Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Clusterings of COVID-19 data for knowledge discovery using c-means and fuzzy c-means

Asif Afzal a,∗, Zahid Ansari b, Saad Alshahrani c, Arun K. Raj d, Mohamed Saheer Kuruniyan e, C. Ahamed Saleel f, Kottakkaran Sooppy Nisar g

a Department of Mechanical Engineering, P. A. College of Engineering (Affiliated to Visvesvaraya Technological University, Belagavi), Mangaluru, India
b Electrical Engineering Section, University Polytechnic, Aligarh Muslim University, Aligarh, India
c Department of Mechanical Engineering, College of Engineering, King Khalid University, PO Box 394, Abha 61421, Kingdom of Saudi Arabia
d Department of Mechanical Engineering, Indian Institute of Technology Bombay, Maharashtra 400076, India
e Department of Dental Technology, College of Applied Medical Science, King Khalid University, Asir-Abha, Kingdom of Saudi Arabia
f Department of Mathematics, College of Arts and Sciences, Wadi Aldawaser, Prince Sattam bin Abdulaziz University, Saudi Arabia
g Electrical Engineering Section, University Polytechnic, Aligarh Muslim University, Aligarh, India

∗ Corresponding authors.

E-mail addresses: asif_mech@pace.edu.in (A. Afzal), saadayed@kku.edu.sa (S. Alshahrani), mkurunian@kku.edu.sa (M. Saheer Kuruniyan), aveetil@kku.edu.sa (C. Ahamed Saleel), n.sooppy@psau.edu.sa (K.S. Nisar).

https://doi.org/10.1016/j.rinp.2021.104639

Received 11 July 2021; Received in revised form 30 July 2021; Accepted 31 July 2021

ABSTRACT

In this work, the partitioning clustering of COVID-19 data using c-Means (cM) and Fuzzy c-Means (Fc-M) algorithms is carried out. Based on the data available from January 2020 with respect to location, i.e., longitude and latitude of the globe, the confirmed daily cases, recoveries, and deaths are clustered. The performance and validity indices of the clusters formed are analyzed to assess the quality of clusters. The validity indices to understand all the COVID-19 clusters’ quality are analysed based on the Zahid SC (Separation Compaction) index, Xie-Beni Index, Fukuyama-Sugeno Index, Validity function, PC (performance coefficient), and CE (entropy) indexes. The analysis results pointed out that five clusters were identified as a major centroid where the pandemic looks concentrated. Additionally, the observations revealed that mainly the pandemic is distributed easily at any global location, and there are several centroids of COVID-19, which primarily act as epicentres. However, the three main COVID-19 clusters identified are 1) cases with value <50,000, 2) cases with a value between 0.1 million to 2 million, and 3) cases above 2 million. These centroids are located in the US, Brazil, and India, where the rest of the small clusters of the pandemic look oriented. Furthermore, the Fc-M technique seems to provide a much better cluster than the c-M algorithm.

Introduction

Coronavirus disease 2019 (COVID-19) is defined as a disease caused by a new strain of coronavirus, which belongs to a large family of viruses known to cause disease ranging from the common cold to more severe illnesses such as Middle East Respiratory Syndrome (MERS-CoV). It is now called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; formerly known as 2019-nCoV), which was identified as the infectious agent causing an outbreak of respiratory illness cases in Wuhan City, Hubei Province, China, in December 2019 [1]. Since January 2020, the COVID-19 confirmed cases and death had been detected outside Wuhan. Currently, the number of reported cases and death have been reported across the world. On March 11, 2020, the World Health Organization (WHO) declared that the coronavirus diseases 2019 (COVID-19) as a global pandemic, and it is an international concern of public health emergency. As of October 22, 2020, there were 41,500,769 confirmed cases, 1,136,526 deaths, and 30,918,817 recovered cases of COVID-19 globally [2]. The high disease transmission rate of COVID-19 and vaccines’ unavailability cause COVID-19 to become a global pandemic. COVID-19 virus is primarily transmitted between people through respiratory droplets inhalation from infected individuals and direct contact with contaminated surfaces. Therefore, the transmission of the COVID-19 virus can occur by direct contact with infected people when an infected person generates droplets by coughing, sneezing, or exhaling. Moreover, the virus can be transmitted by indirect contact with contaminated surfaces in the immediate environment or with the
infected person’s fomites \([3–5]\).

Spreading the novel coronavirus disease COVID-19 has become a challenging global issue. Therefore, to control the spread of COVID-19, the governments have been implementing different control measures such as lockdowns to maintain social distance, curfew, preventing crowded events, and using the precautions such as face masks while in public. These procedures reduced the COVID-19 transmission rate. However, the full lockdown has a negative impact on the economy and may cause a significant financial crisis within a short time. Therefore, there should be a suitable balance between governmental policy lockdown and healthy free conditions. As the COVID-19 is transmitting easily between people in many parts of the world simultaneously, and the outbreak spread quickly across the world, it is important to understand its dynamics and predications. Uncertainties around the spread of COVID-19 have led many scientists, clinicians, and engineers to investigate and suggest many mathematical models, which are necessary to quantify COVID-19.

The study of infectious disease spread mechanism can be done by many models such as Susceptible-Infected-Recovered (SIR), susceptible individuals, asymptomatic infected, symptomatic infected, recovered, and deceased (SEIRD), and Susceptible, Exposed, Infectious, Hospitalized, Dead (SEIHRD). These models use differential equation systems or difference equations (discrete-time) for describing population dynamics concerning infection \([3–8]\). The SIR-model is one of the simplest models, and it divided the entire population into three groups or compartments: the number of susceptible people \((S)\), the number of infectious people \((I)\), and the number of removed people \((R)\) \([8]\). In SIR-model, each person is categorized as susceptible, infectious, or removed, where susceptible indicates healthy person and removed indicates either recovered or dead as a result of the COVID-19. Each person can transition from one group or state (susceptible, infectious, or removed) to another. That is, each person has a unique Markov chain associated with their classification \([9]\).

Mathematical models have been used successfully in understanding the population dynamics of COVID-19 in several countries. Kucharski et al. \([10]\) have developed a stochastic transmission mathematical model combined with data on cases COVID-19. All positive cases of Wuhan, China were considered by this study until March 5, 2020. In another study by Ndaireu et al. \([11]\), an autonomous mathematical model system was performed to investigate the spread of COVID-19 in Wuhan city, China. Another study that focuses on controlling the status of COVID-19 of Wuhan city research was performed by Prem et al. \([12]\) using mathematical modeling. In another work, using the isolation method to investigate the effective procedure of COVID-19 disease was done by Hellewell et al. \([13]\). Mizumoto and Chowell \([14]\) investigated the transmission dynamics of COVID-19 in the Diamond Princess Cruises Ship. To study the nature of infectious diseases. In order to study the nature of the infectious disease, Liu et al. \([15]\) have analyzed the numerical value of the basic reproduction number of COVID-19. Fanelli and Piazza \([16]\) have analyzed and forecasted the nature of COVID-19 spreading in China, Italy, and France until March 2020 by developing mathematical modelling. Chakraborty and Ghosh \([17]\) have considered a hybrid autoregressive integrated moving average (ARIMA) and Wavelet-based forecasting of the number of daily confirmed cases for Canada, France, India, South Korea, and the UK. In another study, Ribeiro et al. \([18]\) have used some stochastic-based regression models to evaluate the task of time series forecasting with one, three, and six days ahead of the COVID-19 cumulative confirmed cases in ten Brazilian states with a high daily incidence.

Yang et al. \([19]\) have developed a mathematical model using a modified susceptible exposed infectious recovered (SEIR) technique to study the epidemic development of COVID-19 in China. They have predicted the magnitude and timing of the COVID-19 epidemic peaks and sizes under various intervention strategies. Peng et al. \([20]\) performed a study on estimating the epidemic of COVID-19 in China by implementing the SEIR model. They have made predictions on the inflection point and possible ending time of confirmed COVID-19 cases for five different regions. Li et al. \([21]\) have used the traditional SEIR model to predict the COVID-19 transmission dynamics in Wuhan and Beijing. They have estimated the effect of lockdowns on the COVID-19 transmission dynamics for these cities. Wu et al. \([22]\) have used an SEIR model to predict the disease transmission dynamics in China and forecasted the national and global spread of COVID-19, based on the basic reproductive number and the outbreak size in Wuhan from December 31, 2019, to January 28, 2020. Tang et al. \([23]\) have developed a mathematical model to determine the potential and severity of an outbreak of COVID-19. They have implemented the disease’s clinical progression, the individuals’ epidemiological status, and intervention measures to their model.

The previous and recent works on COVID-19 modelling and analysis of its spread and further prediction indicate a widespread list of methods. However, the knowledge discovery of the COVID-19 data using partition-based clustering methods has a lot of scopes. Hence in this study, the clusters of COVID-19 based on longitude, latitude, confirmed cases, deaths, and recoveries based on data available from the initial days are formed using c-Means (c-M) and Fuzzy c-Means (Fc-M) algorithm. The optimum number of clusters formed, and their performance and validity indexes are analyzed in detail.

Clustering of COVID-19 data

c-M clustering

In this algorithm, the cluster’s center is obtained by minimizing a cost function with a dissimilarity measure. For a given data points set \(m\) represented as \(X = (x_1, \ldots, x_m)\). This set is a vector with \(n\)-dimension in which the c-M algorithm partitions the data into \(c\) clusters to minimize the dissimilarity \(J(X, \{V\})\). The dissimilarity is chosen by the Euclidean distance, which is the sum of squares. The objective function is mathematically shown as in Eq. (1). The performance Index for c-M is thus given by:

\[
J(X, V) = \sum_{j=1}^{c} \frac{1}{n} \sum_{i=1}^{n} u_{ij}^2 d^2(x_i, v_j)
\]

where,

\[
J(x_i, v_j) = \sum_{i=1}^{n} u_{ij}^2 d^2(x_i, v_j)
\]

In cluster, \(C_k\) is objective function

When, \(x_i \in C_k\), the membership \(u_{ik} = 1\), else it is 0.

The distance \(d^2(x_i, v_j)\) is between \(v_j\) and \(x_i\) which is given by:

\[
d^2(x_i, v_j) = \sum_{k=1}^{n} x_{ik}^2 - v_{jk}^2
\]

The above equation is the Euclidean distance between the centre of cluster and the data point.

The \(k^{th}\) dimensional value of \(x_i\) is given by \(x_{ik}\)

The \(k^{th}\) dimensional value of \(v_j\) is given by \(v_{jk}\)

The membership function \(V = \{v_1, v_2, \ldots, v_c\}\) for cluster centres are defined as:

\[
u_{ij} = \begin{cases} 
1 & \text{if } d^2(x_i, v_j) < d^2(x_i, v_j'), j' \neq j, \forall j' = 1, \ldots, k \\
0 & \text{otherwise}
\end{cases}
\]

According to Eq. (1), the \(x_i\) point belonging to \(C_j\) cluster having a nearest \(v_j\) cluster center should be assigned. Then the optimal \(v_j\) cluster center, which lessens the mean of cluster data points in \(j\) cluster is assigned with \(U = \{|u_{ij}|\}\) membership using Eq. (4). In Algorithm 1, the step-by-step procedure for the partition of COVID-19 data points using c-M is mentioned. In Fig. 1 the c-Means algorithm is shown to form...
clusters before and after its application.

\[ v_j = \frac{1}{|c_j|} \sum_{i \in c_j} x_i \]  \hspace{1cm} (5)

**Algorithm 1: Clustering by c-Means**

**Input:** \( X = \{x_1, \ldots, x_m\} \) data point set with clusters \( k \)

**Output:** \( U = \{u_{ij}\} \) (membership matrix), \( C = \{c_1, \ldots, c_k\} \) (clusters), and \( V = \{v_1, \ldots, v_k\} \) as centroids

- Select \( k \) points randomly to initialize the \( V = \{v_1, \ldots, v_k\} \) centers
- for each cluster in number of clusters \( k \) do
  - Calculate the membership value using Eq. (2) by allocating \( x_i \) to the nearest centre
  - The objective \( J(X, V) \) function is calculated using Eq. (1). Stop the iterations if it reduces below a threshold \( \epsilon \).
  - Update the centre of cluster using Eq. (3)
- end for

- till the centroids don’t change

**Fc-M clustering**

Fc-M clustering is widely used in the data mining technique. Each data point in this technique belongs to a cluster by a membership degree. As with c-M clustering, Fc-M clustering depends on the cost function of a different measure being minimized. Fc-M integrates a simple concept of c-M, whereas each data point in Fc-M belongs to a cluster to some extent. While each data point of c-M either belongs to or does not belong to a specific cluster. Thus, Fc-M performs fuzzy clusters so that a certain data point may belong to several clusters’ membership degrees between 0 and 1. However, when trying to partition the data, Fc-M still uses an objective function to be minimized. The algorithm uses a \( q \) parameter of the fuzziness index that specifies the clusters’ amount of fuzziness. The algorithm operates like a crisp partitioning algorithm when the value of \( q \) reaches 1. The increase in the \( q \) value contributes to further cluster overlap.

Again, data points \( X = \{x_i \mid i = 1 \ldots m\} \) where each \( x_i = (x'_i, x'_1, \ldots, x'_n \mid i = 1 \ldots m) \). The \( n \)-dimensional vector of centres of clusters is \( V = \{v_j \mid j = 1 \ldots c\} \) and each \( v_j = (v'_j, v'_1, \ldots, v'_n \mid j = 1 \ldots c) \). \( u_{ij} \) is the membership of data points with

\[ u_{ij} \in [0,1]\forall i = 1 \ldots m \text{ and } \forall j = 1 \ldots c \]

\( U = \{u_{ij}\} \) is the Fc-M matrix that represents the data point allocation and satisfies the condition:

\[ \sum_{j=1}^{c} u_{ij} = 1, \quad \forall i = 1 \ldots m \]

\[ 0 < \sum_{j=1}^{c} u_{ij} < m, \quad \forall j = 1 \ldots c \]  \hspace{1cm} (6)

In this algorithm the performance index as in c-M is denoted similarly by \( J(U, V, X) \) i.e. weighted sum of data point distances and respective cluster point centres as:

\[ J(U, V, X) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} u_{ij} d_{ij}^q(v_i, v_j) \]  \hspace{1cm} (7)

Here \( q \in [1, \infty] \) is the clustering fuzziness index and distance \( d_{ij}^q(v_i, v_j) \) is between \( v_i \) and \( v_j \).

The updating of clusters centre occurs using Eq. (8).

\[ v_j = \frac{\sum_{i=1}^{m} u_{ij}^q v_i}{\sum_{i=1}^{m} u_{ij}^q} \]  \hspace{1cm} (8)

The membership of the cluster points is calculated using Eq. (9). The entire Fc-M algorithm in the COVID-19 clustering is mentioned in Algorithm 2.

**Algorithm 2: Clustering by Fc-Means**

**Input:** \( X = \{x_1, \ldots, x_m\} \) data point set with clusters \( k \)

**Output:** \( U = \{u_{ij}\} \) (membership matrix), \( C = \{c_1, \ldots, c_k\} \) (clusters), and \( V = \{v_1, \ldots, v_k\} \) as centroids, and Fuzzy membership matrix

- Select \( k \) points randomly to initialize the \( V = \{v_1, \ldots, v_k\} \) centres
- for each cluster in number of clusters \( k \) do
  - Calculate the membership value using Eq. (8) by allocating \( x_i \) to the nearest centre
  - The objective \( J(U, V, X) \) function is calculated using Eq. (7). Stop the iterations if it reduces below a threshold \( \epsilon \).
  - Update the centre of cluster using Eq. (9)
- end for

- till the centroids don’t change

**Validity indices**

For the set of COVID-19 data \( X \), the optimal number of clusters is calculated using validity function \( S \) representing the compactness to separation ratio is given by [24,25]:

\[ S = \frac{\sum_{i=1}^{m} \sum_{j=1}^{c} u_{ij}^q v_i - v_j^q}{m \min_{i \neq j} \sum_{i=1}^{m} \| v_i - v_j \|} \]

For each \( \epsilon = \min, \ldots, \max \)

Reducing the output index is normally accomplished by changing the membership grading of data points and cluster centres differently before convergence is achieved. This index is based on the sum of the squared error criterion. The cluster centres will be updated as follows during each iteration [26,27]:

\[ v_j = \frac{\sum_{i=1}^{m} u_{ij}^q v_i}{\sum_{i=1}^{m} u_{ij}^q} \]  \hspace{1cm} (11)

Intra-cluster membership ratio to inter-cluster separation is the Xie-Beni (XB) validity index. This index is based on the objective function \( J \) and the square of the cluster centres’ minimum distances. This is given as [28]:

\[ XB = \frac{\sum_{i=1}^{m} \sum_{j=1}^{c} u_{ij}^q d^2(s, v_j)}{m \bar{d}^2_{\text{min}}} \]  \hspace{1cm} (12)

Here the term \( \bar{d}^2_{\text{min}} \) is the Euclidean distance between the centres of clusters which is given by:

\[ \bar{d}^2_{\text{min}} = \min_{i,j} \bar{d}^2(s, v_j) \]

The Fukuyama–Sugeno (FS) Index is given as:
\[ FS = \sum_{i,j=1}^{n} \mu_{ij}^2 \delta^2 \left( x_i, \delta \right) - \sum_{i,j=1}^{n} \mu_{ij} \delta^2 \left( y_j, \delta \right) \]  

(13)

The vector of cluster centres mean \( v \) is given by:

\[ v = \frac{\sum_{j=1}^{c} v_j}{c} \]

By considering geometrical properties of membership functions and data structure, the term \( \zeta_1 \) measures the ratio of fuzzy separation to fuzzy compactness. A large value of \( \zeta_1 \) implies well-separated and compact clusters. The term \( \zeta_2 \) also measures the ratio of fuzzy separation to fuzzy compactness, but only considers the fluid value of the membership. In order to obtain the compactness and the fuzzy separation, the term \( \zeta_2 \) uses a fuzzy union and a fuzzy intersection. For well separated fuzzy c-partitions, a small value in the numerator is needed, and the high value in the fuzzy denominator is a fuzzy c-partition. Therefore, the Zahid SC index is given as:

\[ SC = \zeta_1 - \zeta_2 \]

\[ \zeta_1 = \frac{\sum_{j=1}^{c} \sum_{i=1}^{m} \mu_{ij} \delta^2 \left( x_i, \delta \right)}{\sum_{j=1}^{c} \sum_{i=1}^{m} \mu_{ij} / \sum_{i=1}^{m} \mu_{ij}} \]

\[ \zeta_2 = \frac{\sum_{j=1}^{c} \sum_{i=1}^{m} \min \left( \mu_{ij}, \mu_{ik} \right)}{\sum_{j=1}^{c} \sum_{i=1}^{m} \max \left( \mu_{ij}, \mu_{ik} \right)} \]

(14)

The performance coefficient (PC) and entropy (CE) are respectively given as:

\[ PC = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^2 \]

(15)

\[ CE = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij} \log(\mu_{ij}) \]

(16)

To evaluate the overlapping clusters in Fuzzy partitioning, the PC index reduces with the increase in cluster size. CE is just another parameter to calculate the fuzziness of the cluster partition.

In the Silhouette index \( (S_i) \), the method calculates the silhouette width, the average silhouette width for each cluster, as well as the total data set for the average silhouette width. Eq. (17) is used to measure the silhouette width of the data point, where \( a_i \) is the average dissimilarity it has with every other data point in that cluster. \( b_i \) is a minimum of average dissimilarity of a data point with respect to every other point in other clusters.

\[ S_i = \frac{b_i - a_i}{\max(a_i, b_i)} \]

(17)

Results and discussion

The clustering using c-M and Fc-M algorithms of COVID-19 data from January 2020 has more than 67,000 entries with longitude, latitude, date, reported cases, recovered, and deaths as the attributes. The clusters are varied from 5 to 50 in separate runs of both the algorithms. The clusters are kept at a minimum of 5 and a maximum from 5 to 50. The optimal number of clusters is obtained based on the algorithms’ calculations, as mentioned in the previous section. The clusters of COVID-19 are based on 1) longitude 2) latitude 3) date 4) reported cases, 5) recovered, and 6) deaths. Later, the validity indices to understand all the COVID-19 clusters’ quality are analyzed based on the Zahid SC index, Xie-Beni Index, Fukuyama–Sugeno Index, Validity S function, PC, and CE indexes.

The cluster centers correspond to the longitude of the globe where the COVID-19 cases are centered are shown in Fig. 2 during different upper limits of cluster inputs, i.e., from 5 to 50. The centroids of COVID-19 longitudes are largely in the zone of \(-10^\circ \) to \(20^\circ \). Three of the centroids are placed at approximately \(-30^\circ \), and three more are \(50^\circ \). These all clusters are obtained from the c-M and Fc-M algorithms. Fig. 3 shows the latitude position of the COVID-19 centroids obtained using the c-M and Fc-M algorithms. Mainly the concentrations are in \(-50^\circ \) to \(-100^\circ \) indicating the affected areas are the American continent. The next three zones largely affected by the COVID-19 are found between \(0^\circ \) to \(30^\circ \), primarily the Indian subcontinent. These latitude positions indicate the centroids of the COVID-19 pandemic. Basically, the knowledge discovered from this analysis is that mainly the pandemic is distributed easily at any global location. At the same time, there are several centroids of COVID-19, which primarily act as epicenters. In such centroids of COVID-19, the newly added data is easily allocated with a cluster center. Fig. 4 shows that the centres of confirmed COVID-19 cases are depicted obtained from the c-M and Fc-M algorithms. Except for three centres the remaining centres are more than 50,000 values. The remaining indicate the US, India, and Brazil as COVID-19 centres separately. The trend is nearly similar for both the algorithms but the peaks are largely different. However, the Fc-M technique seems to provide a much better cluster than the c-M algorithm. Instead of 9 clusters shown here with centres having values in the range of \(<50,000 \) and \(>50,000 \), the partition can be made into three main clusters: 1) COVID-19 cases with value < 50,000, 2) cases with a value between 0.1 million to 2 million, and 3) cases above 2 million. The finding mainly here is the distribution of cases in these three cases where the COVID-19 confirmed cases fall.

The recovered cases shown in Fig. 5 has primarily two main centres, which can be extended with three centres, as in the previous case. Here
also the three main centres contributing to COVID-19 recovered cases are nearly the same. But in this case, the recovered cases are higher in number for cluster 6. Previously, the confirmed US soil cases were more while the Indian people have shown more recovery in the recovery rate. The rest, the other centres remain locally in the same range of <50,000.

In Fig. 6, the death cases due to the COVID-19 pandemic are of a very similar trend. The pattern of all the cases looks much similar. The confirmed, recovered, and death centres can be easily mapped, referring to Figs. 2 and 3. As the recovery is high for cluster number 6, the deaths pertinent to this cluster are also less than cluster number 4. In cluster centre 4, the deaths are more than 12,000. The centres from c-M and Fc-M nearly have the same partitioned clusters in the US and India region. These two are mainly the centres where the spread, propagation, and related recovery and deaths are concentrated. It is further required to cluster the US and Indian COVID-19 cases.

The optimal cluster validity values for c-M and Fc-M is provided in Fig. 7 with increasing cluster size. The trend of rising and falling using the c-M and Fc-M algorithms is quite separated, showing an abrupt cluster validity value. The cluster validity using Fc-M for the COVID-19 data is highest for a maximum cluster size of 50 and lower for lower cluster sizes (<25 maximum clusters). c-M has provided the highest value at 5 maximum clusters as it is much clear from the preceding results that lower cluster numbers represent a much appropriate scenario of COVID-19.

In Fig. 8, the performance index of the c-M algorithm with the increase of cluster size from 5 to 50 is presented. Eq. (1) is used to measure the performance index of c-M clustering. It is the sum of distances from the data points to the appropriate cluster centres. The output minimization of Index J is achieved by using Eqs. (5) and (8), respectively, to change the membership of the data points and cluster centres, before convergence is achieved. The performance index continuously reduces with the increase of cluster size, which is the indication of convergence.

The validity index S of c-M obtained using Eq. (10) with an increasing number of cluster sizes is shown in Fig. 9. The Silhouette index obtained using Eq. (17) is shown in the secondary axis of the same Fig. 9. The S validity index is highly fluctuating with the cluster number. There is no specific reason for the variations. However, the validity index is also low for smaller cluster sizes, indicating a lower number of optimal clusters. The values range from −1 to 1 for the Silhouette index. A value close to 1 means that a very suitable cluster is allocated to the data point. A value near zero means that the data point may also be allocated to another nearest cluster since both clusters are equidistant. If the value is similar to −1, the data is wrongly labelled and lies between the clusters. Our findings demonstrate that all technologies produce values from 0 to 1. Single link hierarchical clusters and leader clusters (with a threshold of dissimilarity β = 1.0) have closer values to 1 and therefore outdo the other techniques.

*Fig. 9 shows the performance index of Fc-M and its validity index S is represented with increasing cluster size. The cluster size’s performance index pattern is having the same pattern as we have observed with the c-M algorithm. We have estimated how many optimal clusters we have identified Index of validity (S), the ratio of the separation, and compactness given by Eq. (10). The validity index S is a minimum for the initial cluster size and fluctuates with peak values randomly between the clusters. The variations in the results have no role in the COVID-19 cluster formation. However, this index helps obtain the
optimum number of clusters, as discussed previously concerning Fig. 7. The partition entropy (PE) and coefficient (PC) index of the Fc-M algorithm is shown in Fig. 10. PC measures the overlapping clusters. PC’s main characteristics are the decreasing trend with the increasing cluster size, as also noted in Fig. 10. Another measure similar to the partition coefficient described by Bezdek is entropy classification for the cluster partition fuzziness. Only the elements of the membership matrix are used to measure partition and classification entropy. Both of them may lack a clear relation to their geometry.

The Xie Beni index variation is shown in Fig. 11. Xie Beni has proposed a well-known validity index. This index incorporates the features of the degree of membership and the geometry of the dataset. XB index calculates the average total separation between the clusters and intra-compactness. Smaller XB index values mean more compact clusters with greater spacing. When the cluster numbers are very high, the index decreases monotonically. The Zahid Separation Compaction (SC) index values determined using Eq. (14) can be found in Fig. 10 with the number of clusters as a function. This index shows a higher value for improved cluster consistency. Fig. 12 displays the index scores determined using Eq. (13) for Fukuyama Sugeno (FS) with the number of clusters as a function. Since FS reflects the difference between the cluster’s compaction and the separation between the clusters, a lower value of the index is better as shown in Fig. 13.

Comparative analysis

In Table 1 the most important quantitative analysis between the c-Means and Fc-Means algorithm results are comparatively shown. It can be seen that the optimum value of the c-Mean cluster is highest at the lowest cluster size while it reduces with increasing cluster size. The optimum validity is at the lowest cluster indicating the epicentre of the epidemic. In Fc-Mean the increasing number of cluster sizes shows an increasing cluster validity which has a direct effect on the fuzzy index in the clustering of COVID-19 data. The maximum validity obtained using Fc-Mean is at the largest cluster size. However, it can be mentioned that the c-Mean has an appropriate clustering technique which is helpful in the insight of the epicentres compared to the Fc-Mean algorithm predictions.
Conclusion

In this work, the clustering of COVID-19 data to discover useful patterns that can help understand the pandemic’s spread was carried out. For the clustering of data, c-Means and Fc-Means algorithms were employed. The c-Means and Fc-Means performance and validity indices obtained for the clusters formed are analyzed to assess the quality of clusters. The below conclusions are drawn from the clustering analysis:

1. The cluster’s optimum validity is obtained at the lowest cluster size using c-Means while the maximum validity obtained using Fc-Means is at the maximum cluster size.
2. The performance index using c-Means and Fc-Means obtained is of the same nature (reducing with increasing cluster size) while the validity index obtained using both though same was of very random nature with increasing cluster size.
3. The clustering of COVID-19 data from the available data revealed that there were five optimal clusters based on the location and the cases observed so far.
4. The three main COVID-19 clusters have been identified: 1) cases with value < 50,000, 2) cases with a value between 0.1 million to 2 million, and 3) cases above 2 million.
5. In cases where there were 0.1 to 2 million cases, the recovery was faster, and the number of deaths was lower. The first and third cases were mainly located in the US and Brazil, while the third is in the Indian sub-continent.
6. Even if the pandemic spreads faster, these three centres will remain as centroids. These three main centres need to adopt stringent rules to eradicate pandemic transmission.

Conclusion of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University, Saudi Arabia for funding this work through Research Group Program under Grant No: RGP 2/191/42.

References

[1] WHO. Coronavirus disease (COVID-19) situation reports n.d.
[2] Worldometers. COVID-19 Coronavirus Pandemic n.d.
[3] Panda SK, Ravichandran C, Hazarika B. Results on system of Atangana–Baleanu fractional order Willis aneurysm and nonlinear singularly perturbed boundary value problems. Chaos Solitons Fractals 2020;135:99863. https://doi.org/10.1016/j.chaos.2020.110390.
[4] Ravichandran C, Logeswari K, Jayal F. New results on existence in fractional integro-differential equations. Chaos Solitons Fractals 2019;125:194–200.
[5] Logeswari K, Ravichandran C NK. Mathematical model for spreading of COVID-19 virus with the Mittag–Leffler kernel. Numer Methods Partial Differ Equ 2020. https://doi.org/10.1002/num.22652.
[6] Hethcote HW. The mathematics of infectious diseases. SIAM Rev 2000;42(4):599–653. https://doi.org/10.1137/S0036144500371907.
[7] Bonyah E, Khan MA, Okosun KO, Islam S. A theoretical model for Zika virus transmission. PLoS One 2017;12:e0185540.
[8] Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc R Soc London Ser A 1927;115:700–21.
[9] Aleman DM, Wibisono TG, Schwartz B. Accounting for individual behaviors in a pandemic disease spread model. In: Proc. 2009 winter Simul. Conf., IEEE; 2009. p. 1977–85.
[10] Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis 2020.
[11] Ndairou F, Area I, Nieto JJ, Torres DFM. Mathematical modelling of COVID-19 transmission dynamics with a case study of Wuhan. Chaos Solitons Fractals 2020;135:109846. https://doi.org/10.1016/j.chaos.2020.109846.
[12] Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. Lancet Public Heal 2020.
[13] Hellewell J, Abbott S, Gimma A, Bosse N, Jarvis CI, Russell TW, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Global Health 2020;8(4):e488–96. https://doi.org/10.1016/S2214-109X(20)30074-7.
[14] Mizumoto K, Chowell G. Transmission potential of the novel coronavirus (COVID-19) on board the diamond Princess Cruises Ship, 2020. Infect Dis Model 2020;5:264–74. https://doi.org/10.1016/j.idm.2020.02.003.
[15] Liu Y, Gong M, Liu S, Pan Y, Huo Y. Effects of blood glucose on vasoconstriction in patients with gestational diabetes mellitus. Gynecol Endocrinol 2021;37(3):221–4.
[16] Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos Solitons Fractals 2020;134:1–12. https://doi.org/10.1016/j.chaos.2020.109761.
[17] Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. Chaos Solitons Fractals 2020;135:109850. https://doi.org/10.1016/j.chaos.2020.109850.
[18] Ribeiro MHDMM, da Silva RG, Mariani VC, Coelho LDSS. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos Solitons Fractals 2020;135:109883. https://doi.org/10.1016/j.chaos.2020.109883.
[19] Yang Z, Zeng F, Wang K, Wong S-S, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 2020;12(3):165–74.
[20] Peng L, Yang W, Zhang D, Zhuge C, Hong L. Epidemic analysis of COVID-19 in China by dynamical modelling; 2020.
[21] Li X, Zhao X, Sun Y. The lockdown of Hubei Province causing different transmission dynamics with a case study of Wuhan, China by dynamical modelling; 2020.
[22] Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV and its implication for public health interventions. J Clin Med 2020;9:462.
[23] Islam MZ, Estivill-Castro V, Rahman MA, Bossonmaier T. Combining K-Means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering. Expert Syst Appl 2018;91:402–17. https://doi.org/10.1016/j.eswa.2017.09.005.
[24] Ding Y, Fu X. Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm. Neurocomputing 2016;188:233–8.
[26] Žalik KR, Žalik B. Validity index for clusters of different sizes and densities. Pattern Recogn Lett 2011;32(2):221–34.

[27] Zhang J, Chung H-H, Lo W-L. Clustering-based adaptive crossover and mutation probabilities for genetic algorithms. IEEE Trans Evol Comput 2007;11(3):326–35. https://doi.org/10.1109/TEVC.2006.880727.

[28] Ansari ZA, Sattar SA, Babu AV. A fuzzy neural network based framework to discover user access patterns from web log data. Adv Data Anal Classif 2017;11(3):519–46. https://doi.org/10.1007/s11634-015-0228-4.