meet the goal of this study.

- To develop clusters to for daily infected and death cases of COVID-19 globally
- To identify the temporal breakpoints for each cluster
- Link the breakpoints with the incidents related to COVID-19 around the world

2. Dataset

COVID-19 data was collected for all countries of the world, for periods ranging from 31 December 2019 to 26 May 2020. The data was collected from European Centre for Disease Prevention and Control (https://www.ecdc.europa.eu). This center provides data for infected and death cases for each day in all countries of the world. Figure 1 presents the data collected from their website. The figure shows that the data shows varying trends in specific time periods. For example, the initial period is relatively straight then there is a steep upward gradient, and some seasonal patterns could be observed towards the end. This further justifies the need for identifying these temporal breakpoints in data with the application of clustering analysis. This would be the first time that global COVID-19 data has been subjected to such analysis. Other researchers have focused on modeling data from specific countries.

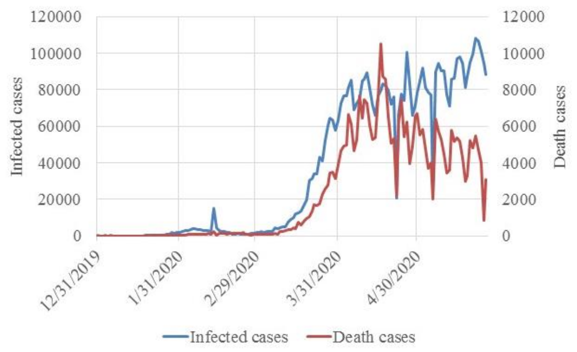

Fig.1. Available dataset

3. Clustering Method

There are two types of clustering algorithms which can be utilized for unsupervised learning of data, namely, hierarchical and partitioning. K-means clustering is one of the partitioning methods [5]. In this method, feature space is divided into classes maximizing the distance between their centroids, where 'k' refers to the number of classes or clusters [6]. The algorithm starts with a set of random initial centroids and data points are assigned to the nearest centroid. Centroids are, then, recomputed as the arithmetic mean of points assigned to that cluster [7]. The squared Euclidean distance was used as the metric for developing the clusters and defining their accuracy, which can also be referred to as Mean Square Error (MSE) [8].

There were 148 samples available in the dataset, out of which 103 were randomly selected for developing the clusters while remaining were kept for testing the accuracy of the clusters. The optimum number of clusters were found by observing the reduction in the average squared Euclidean distance, starting with the simplest case of 2 clusters. If adding an extra cluster did not reduce the distance metric significantly then the algorithm was stopped there. Analysis of Variance (ANOVA) was also performed on the training and test samples to investigate if the clusters had a significant impact on the distribution of data [9]. Each cluster can be considered as a separate distribution or dataset for modeling.

4 Results

4.1 Clustering for infected cases

It was found that four clusters were optimum for dividing the temporal series of infected cases. The change in the squared Euclidean distance metric with respect to number of clusters is given in figure 2. It can be observed that there is no significant change in the distance metric when the number of clusters is increased from four to five.

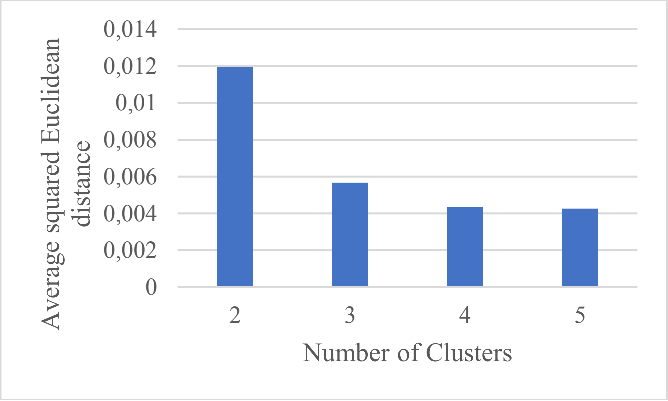

Fig. 2. Determining optimum number of cluster for infected cases

The statistics of clusters for infected cases are given in table 1. The clusters are arranged in ascending order, hence, cluster 1 has the lowest mean and cluster 4 has the highest mean for infected cases. The accuracy of the proposed clusters in terms of average Euclidean distance for training cases (103) was 0.004, while that for test cases (45) was 0.005.

Table 1. Cluster statistics for infected cases

Parameter	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Minimum	0	57810	9021	24297
Maximum	7535	107909	20722	52224
Mean	1553	81487	14362	36259
Standard deviation	1588	11602	4081	8719

Table 2 shows the parameters of ANOVA test performed on the training and test cases with the proposed clusters. The f-statistic was found to be significant for both datasets, so it could be concluded that the proposed classification of clusters has a significant impact on the distribution of data for training as well as test data.

Table 2. ANOVA for clusters of infected cases

Parameter	Training	Test
Sum of squares between clusters	2E+13	1E+11
Degrees of freedom	3	3
Sum of squares within clusters	6E+09	2E+09
Degrees of freedom	99	41
F	84213	472
p value	0.00	0.00

Figure 3 shows the temporal occurrence of each cluster with the breakpoints. It should be noted that there were few outliers between each break point which were ignored, and the general trend of the data is plotted. The first breakpoint occurs on 12 March 2020, after which the infected cases increase significantly and the points from cluster 3 started to occur. The trend goes to further growth on 20th March 2020 where points from cluster 4 appear. There is a sharp decline observed on 27 March 2020 which resulted in points from cluster 2 which continues till 26 May 2020.

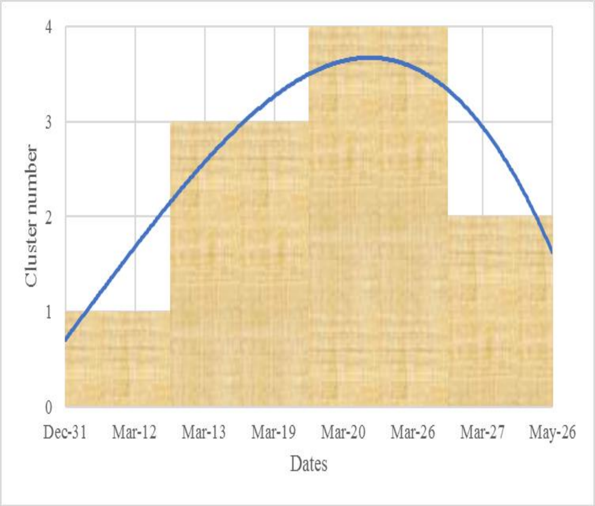

Fig. 3. Infected cases' cluster temporal trend

4.2 Clustering for Death Cases

Similar to infected cases, four clusters were also found optimum for dividing the temporal series of death cases. The change in the squared Euclidean distance metric with respect to number of clusters is given in figure 4. Although, there is a change in the distance metric when the number of clusters are increased from four to five, but the number of points in the fifth cluster were less than 5% of the total dataset. Therefore, four clusters were considered optimum for meaningful evaluation of data.

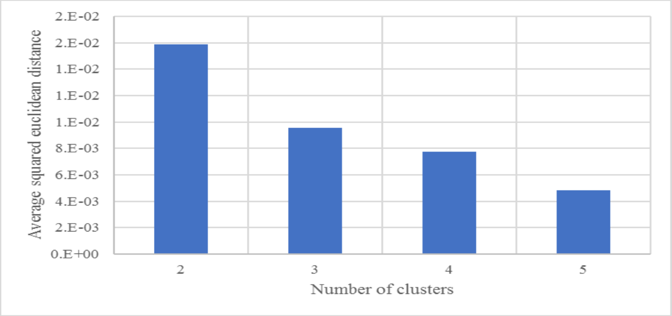

Fig. 4. Determining optimum number of cluster for death cases

The statistics of clusters for death cases are given in table 3. Cluster 1 has the lowest mean and standard deviation while cluster 4 has the highest of these values. The accuracy of the proposed clusters in terms of average Euclidean distance for training cases (103) was 0.004, while that for test cases (45) was 0.008.

Table 3. Cluster statistics for death cases

Parameter	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Minimum	0	800	2750	5036
Maximum	746	2573	4956	10520
Mean	91	1596	3984	6326
Standard deviation	133	605	701	1234

ANOVA test was performed on the training and test datasets, with the proposed clusters, is shown in table 4. The f-statistic was found to be significant for both datasets, so it could be concluded that the proposed classification of clusters were able to classify the datasets with significant variability.

Table 4. ANOVA for death cases

Parameter	Training	Test
Sum of squares between clusters	9E+10	468184454
Degrees of freedom	2	2
Sum of squares within clusters	39136725	39219753
Degrees of freedom	100	42
F	81105	163
p value	0.00	0.00

The temporal distribution of these clusters (in figure 5) shows that the breakpoints occur that 18 and 27 March and 4 April 2020, with a rise in the number of death cases. However, there is a decrease in the trend on May 17 which has continued till now. Unlike infected cases, death cases have a shown a more gradual change in trends on the breakpoints.

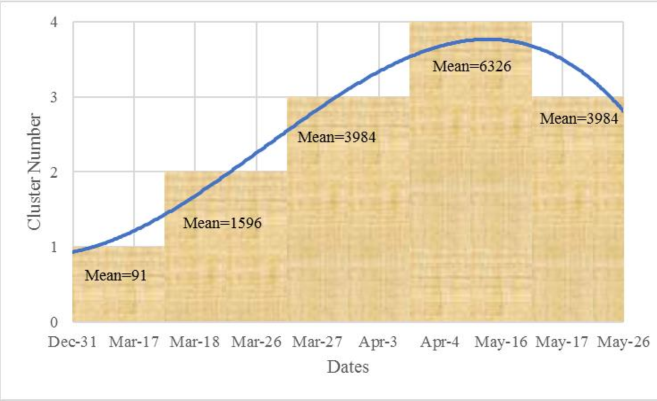

Fig. 5. Death cases' cluster temporal trend

5. Discussion

Comparison of figures 3 and 5 shows that breakpoints for increase in death cases generally follow those for infected cases with an approximate gap of 7-10 days. This is understandable since a portion of infected people are in the death cases, so drop or increase in the former will result in the same trend for the later with some lag. There is a decrease in the trend of infected cases since April 2020 which has also contributed towards the decrease in death cases in May 2020. This has been predicted by some researchers for specific countries, such as Italy, which is amongst the most infected countries [10] (Ding et al., 2020). However, these predictions were made based upon a time series models which do not incorporate the factors which could affect the change in trend.

The World Health Organization (WHO) has been publishing daily situational reports for COVID-19 data since January 2020. In addition to that, reports have also been published by ECDC on a periodic basis. These are the most authentic sources of information from a global perspective. Hence, these reports were consulted to investigate the events around the breakpoints found in this study. The details are provided in table 5.

Table 5 shows that the initial increases in the infected cases could be due to the increased surveillance for the disease including development and rigorous application of tests. It could also be due to the spread of virus in other parts of the world including Europe and the USA which resulted in a shortage of healthcare facilities. Alternatively, the reduction in the trend of infected cases can be attributed to greater efforts and investment in the combat of disease through lockdowns, work-from-home and provision of better facilities and funding. Apart from being a response to the infected cases, the increase in death cases could be attributed to the arrival of falsified medications and equipment. On the other hand, the decrease in death cases could be attributed to better management of health facilities and infected people which was a result of continuous research efforts.

6. Conclusions

This study was focused on determining the important temporal clusters for COVID-19 worldwide representing significant increase/decrease of infected and death cases. An unsupervised learning approach of cluster analysis was applied using K-means algorithm with squared Euclidean distance metric.

It was found that infected and death cases' data can be divided in to four distinguishable clusters. The temporal arrangement of these clusters showed a decrease in trend for infected cases since April 2020 and that for death cases since 17 May 2020. The increase in infected cases changed significantly on 13 and 20 March 2020 while a significant decrease started to appear since 27 March 2020. The possible reasons for increase could be increased surveillance and testing, and spread of virus to multiple countries before travel restrictions were in place. The decrease could be attributed to severe restrictions on work place, education and travel activities and more funding provided, for combating COVID-19, through United Nations and other international agencies. Trend in death cases has mostly followed that for infected cases in terms of increase. However, the breakpoint for increase in death cases from 4 April 2020 could be attributed to the floating of falsified medicines and equipment. A decrease in death cases have started to appear since 17 May 2020 which could be attributed to the development of better guidelines for handling of virus, infected patients and health care facilities.

The findings of this study could be beneficial for other researchers as they could study the trend of COVID-19 within each cluster to get more insights. It is recommended to analyze and model the clusters as separate distributions for avoiding over or underestimation in specific periods.

Table 5. Temporal Breakpoints and their possible reasons

Date of breakpoint	**Change in trend**	**Possible events**	**Sources**
13 March 2020	Significant increase in infected cases	• Spread COVID-19 became more prominent outside China with an increase of 13 times in the first two weeks of March • Higher coordination at the international level for surveillance of COVID-19 • Several countries in EU/EEA and UK reported being on high risk of exceeding the capacity of their health care system	[10, [11]
20 March 2020	Increase in infected cases	• A new protocol, to investigate the extent of COVID-19 infection in the population, has been developed	[12]
27 March 2020	Significant Decrease in	• The United Nations launched a US$2 billion	[13], [14]

	infected cases	COVID-19 Global Humanitarian Response Plan to support the world's most vulnerable countries • Many countries closed down schools and businesses and major sports events • Severe restrictions on gathering of people and travel across borders	
18 March 2020	Increase in death cases	• In response to increase in infected cases	N/A
27 March 2020	Increase in death cases	• In response to increase in infected cases	N/A
4 April 2020	Increase in death cases	• WHO reported use of in vitro falsified equipment and medicines for detection and treatment of COVID-19 • With spread of virus increasing in developing countries; there was a greater need for medical equipment, testing kits, personal protective equipment for health workers and enhancing health systems capacities	[15], [16]
17 May 2020	Decrease in death cases	• In response to decrease in infected cases • More clear guidelines were available for disinfection and protection against COVID-19	[17], [18]

Acknowledgments

The author acknowledges the cooperation from Dr. Ayman Alkhazraji for identifying the data sources.

Declaration

This paper was published as a preprint, and can be found with the following reference:
Uneb Gazder, "Identifying Global Temporal Patterns for COVID-19 Using Cluster Analysis", May 2021, DOI: 10.13140/RG.2.2.22940.62083, LicenseCC0, available on https://www.researchgate.net/publication/351283194_Identifying_Global_Temporal_Patterns_for_COVID-19_Using_Cluster_Analysis

References

[1] Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., Constrained k-means clustering with background knowledge, In *Icml*. 1, June, 2001, 577-584).

[2] Capó, M., Pérez, A., Lozano, J. A., An efficient approximation to the K-means clustering for massive data. *Knowledge-Based Systems*, 117, 2017, 56-69.

[3] Abdulrahman, I. K., SimCOVID: An Open-Source Simulation Program for the COVID-19 Outbreak. [29 May 2020]. Doi: https://doi.org/10.1101/2020.04.13.20063354 medRxiv .

[4] Bayyurt, L., and Bayyurt, B., Forecasting of COVID-19 Cases and Deaths Using ARIMA Models. [29 May 2020]. https://doi.org/10.1101/2020.04.17.20069237.

[5] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., et al., A review of clustering techniques and developments. *Neurocomputing*, 267, 2017, 664-681.

[6] Jamal, A., Handayani, A., Septiandri, A. A., Ripmiatin, E., Effendi, Y., Dimensionality reduction using PCA and K-Means clustering for breast cancer prediction, *LONTAR KOMPUTER: Jurnal Ilmiah Teknologi Informasi*, 2018, 192-201.

[7] Chakraborty, S., Das, S., k− Means clustering with a new divergence-based distance metric: Convergence and performance analysis, *Pattern Recognition Letters*, 100, 2017, 67-73.

[8] Oyelade, O. J., Oladipupo, O. O., Obagbuwa, I. C., Application of k means clustering algorithm for prediction of students academic performance. [29 May 2020] arXiv preprint arXiv:1002.2425.

[9] Cardinal, R. N., Aitken, M. R., *ANOVA for the behavioral sciences researcher*, Psychology Press. Mahwah, New Jersey USA, 2013.

[10] Ding, G., Li, X., Shen, Y., Fan, J., Brief Analysis of the ARIMA model on the COVID-19 in Italy, [29 May 2020] doi: https://doi.org/10.1101/2020.04.08.20058636, medRxiv.

[11] ECDC, *Risk assessment report*, 12 March 2020. [29 May 2020]. https://www.ecdc.europa.eu/en/covid-19-pandemic

[12] WHO, *COVID-19 situation report*, 11 March 2020, [29 May 2020]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[13] ECDC, *Communicable disease threats report*, 28 March 2020, [29 May 2020]. https://www.ecdc.europa.eu/en/covid-19-pandemic

[14] WHO, *COVID-19 situation report*, 12 March 2020, [29 May 2020]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[15] WHO, *COVID-19 situation report*, 19 March 2020, [29 May 2020]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports\

[16] ECDC, *ECDC communicable disease threats report*, 4 April 2020, [29 May 2020]. https://www.ecdc.europa.eu/en/covid-19-pandemic

[17] WHO, *COVID-19 situation report*, 26 March 2020, [29 May 2020]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[18] ECDC, *ECDC technical report*, 13 May 2020, [29 May 2020]. https://www.ecdc.europa.eu/en/covid-19-pandemic

[19] WHO, *COVID-19 situation report*, 3 April 2020, [29 May 2020]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[20] WHO, *COVID-19 situation report*, 16 May 2020, [29 May 2020]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports