Data Clustering Mapping of Covid-19 Pandemic Based On Geo-Location and Machine Learning

Mustafa Abdul Slaam [1,2]* **Karam Gouda**[3] **Ahmad Naguib**[3]

[1]Artificial Intelligence Dept., Faculty of Computers and Artificial Intelligence, Benha University Egypt
[2]Faculty of Computers Studies, Arab Open University, Egypt
[3]Information Systems Dept., Faculty of Computers and Artificial Intelligence, Benha University Egypt
* Corresponding author's Email: Mustafa.abdo@fci.bu.edu.eg

Abstract: The covid-19 virus pandemic has spread quickly, with the virus beginning in Wuhan, Hubei Province, China, and quickly spreading to practically every country in the world. Because of the Covid-19 pandemic's rapid spread, the disease's contagious nature, and the delay of vaccine production, the only option is to prevent people from mingling in a mob. Using data mining techniques to clustering hotspots zones for impacted Corona positive patients and narrowing the focus to only those zones can be one of the suitable solution against the spread of pandemic. Up-to-date and reliable information about hotspot zones can help the government efficiently implement the measures by focusing resources on the zones, as well as notify other residents about such hotspot zones. The most typical use of hotspot detection in public health is to identify the outbreaks of diseases. For clustering COVID-19 Pandemic, there are a variety of algorithms used by researchers, this research aims to compare between three method of clustering technique DBSCAN (Density-Based Spatial Clustering of Applications with Noise), K-Means and hierarchical clustering. The performances of these clustering algorithms are evaluated using Silhouette score values and elbow method for K-means. Producing a powerful hotspot map can help the decision makers to solve the problem quickly. The proposed approaches are applied to the Data Science for COVID-19 (DS4C) dataset. The dataset is available on the Korea Centre for Disease Control and Prevention's official repository (KCDC). Experimental results show that DBSCAN method separate the data to 4 main cluster noise points with eps=0.45 and minimum pts=20, and for K-Means method with k = 4, including all points, as no noise points in K-mean, and to 5 clusters in hierarchical cluster method with no noise points.

Keywords: DBSCAN, K-Means, Hierarchical clustering, Geo-location, COVID-19

1. Introduction

At the end of 2019, the new species virus called as covid-19 infected started from Wuhan City in Hubei Province, China. As fast as spreading of covid-19, by March 11, 2020, World Health Organization (WHO) announced the covid-19 became pandemic [1].

in order to spatially restrict the impact of a pandemic. The primary infection sources (hot spots) as well as the most vulnerable population regions must be identified as soon as possible. National or regional confinement measures have been found to be successful, despite the fact that they have had substantial social and economic effects in many nations. A more concentrated confinement around infection areas with a higher infection risk could help to prevent worldwide measures that harm the economic development of these countries, especially the poorer ones.

Disease maps or hotspot map one of the most effective way that can visualize the location of the most infected area in specific country or city, it may uncover subtle patterns in data that are lost in tabular presentations and provides a quick visual overview of complicated geographic information. The objective of this research is comparing between three popular Clustering techniques (Kmean, DBSCAN and Hierarchical) to provide a methodology for hotspot detection and locate the places that are the major sources of infection or that may become high-infection zones early. Clustering is Unsupervised learning that consist of sequence technic for structure identification in data set without refer label training set that already known of data vector [2].

Clustering is data grouping process into each group that have high similarity of data and the others among of groups also have low similarity of data. Use DBSCAN method is good work for it in low dimension space, such as two dimensions' feature in case of geo-spatial [3]. In DBSCAN method, conduct grouping of data according to minimum size the object that participate in each cluster and with minimum distancing that needed among them. but k-means conduct grouping the object of data as a group's number that determine before, so the iteration number with cluster centroid will influenced by first cluster centroid randomly. Therefore, it can be fix by determine of cluster centroid at the first high data for obtain higher performance.

The goal of this study is to demonstrate how to use the clustering approach from Data Science to find and graphically plot hotspot zones (clusters) on a map using the gathered patient's geo location. [4]. This Clustering hotspot or clustering map of COVID-19 will assist in determining the pandemic's scope and impact. [5]. It also aids in identifying areas where health-care services should be improved. It's a significant method for making decisions, planning, and taking action in the community. The remainder of the paper is structured into four sections. Literature review briefly explain in section 2. section 3 explains the methodology. Section 4 discusses the impact into the experimental study of the stated purpose. The study results and discussions discussed in section 5. Finally, section 6 contain the conclusion and future work of this study.

Manuscript received April 5, 2022
Manuscript revised April 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.4.56

1.1 Motivation and contribution

There are many challenges to detect hotspot of covid-19 and clustering containment zones , one of the most challenges is collecting accurate geolocation data for each patient , most of hotspot map of covid-19 work with accumulative number of infected cases in each country or city , and Very few institutions are interested in recording geolocation data of infected cases individually , this study focus on clustering technique to detect hotspot area based on data of latitude and longitude for each infected case ,and evaluate each cluster to find the optimum algorithm, so the result that gain from this study is very accurate to implement a hotspot map for covid-19 infected cases and monitoring the spread of the diseases

2. Related work

Valentine Seaman, who published two maps of the 1795 Yellow Fever Outbreak in New York City [6], was among the first to perform studies regarding such a spread in terms of its spatial characteristics. Seaman mapped all deaths caused by the yellow fever in this area and was able to successfully relate these to waste places, which were frequent repositories of rubbish and filth of every variety. Others have made similar maps displaying yellow fever outbreaks in several of the country's main cities (Shannon, 1981; Stevenson, 1965) [7]. Another well-known example is John Snow's map of the 1854 Cholera Outbreak (Snow, 1854) [8, which depicts the number of deaths along Broad Street in Soho, London. While his work has several methodological flaws (Koch & Denike, 2009) [9], it once again shows the importance of visualization of disease transmission could be. In the age of covid-19 pandemic, there are Several studies in the literature began to apply different models to try to identify the epidemic behavior of COVID-19 in the beginning of the outbreak, even before the WHO declared the disease a pandemic. There have been several efforts to combat the spread of COVID-19 and any pandemic crisis. The majority of them are interested in medical issues such as possible vaccinations, early COVID-19 detection utilising X-ray images, and so on. In addition, several solutions have been devised to reduce COVID-19's negative effects. However, just a few studies have looked into spatial growth. These articles were limited to reporting the number of illnesses or deaths by region without providing pertinent information that could lead to successful actions, such as local containment or other measures.

Among the papers identified in the literature, Kramer's [10] published one described a strategy for predicting the spread of a disease by assessing the relative probability of possible epidemic routes. This study examined multiple models that identified the Ebola virus epidemic's network space movement in West Africa. To calculate the transmission probability across cities, the suggested model used a generalized gravity model with distance and population density. Poon [11] published a paper that presented an automated approach for monitoring and identifying hot spots of HIV transmission in British Columbia, Canada. This system's database comprises over 32 000 genotypes for about 9000 HIV-positive people. The data was clustered by the monitoring system, which extracted groups of five or more individuals with phylogenetic distances. In Rajasthan, India, the cluster containment method for the Zika virus outbreak (Singh et al., 2019) [12] was proven to be successful. In their research, Singh et al (2019) discuss how monitoring tactics are utilised to keep the illness from spreading beyond 3 kilometre containment zones. The study emphasises the need of constructing containments to avoid disease outbreaks, however it does not explain how to create these zones fast and accurately.

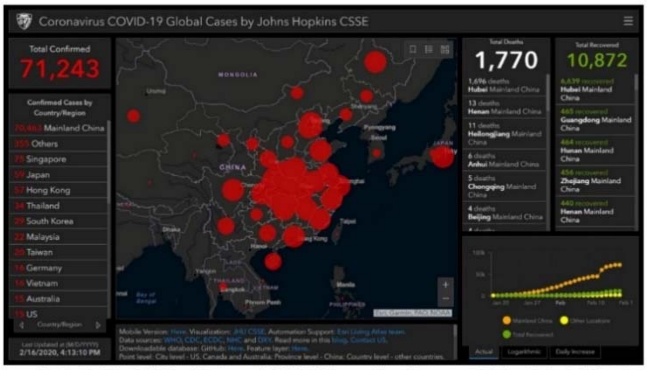
In their study (Maier & Brockmann, 2020) [13], they discussed how to reduce COVID-19 infections in China with successful containment. Quarantine of symptomatic infected cases as well as other community isolation measures are captured by the model they described in their research. The research focuses on the contagion process and its impacts in general, as well as the need of containment. Their research suggests and supports the necessity to accurately determine confinement zones.

Kang [14] used Moran's I statistical approach to illustrate the geographical epidemic dynamics of COVID-19 in Mainland China. This study looked at instances that were geographically near together to see whether there was a relation between viral infection locations. The geographical analysis was used to figure out how infectious diseases spread.

To estimate the number of infections in Wuhan from December 1, 2019 to January 25, 2020, Wu et al. [15] evaluated data from December 31, 2019 to January 28, 2020 on the number of cases exported from Wuhan globally (known days of symptom start from December 25, 2019 to January 19, 2020). The authors focused their forecast on the number of cases transmitted from Wuhan to other important Chinese cities such as Beijing, Shanghai, Guangzhou, and Shenzhen. Using Markov Chain Monte Carlo techniques, the fundamental reproductive number was calculated using the SEIR model. Domestic infections have an impact on these cities, However, the government's actions to lock Wuhan's borders and restrict movement inside China were effective in containing the rates, and the contamination rates did not follow the exponential rates forecasted for these cities.

The authors of Zhong et al. (2020) [16] presented a predicting model for COVID-19 patients based on a simple mathematical model with minimal epidemiological data. Even while epidemiological models aid in assessing the dynamics of escalation, they rely on distinct hypotheses and require knowledge of particular parameters. Also there are number of helpful online/mobile GIS and mapping dashboards and apps for tracking the coronavirus pandemic and associated events as they spread throughout the world. Johns Hopkins University Centre for Systems Science and Engineering dashboard, the spread of 2019-nCoV is being closely monitored by John Hopkins University (JHU). Using ArcGIS Online, JHU created a map-sensitive dashboard that gathers pertinent data from WHO, the US Centres for Disease Control and Prevention (CDC), the ECDC China CDC

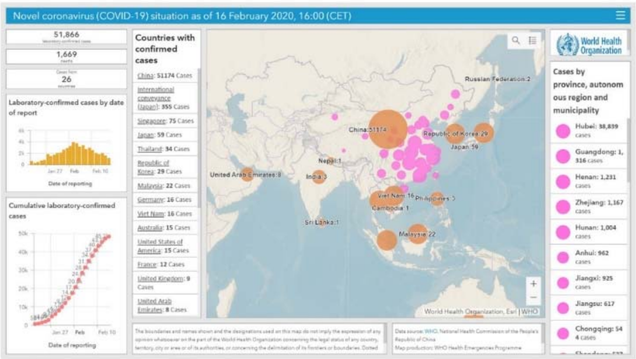
(CCDC), NHC, and Dingxiangyuan. Every day, the online mapping tool is updated with new information regarding the spread of 2019-nCoV, as shown in Fig.1.



Johns Hopkins University CSSE is tracking the spread of SARS-CoV-2 in near real time with a map-centric dashboard (using ArcGIS Online) that pulls relevant data from the WHO, US CDC (Centers for Disease Control and Prevention), ECDC (European Centre for Disease Prevention and Control), Chinese Center for Disease Control and Prevention (CCDC), NHC (China's National Health Commission), and Dingxiangyuan (DXY, China). Screenshot date: 16 February 2020

Fig. 1 COVID-19 Global cases map by John Hopkins University

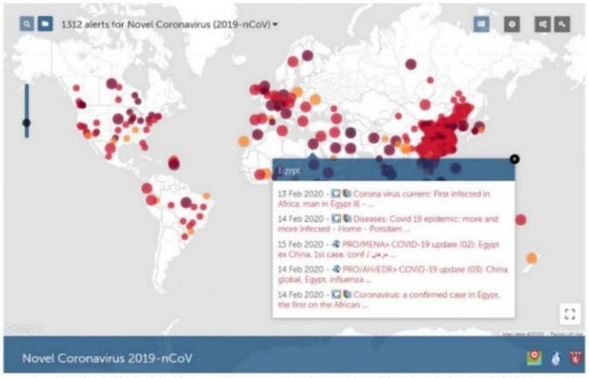
The World Health Organization dashboard(WHO) released its ArcGIS Operations Dashboard for COVID-19 on January 26, 2020, which maps and displays coronavirus infections and total deaths by nation and Chinese province, as well as informational panels regarding the map and its data resources, as shown in Fig 2.



The WHO COVID-19 situation dashboard. Screenshot date: 16 February 2020

Fig. 2 The WHO COVID-19 situation dashboard

HealthMap: developed in 2006 and employs internet media sources for real-time surveillance of emerging public health hazards, is maintained by a team of academics, epidemiologists, and software engineers at Boston Children's Hospital in the United States. HealthMap gathers epidemic information from a variety of sources, including news media (e.g., via Google News), social media, confirmed official warnings (e.g., from the World Health Organization), and expert-curated reports [19]. [20] Health-interactive Map's map for SARS-CoV-2 provides near-real-time geo-located updates from several sources to help clarify the pandemic's spread., as shown in Fig. 3


1312 alerts for Novel Coronavirus (2019-nCoV)
Egypt
Novel Coronavirus 2019-nCoV

13 Feb 2020 - Corona virus current: First infected in Africa, man in Egypt ill - ...

14 Feb 2020 - Diseases: Covid 19 epidemic: more and more infected - Home - Potsdam ...

15 Feb 2020 - PRO/MENA> COVID-19 update (02): Egypt ex China, 1st case, conf ...

14 Feb 2020 - PRO/AH/EDR> COVID-19 update (03): China, global, Egypt, influenza ...

14 Feb 2020 - Coronavirus: a confirmed case in Egypt, the first on the African ...

Screenshot of HealthMap for Wuhan Coronavirus showing a number of news articles and alerts about the first case of COVID-19 reported in Africa (Egypt) on 13 February 2020. Screenshot date: 17 February 2020. (HealthMap uses base map data from Google.)

Fig. 3 Health Map Dashboard.

3.Preliminaries

3.1 K-Means Clustering Technique

K-Means Clustering [17] is one of the clustering techniques based on partitions. The K-Means approach was chosen because it is one of the simplest and clear data clustering algorithms. The K-Means algorithm divides a data set into a specified number of K distinct subgroups. The partitioning method minimizes the square-error for each cluster in order to maximize homogeneity within the cluster. The square error calculated using the following formula:

$$E = \sum_{i=1}^{K} \sum_{j=1}^{n} |dist(x_j, c_i)|^2 \qquad (1)$$

The square error is determined as the square of the distance x between each item and the cluster center (or mean). The center of each cluster is represented by object c. K-Means uses the following approach to reduce the square error. The centers of the K clusters are selected randomly from within the subspace at first. The data set's objects are then partitioned into the clusters that are closest to them. K-Means iteratively computes the new cluster centers and then repartitions them based on the new centers. The K-Means method repeats this procedure until the cluster membership stabilizes, resulting in the final partitioning.

3.2. DBSCAN Clustering Algorithm

The DBSCAN [18] approach's aims are to categories candidate points as main points, border points, or outliers. DBSCAN will basically return clusters of any shapes given the two parameters eps and minpts. If the application assumes a collection of sample points, DBSCAN's clustering method determines the number of points within the chosen eps neighborhood distance of a specific point. The most notable density-based clustering approach is DBSCAN. Clusters are found using this approach as areas with higher densities than the rest of the data. Objects required to divide clusters in these sparse areas are mostly considered noise and boundary points.

3.3. Hierarchical clustering

Hierarchical clustering is similar to partition-based clustering in that it classifies data points in a different way. Each data point is initially considered an independent cluster. The most comparable clusters that are near together are then merged. It continues to iterate until all clusters have been merged. Unlike the k-means technique, it does not need to define the number of clusters to be produced in advance. Furthermore, compared to K-means clustering, hierarchical clustering produces a dendrogram, which is an appealing tree-based representation of the observations.

4.Data and study area

South Korea is divided into 17 administrative divisions (provinces) and 250 districts. Daily verified cases of COVID-19 by district were collected from the Korea Centers for Disease Control and Prevention (KCDC) (Korea Centers for Disease Control and Prevention, 2020) and each province website from January 20 to May 31, 2020. (Seoul Metropolitan Government, 2020). Furthermore, Data Science for COVID-19 provides aggregated data based on all publicly available information in Korea (DS4C, 2020). The case number, province, district (si/gun/gu), date of diagnosis, date of discharge from hospital/community treatment center, result (released from hospital/isolated for treatment/died), sex, and age were all included in the data. Statistics Korea provided the total population per district in 2020 as well as shape files for mapping were obtained from Statistics Korea.

The dataset includes information such as per-patient symptom starts and confirmed date, travel frequency, hospital accessibility, and 61 preventative initiatives taken in South Korea, among other things. [19] The DS4C-PPP dataset provides a comprehensive and unique insight of COVID-19 in Korea. There are three types of data in the dataset:

1) Patient data: illness start and confirmed date for each patient, as well as travel frequency

2) Policy data: comprehensive descriptions and implementation dates for the 61 COVID-19 intervention policies,

3) Provincial data: the number of COVID-19 screening facilities in each municipal district, as well as the size and population density of each municipal district

Our approach on the case dataset file regard its contain the data of COVID-19 infection cases and its location,

- case_id: the ID of the infection_case, case_id(7) = region_code(5) + case number(2)
- province: Province name
- city: City name, if the value 'from another city' indicates that the group infection began in another location.
- group: TRUE: infection group / FALSE: not group
  - If the value is 'TRUE' in this column, the value of 'infection cases' means the name of group.
  - The values named 'contact with patient', 'overseas inflow' and 'etc.' are not group infection.
- infection case: the infection case (the name of group or other cases)
  - The value 'overseas inflow' indicates that the infection originated in a different nation.
  - The value 'etc' includes individual cases, cases where relevance classification is ongoing after investigation, and cases under investigation.
- Confirmed:the total number of confirmed cases
- latitude: the group latitude (WGS84)
- longitude: the group longitude (WGS84)

Table 1: structure of DS4C dataset

Case_id	Province	City	Infection_case	Latitude	Longitude
1000001	SEOUL	Yongsan-gu	Itaewon clubs	37.538621	126.992652
1000002	SEOUL	Gwanak-gu	RICHWAY	37.48208	126.901384
1000003	SEOUL	Guro-gu	Guro-gu call center	37.508163	126.874209

5. Experimental results

To evaluate the quality of each clustering algorithm to find the optimum one there many methods to check the quality of outcome result, Silhouette analysis is one of this method that used to investigate and comprehend the distance between the resulting clusters This method is used to determine how close each object in one cluster is to another cluster's objects. The silhouette score ranges from -1 to +1. A score of +1 indicates that items are appropriately clustered, whereas a value of -1 indicates that objects are not properly clustered.

The following steps demonstrate how to determine the silhouette value for each object.

a) Calculate the average distance between the jth object and the other objects in its cluster. Let's call this number xj.

b) Calculate the average distance between the jth object and all other objects in other clusters. c)Calculate the smallest value over all clusters. Let's call this number yj.

d)The silhouette value for the jth is sj = (yj-xj) / max (xj,bj).

Another way for determining the appropriate partition was the elbow method. For varying values of k, the elbow technique calculates the clustering algorithm. The total sum of squares is then calculated for each k. We were able to find the proper number of clusters using the sum of squares representation for the number of clusters k.

5.1. Working Platform

This work was done on R Studio Version 1.4.1106f using R languages which mostly related to be used in data manipulation, data science, and statistical analysis.

5.2. DBSCAN clustering analysis

DBSCAN clustering method is used to detect the hot spot by clustering the area with high density of infected patient and consider the other area with low density as a noise points, using Silhouette Coefficient index to optimize the value parameter of DBSCAN eps and minpts to determine optimum value, the calculation results are show in Table 2, and show the silhouette method in Fig.4, Fig. 5, then cluster the points into 4 cluster as shown in Fig-6.

Table 2: Quality of Cluster use silhouette Score Method

Eps	Min Pts	Silhouette score	cluster	Cluster points	noise
0.45	20	0.68	4	140	74
0.45	10	0.44	2	222	22
0.45	25	0.86	2	109	135

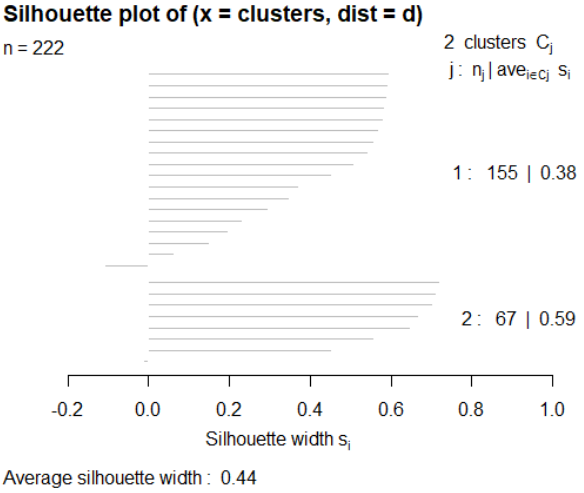


Fig. 4 Silhouette score results for DBSCAN with n=222

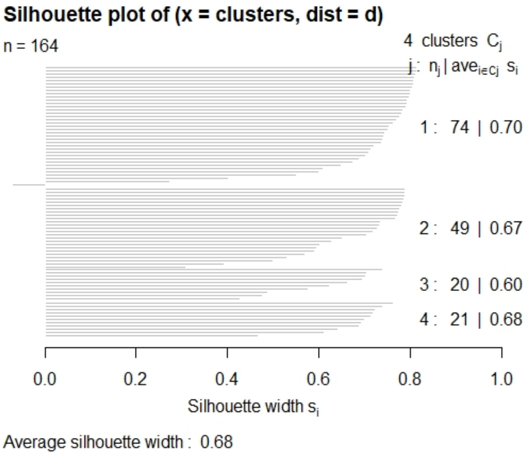


Fig. 5 Silhouette score results for DBSCAN with n=164

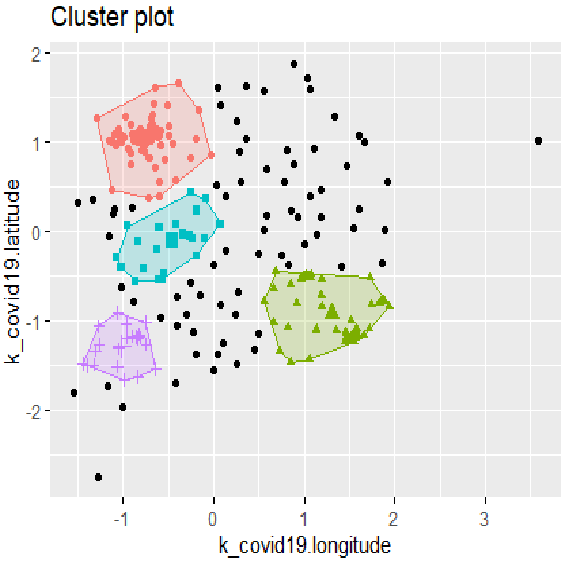


Fig. 6 clustering using DBSCAN

5.3. K-means clustering analysis

By using the elbow method to find the optimal number of cluster K in k-mean algorithm, finding the optimal number of cluster is 4, as shown in Fig. 7.

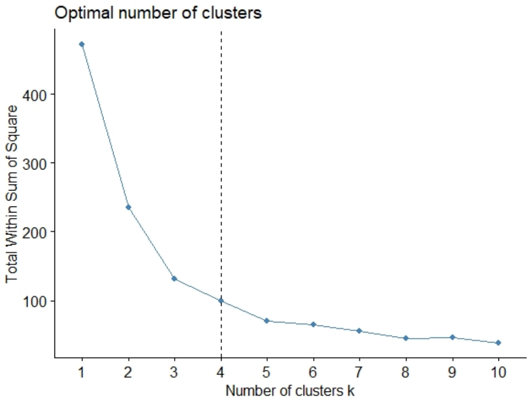
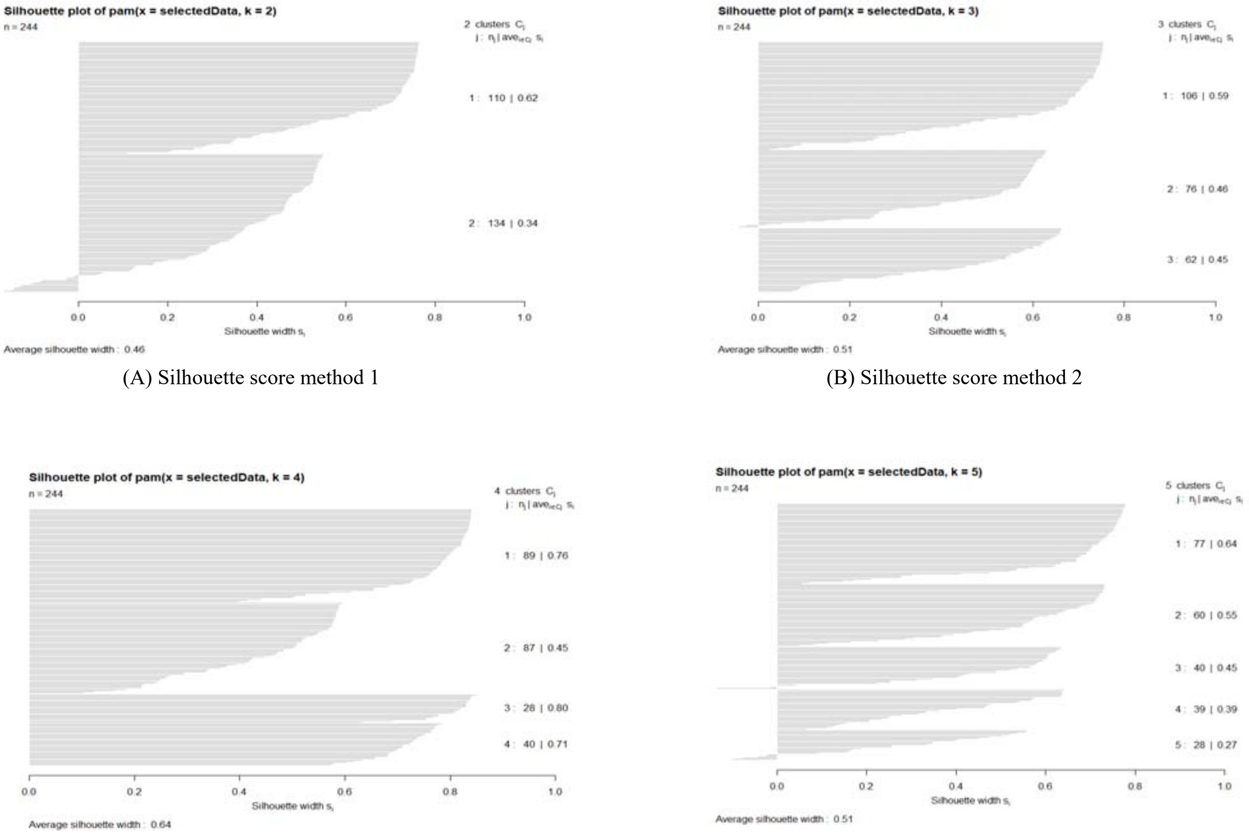


Fig. 7 Application of the elbow method to obtain the optimum number of clusters (k = 4).

Using silhouette Coefficient method to evaluate the k-mean cluster. According to table 3, from silhouette coefficient calculation with average from minimum and maximum results in each clusters, found that 4 cluster reach the best of silhouette Coefficient by optimum number cluster in Kmean algorithm, this



(A) Silhouette score method 1

(B) Silhouette score method 2

(C) Silhouette score method 3

(D) Silhouette score method 4

Fig. 8 Silhouette score results for Kmean

result is matched with elbow method and give same number of cluster, in Fig. 8 its show Silhouette score results for Kmean for the different cluster number in table 3, then in Fig. 9 show the 4 clusters using Kmean algorithm.

Table 3: Quality of K Cluster use silhouette Score Method

Cluster No	Silhouette Score
2	0.46
3	0.51
4	0.64
5	0.51

5.4. . Hierarchal clustering analysis

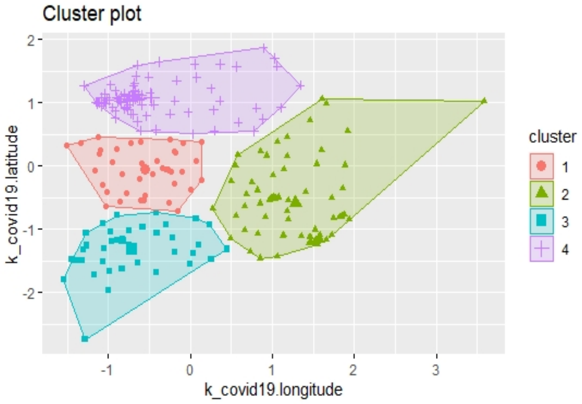


Fig. 9 clustering using Kmean

In R, there are several functions for computing hierarchical clustering. For agglomerative hierarchical clustering, the hclust function is employed (HC)

With hclust, we perform agglomerative HC. We first compute the dissimilarity values using Euclidean distance, then feed them into hclust, define the agglomeration technique, and plot the dendrogram.

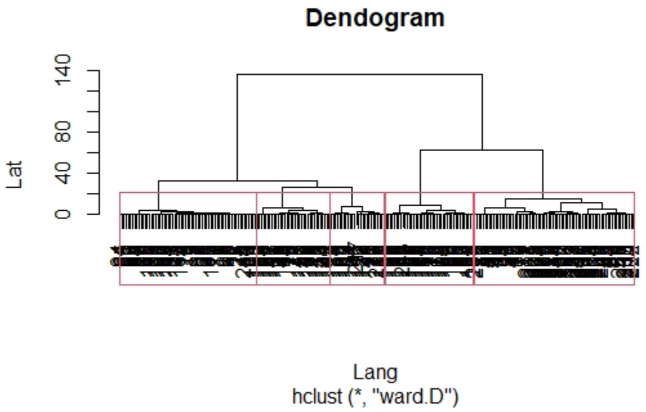


Fig.10 Dedogram for HC

Based on the hierarchical structure of the dendrogram, we choose the ideal number of clusters, as illustrated in Figure 10. clusters can be considered for the agglomerative hierarchical as 5 cluster, then by using silhouette coefficient to check the score

Table 4: Quality of HC using silhouette score

Cluster No	Silhouette Score
3	0.41
4	0.43
5	0.47
6	0.43

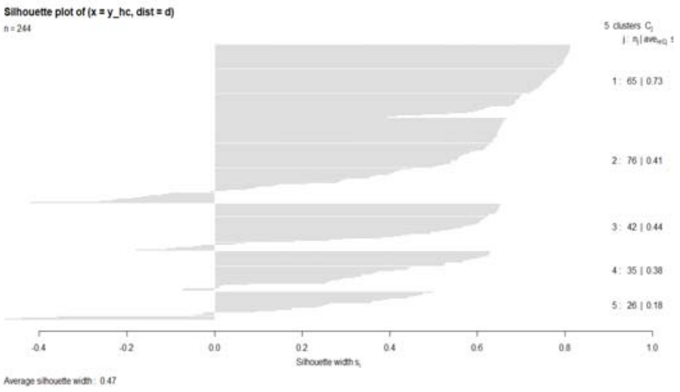


Fig. 11 Silhouette score results for HC

As it shown in figure 12 all points are grouped in different cluster although some point is far enough from the centroid to consider it in the cluster, but in DBSCAN it considers the outlier point as noise and focus on the density of each cluster, so DBSCAN is recommended to detect hotspot of covid -19.

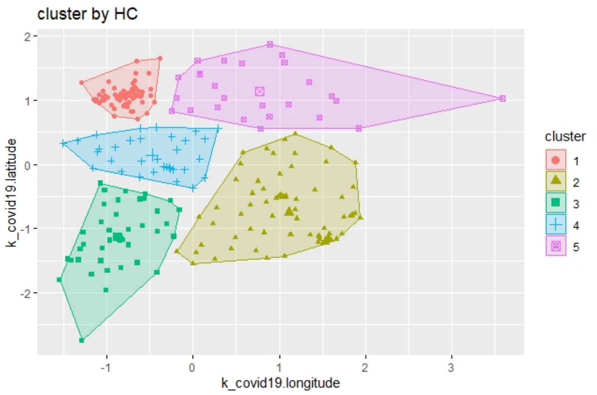


Fig. 12 Clutring using HC

To compute the total run time we started a timer function provided by R language start.time <- Sys.time() just before the clustering call. At the end of clustering we stopped the timer end.time <- Sys.time(). Then we calculate the difference time.taken <- end.time - start.time , All other user programmes were closed, and each test case's execution time was measured 10 times before the average was computed.

Table 5: Runtime analysis results

Algorithm	Execution time (Sec.)	No. of cluster	Sample size	Noise points
Kmean	0.5012929	4	244	No noise points
DBSCAN	0.483701	4	244	Noise Point considered as outlier
HC	0.4526381	5	244	No noise points

5.5. Results Discussions

DBSCAN is a unique clustering algorithm. As the name implies, this strategy focuses on the closeness and density of data to generate clusters. In contrast to K-means and hierarchal clustering, in which an observation becomes a member of the cluster represented by the closest centroid, K-Means and hierarchal clustering may group together weakly related observations. Every observation becomes a part of some cluster eventually, even if the observations are scattered far away in the vector space. Since clusters depend on the mean value of cluster elements, each data point plays a role in forming the clusters. Slight change in data points might affect the clustering outcome. This problem is greatly reduced in DBSCAN due to the way clusters are formed, The DBSCAN algorithm is the only algorithm considered that can label connections as noise. The K-Means and hierarchal clustering algorithm place every connection into a cluster, each point visualizes in the map represent an infected case stored in the dataset, not all represented point could be considered as a hotspot, so DBSCAN is fit to Implement the hotspot map.

6. Conclusion and future work

This study applied three clustering algorithms (K-mean, Hierarchal and DBSCAN) to recognize the covid-19 hotspot in a specific area. The clusters are formed based on COVID-19 patient's locational data. The used algorithms were evaluated using silhouette coefficient method for each algorithm (K-mean, hierarchal and DBSCAN), and with the elbow method for K-mean. Simulation results showed that the DBSCAN method is more relevant with this pandemic case and could detect outliers that do not fit into any of the clusters. DBSCAN clustering is more useful with data when we don't know how many clusters there could be.

In future work, optimization methods will be used to self-adapting the algorithms parameters. Also, federated learning for data privacy model will be used [20].

7. REFERENCES

[1] I. Franch-pardo, B. M. Napoletano, F. Rosete-verges, and L. Billa, "Spatial analysis and GIS in the study of COVID-19. A review," Sci. Total Environ., vol. 739, p. 140033, 2020, doi: 10.1016/j.scitotenv.2020.140033.

[2] S. S. Vallery, Happy Novita, "Data Cluster Mapping Of Global Covid-19 Pandemic Based On Geo-Location," J. Mantik, vol. 3, no. January, pp. 31–38, 2019.

[3] R. Hermawati and I. S. Sitanggang, "Web-Based Clustering Application Using Shiny Framework and DBSCAN Algorithm for Hotspots Data in Peatland in Sumatra," Procedia Environ. Sci., vol. 33, pp. 317–323, 2016, doi: 10.1016/j.proenv.2016.03.082.

[4] S. Chinchorkar, "Defining Covid 19 containment zones using K- means dynamically," pp. 1–8.

[5] A. Islam, M. A. Sayeed, M. K. Rahman, J. Ferdous, S. Islam, and M. M. Hassan, "Geospatial dynamics of COVID-19 clusters and hotspots in Bangladesh," Transbound. Emerg. Dis., no. December 2020, pp. 3643–3657, 2021, doi: 10.1111/tbed.13973.

[6] V. Seaman, "An inquiry into the cause of the prevalence of the yellow fever in New-York," Evolution (N. Y)., vol. 44, no. 0, pp. 1–26, 1798.

[7] L. G. Stevenson, "Putting Disease on the Map," J. Hist. Med., vol. 20, no. 3, pp. 226–261, 1965.

[8] C. Darwin, "MODE OF COMMUNICATION OF CHOLERA," Oxford Univ., vol. XXX, p. 60, 1895.

[9] T. Koch and K. Denike, "Crediting his critics' concerns: Remaking John Snow's map of Broad Street cholera, 1854," Soc. Sci. Med., vol. 69, no. 8, pp. 1246–1251, 2009, doi: 10.1016/j.socscimed.2009.07.046.

[10] A. M. Kramer, J. T. Pulliam, L. W. Alexander, A. W. Park, P. Rohani, and J. M. Drake, "Spatial spread of the West Africa Ebola epidemic," R. Soc. Open Sci., vol. 3, no. 8, 2016, doi: 10.1098/rsos.160294.

[11] A. F. Y. Poon et al., "routine HIV genotyping : an implementation case study," vol. 3, no. 5, pp. 1–15, 2017, doi: 10.1016/S2352-3018(16)00046-1.Near.

[12] R. Singh et al., "Cluster containment strategy: Addressing Zika virus outbreak in Rajasthan, India," BMJ Glob. Heal., vol. 4, no. 5, pp. 1–4, 2019, doi: 10.1136/bmjgh-2018-001383.

[13] B. F. Maier and D. Brockmann, "Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China," Science (80-.)., vol. 368, no. 6492, pp. 742–746, 2020, doi: 10.1126/science.abb4557.

[14] D. Kang, H. Choi, J. H. Kim, and J. Choi, "Spatial epidemic dynamics of the COVID-19 outbreak in China," Int. J. Infect. Dis., vol. 94, no. January, pp. 96–102, 2020, doi: 10.1016/j.ijid.2020.03.076.

[15] X. Zhou et al., "Forecasting the Worldwide Spread of COVID-19 based on Logistic Model and SEIR Model," medRxiv, 2020, doi: 10.1101/2020.03.26.20044289.

[16] L. Zhong, L. Mu, J. Li, J. Wang, Z. Yin, and D. Liu, "Early Prediction of the 2019 Novel Coronavirus Outbreak in the Mainland China Based on Simple Mathematical Model," IEEE Access, vol. 8, pp. 51761–51769, 2020, doi: 10.1109/ACCESS.2020.2979599.

[17] G. Ogbuabor and U. F. N, "Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value," Int. J. Comput. Sci. Inf. Technol., vol. 10, no. 2, pp. 27–37, 2018, doi: 10.5121/ijcsit.2018.10203.

[18] H. A. Hussein and A. M. Abdulazeez, "COVID-19 PANDEMIC DATASETS BASED ON MACHINE LEARNING CLUSTERING ALGORITHMS: A REVIEW PJAEE, 18 (4) (2021) COVID-19 PANDEMIC DATASETS BASED ON MACHINE LEARNING CLUSTERING ALGORITHMS: A REVIEW Covid-19 Pandemic Datasets Based On Machine Learning Clustering," J. Archaeol. Egypt/Egyptology, vol. 18, no. 4, pp. 2672–2700, 2021.

[19] J. Kim, "DS4C Patient Policy Province Dataset: A Comprehensive COVID-19 Dataset for Causal and Epidemiological Analysis," no. NeurIPS, 2020.

[20] M Abdul Salam, S Taha, M Ramadan, (2021) COVID-19 detection using federated machine learning. PLOS ONE 16(6): e0252573.