Baghdad Science Journal

Volume 22 | Issue 8 Article 24

8-27-2025

A Clustering Technique Based on the Hard K-Means (H.KM.) Method to Determine the Governorate That Have More Influence for Spreading COVID-19 in the Kingdom of Saudi Arabia

Rand Muhaned Fawzi
Department of Mathematics, College of Education for pure Science ibn Al-Haitham, University of Baghdad, Baghdad, Iraq

Wurood R. Abd Al-Hussein
Department of Mathematics and Computer Applications, College of Sciences, Al-Nahrain University, Baghdad, Iraq

Iden Hassan Alkanani
Department of Mathematics, College of Science for Women, University of Baghdad, Baghdad, Iraq

How to Cite this Article

Fawzi, Rand Muhaned; Al-Hussein, Wurood R. Abd; and Alkanani, Iden Hassan (2025) "A Clustering Technique Based on the Hard K-Means (H.KM.) Method to Determine the Governorate That Have More Influence for Spreading COVID-19 in the Kingdom of Saudi Arabia," *Baghdad Science Journal*: Vol. 22: Iss. 8, Article 24.
DOI: https://doi.org/10.21123/2411-7986.5035

BAGHDAD SCIENCE JOURNAL 2025;22(8):2752–2760

Scan the QR to view the full-text article on the journal website

RESEARCH ARTICLE

A Clustering Technique Based on the Hard K-Means (H.KM.) Method to Determine the Governorate That Have More Influence for Spreading COVID-19 in the Kingdom of Saudi Arabia

Rand Muhaned Fawzi [1,*], **Wurood R. Abd Al-Hussein** [2], **Iden Hassan Alkanani** [3]

[1] Department of Mathematics, College of Education for pure Science ibn Al-Haitham, University of Baghdad, Baghdad, Iraq
[2] Department of Mathematics and Computer Applications, College of Sciences, Al-Nahrain University, Baghdad, Iraq
[3] Department of Mathematics, College of Science for Women, University of Baghdad, Baghdad, Iraq

ABSTRACT

The Kingdom of Saudi Arabia is the gathering place of most of the nationalities of the Islamic world. When a certain disease spreads, it will be important to know which governorate has the greatest influence on the spread of the disease and take the necessary precautions to limit its spread, and this is the goal of this study. COVID-19, The SRS-COV-2 coronavirus that caused the most recent pandemic is known as the Corona pandemic. To determine which Saudi governorates had the greatest influence on the epidemics spread, data was gathered for thirteen governorates over two months (July and August). The data was analyzed by using cluster analysis. The Saudi governorates were divided into cluster (groups) and cluster centers, these centers represent the main characteristics of each cluster (group) by using the Hard K-Means (H.KM.) clustering technique, and the optimal number of clusters (groups) was calculated by applying the validity clustering methods to identify the group that has the greatest influence on the epidemic's propagation. We employ variance analysis (ANOVA table) to determine the governorate that has the greatest influence on the spread of the disease by knowing the variances within each cluster (group) and between clusters. The goal of ANOVA is to determine whether there are statistically significant differences between the governorates (clusters) in terms of the spread of the disease. The conclusion of the study suggests that the governorates of (Riyad, Maka and Eastern) have had the most impact on the COVID-19 pandemic spread.

Keywords: ANOVA, Clustering, Covid-19, Hard K-Means, Validity

Introduction

Cluster analysis is one unsupervised machine learning technique; it divides the observations into groups such that the similar ones take the same group and the dissimilar ones into another group. It's an important technique for separating observations. [1]

Clustering plays every vital role in exploring data, creating predictions and overcoming anomalies in the data, to satisfy this goal there are two types of cluster analysis:

1- Fuzzy or (soft) cluster analysis 2- Hard or (crisp) Cluster analysis:

In fuzzy (soft) cluster analysis approach assigns each observation in the dataset X to different clusters

Received 10 March 2024; revised 25 October 2024; accepted 27 October 2024.
Available online 27 August 2025

* Corresponding author.
E-mail addresses: rand.M.f@ihcoedu.uobaghdad.edu.iq (R. M. Fawzi), worood.riad@nahrainuniv.edu.iq (W. R. A. Al-Hussein), idedalkanani58@gmail.com (I. H. Alkanani).

https://doi.org/10.21123/2411-7986.5035

with different memberships which vary between 0 and 1. In another type assign each observation to exactly one cluster with membership exactly (either 0 or 1).[2–4]

The analysis of variance (ANOVA) called the fisher analysis of variance this term became well-known in 1925. It is used to compare variances across the means (cluster centers) of different clusters. A range of scenarios is used to determine if there is any difference between the means (cluster centers) of different clusters.[5]

The outcome of ANOVA is the 'F statistic'. This ratio shows the difference between the within-group variance and the between- group variance, which ultimately produces a figure which allows a conclusion that the null hypothesis is supported or rejected.[5]

One-way aims to determine the existence of a statistically significant difference among several cluster centres. This test uses the variances to help determine if the centres are equal or not.

The aim of clustering is to find similar clusters of observations in the dataset, but the important question is how to evaluate results without missing the auxiliary information, this is one of the fundamental problems of clustering?, It can be shown that there are no absolute standards for the best clustering, but it depends on the research's problem and researcher thought that he should decide whether the observations are correctly clustered or not.[6] Therefor we can use the validity clustering to evaluate the clustering results from finding the optimal number of clusters which are the best description of the data structure without any loss information.[7]

The rest of this study is organized as follows: section 2 contains the Hard K-Means (H.KM.), section 3 contains the Experiment, section 4 contains the Result and Discussion and finally section 5 contains conclusions.

Materials and methods

The Hard K-Means clustering technique

The H.KM. Technique clustering called (Lloyd Forgy algorit technique) was developed by J. B. Macqueen in 1967 as a simple centroid-based method.[8]

This approach, which divides the dataset into clusters, is the oldest and most widely used partitional technique. Its merits include efficiency, speed, and brevity. The k-means clustering algorithm has been extensively researched and utilized in several fields, including medical, engineering, programming, and image processing.[7,9]

This algorithm's goal is to minimize the square error in each cluster and the error measure, which form the foundation of this methodology. This method seeks to identify K partitions that meet a given set of requirements to obtain the best clustering:

1. Select a few observations from the dataset to serve as the initial cluster centers
2. The remaining observations are gathered at their initial centers based on the minimum distance criterion. After that, we obtain the initial classification.

If the classification is deemed unreasonable, we revise it by recalculating each cluster center. This process is repeated until we obtain a classification that makes sense.[9]

To compute the minimizing objective function and centroid of H.KM the following formula is used:

$$J_{H.KM.} = J(X;C) = \sum_{i=1}^{n}\sum_{j=1}^{k} \|x_i - c_j\|^2 \quad (1)$$

Where i is the dataset and j is the number of clusters.

$$c_j = \frac{1}{m_j}\sum_{x\in C_j} X \quad (2)$$

Where c_j the j^{th} cluster, c_j is the centroid of cluster C_j, m_j the number of observations in the j^{th} cluster and X all the observations.

The Steps of Algorithm:

1- Assume the number of clusters K and centers C (in this study we choose the centers randomly)
2- Compute the distance between cluster center and the observations by using Eq. (1).
3- Distributed the observations on the closet centroid cluster based on minimum distance.
4- Recomputed the centroids of each cluster by using Eq. (2).
5- Repeat stages (1–4) until centroid does not change[10] as Fig. 1.

Real data

This section provides a real and numerical dataset to illustrate the Hard K-Means method's performance. Table 1 displays the numerical and authentic dataset that was derived during two-months from thirteen governorates in the Kingdom of Saudi Arabia during the COVID-19 pandemic. This procedure is often used to produce varying numbers of clusters with

Table 1. The real data.

No.	Albahah	Hail	Aljouf	Northren	Tabuk	Najran	Alqasim	Asir	Madina	Jazan	Easrern	Maka	Riyad
1	29	19	7	11	20	1	62	156	74	96	337	377	310
2	29	13	2	37	21	16	115	177	109	55	327	208	279
3	9	13	6	17	19	28	14	137	45	62	314	265	219
4	39	14	8	40	18	18	115	77	90	32	273	222	227
5	34	16	8	8	14	33	23	273	68	110	234	220	206
6	18	8	8	50	18	88	27	134	35	59	264	240	328
7	19	23	6	12	28	46	6	184	23	67	264	222	307
8	19	4	30	18	21	45	26	93	59	94	277	252	319
9	18	15	5	1	18	24	26	108	63	76	206	245	328
10	17	48	5	18	14	41	43	96	59	56	178	297	305
11	37	44	3	7	17	29	26	196	51	40	209	260	193
12	12	37	5	10	11	64	47	125	71	64	254	200	344
13	26	49	8	23	24	39	55	135	45	92	189	283	327
14	23	38	4	18	14	47	93	140	56	41	158	260	354
15	29	45	5	18	23	32	48	122	56	24	185	265	313
16	32	41	4	23	18	45	83	143	69	86	240	240	274
17	50	40	8	15	23	39	77	124	43	82	128	211	258
18	16	32	3	13	20	33	66	114	55	23	170	187	323
19	25	64	5	17	22	33	60	129	62	79	219	262	316
20	31	36	7	20	22	43	77	174	63	54	207	200	339
21	25	43	6	17	20	45	59	143	49	69	176	188	302
22	18	44	4	14	16	56	96	131	57	72	158	211	285
23	21	55	6	24	19	51	73	157	68	90	211	209	263
24	21	47	6	25	23	41	83	150	59	107	170	244	280
25	24	53	5	14	19	39	70	122	54	92	226	183	293
26	20	61	3	6	12	43	81	145	62	69	169	316	265
27	21	49	4	40	32	57	104	154	62	117	224	242	273
28	33	59	12	32	25	41	86	127	67	82	271	239	260
29	23	73	11	30	42	41	97	76	63	100	220	260	253
30	36	58	7	31	28	52	92	59	64	118	174	212	256
31	34	56	8	33	19	53	68	84	64	79	209	196	243
32	13	58	3	17	23	40	170	94	69	64	154	207	235
33	17	44	5	20	18	45	35	108	70	88	152	244	217
34	13	46	10	23	56	18	62	89	70	107	188	209	184
35	16	43	11	18	19	39	39	126	65	92	169	214	192
36	17	41	10	24	35	29	48	98	55	101	162	189	177
37	16	32	10	16	20	28	68	102	66	98	152	182	164
38	22	32	10	17	23	37	55	88	48	81	128	166	143
39	7	32	7	7	16	35	49	60	39	67	132	151	129
40	11	38	11	8	10	39	71	74	43	83	132	142	134
41	18	30	12	19	23	36	71	115	50	85	125	159	121
42	17	31	11	14	19	35	54	86	47	72	111	147	107
43	18	26	11	14	17	41	73	80	41	72	113	127	133
44	11	22	12	14	15	32	51	58	41	66	92	119	148
45	15	14	10	14	12	28	42	48	30	55	74	106	161
46	6	16	3	5	9	26	40	48	28	46	62	85	168
47	6	22	9	5	10	34	48	50	27	59	67	97	170
48	11	17	14	15	15	25	43	41	35	61	59	101	132
49	9	16	10	12	14	24	41	41	30	46	56	88	159
50	7	14	11	12	9	17	34	43	28	49	46	84	145
51	8	14	11	10	13	25	45	30	27	39	42	71	123
52	7	13	5	10	9	16	29	29	20	39	40	66	126
53	3	12	2	3	5	21	30	23	21	31	37	62	134
54	3	12	5	5	4	17	32	30	22	33	39	64	94
55	7	10	7	11	10	16	23	30	26	34	41	66	72
56	6	8	7	8	10	16	32	26	20	25	32	57	74
57	5	6	9	7	8	15	29	23	18	27	30	51	62
58	2	7	8	8	4	13	16	19	17	23	27	46	54
59	4	6	6	7	6	10	22	13	14	22	21	36	67
60	1	5	2	2	5	10	15	14	13	14	22	36	69
61	1	8	3	3	5	9	19	15	15	20	21	36	66
62	5	6	7	6	5	13	16	12	12	15	20	34	73

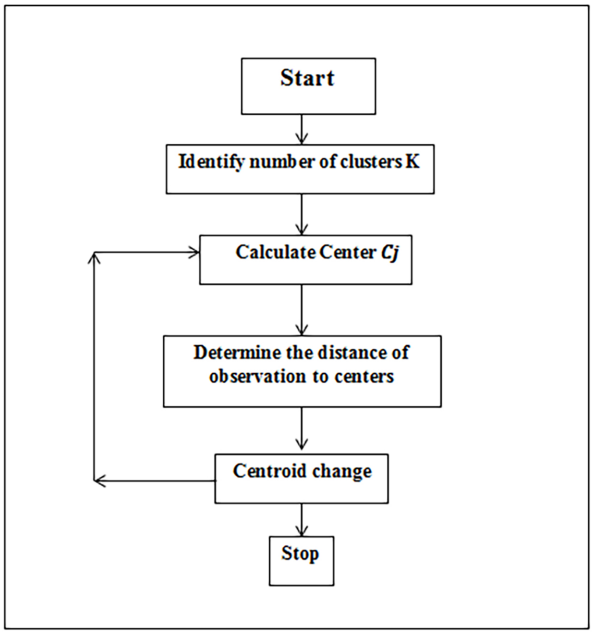


Fig. 1. The HK.M algorithm.

corresponding cluster memberships, after which the clustering outcomes are assessed.

After using the H.KM. Technique to analyze the data and cluster the data into three groups, we will utilize a one-way ANOVA table to identify which governorates are most important in the disease's spread.

Results and discussion

We analyze the dataset X by using the Hard K-Means method in SPSS:

1- We calculate the distance between observations and centers (assign the point to nearest center) after choosing the initial centroid (randomly). as shown in the Table 2 below:

Table 2. The initial value of cluster center.

	Number of Cluster		
	1	2	3
Riyad	310	73	206
Maka	377	34	220
Eastern	337	20	234
Jazan	96	15	110
Madina	74	12	68
Asir	156	12	273
Alqasim	62	16	23
Najran	1	13	33
Tabuk	20	5	14
Northern	11	6	8
Aljouf	7	7	8
Hail	19	6	16
Albahah	29	5	34

Table 3. Iteration history of change in cluster center.

	Change in Cluster Centers		
Iteration	1	2	3
1	162.869	97.805	164.636
2	9.963	0.000	14.008
3	4.780	0.000	5.045
4	0.000	0.000	0.000

2- We update the cluster center and repeat the procedure of distributing the observations to the new centers (update centroids). We do this until convergence is reached since the cluster center has not changed significantly. Upon reaching the final centers (the centroids approach stapelty), where the maximum absolute coordinate change for each center is 0 as shown in Table 3.

In this table the cluster center was shapely in iteration 4. The minimum distances between initial centers is 4.780.

We can evaluate the best clustering from Table 3 is three clusters where the cluster centers was stable in iterative 4 that mean the choices of initial cluster centers was perfect.

From Table 4 calculated the final cluster centers to 3 clusters.

3- We use ANOVA Table (One-Way) to calculate the F and significance for each governorate:

The null hypothesis is $H_0 = \mu_1 = \mu_2 = \ldots = \mu_{25} = 3$

Alternative hypothesis $H_1 \neq \mu_1 \neq \mu_2 \ldots \neq \mu_{25} \neq 3$

We must calculate the value of the F = (0.01) and compare it with the significant value for each governorate to ascertain whether or not that governorate has an effect on the spread of the disease in the Kingdom of Saudi Arabia. If the significant value for each governorate is less than the F, then that governorate is considered to have an effect on the spread of the disease. However, this governorate has no impact on the spread of the illness when the value is more than the F = (0.01).

From Table 5, Aljouf has no effect on the disease's spread because its significant value (0.5755026120000) was more than F = (0.01). However, the twelve governorates—Riyad, Maka, Eastern, Jazan, Madina, Asir, Alqasim, Najran, Tabuk, Northern, Hail, and Albahah—had an effect because the significant values of it less than F = (0.01), but with different degree as shown:

Fig. 2(a), (b), and (c) show the distributions of the observations on three governorate clusters that were more important in the disease's spread (Riyad, Maka, and Eastern).

Table 4. The final cluster centers.

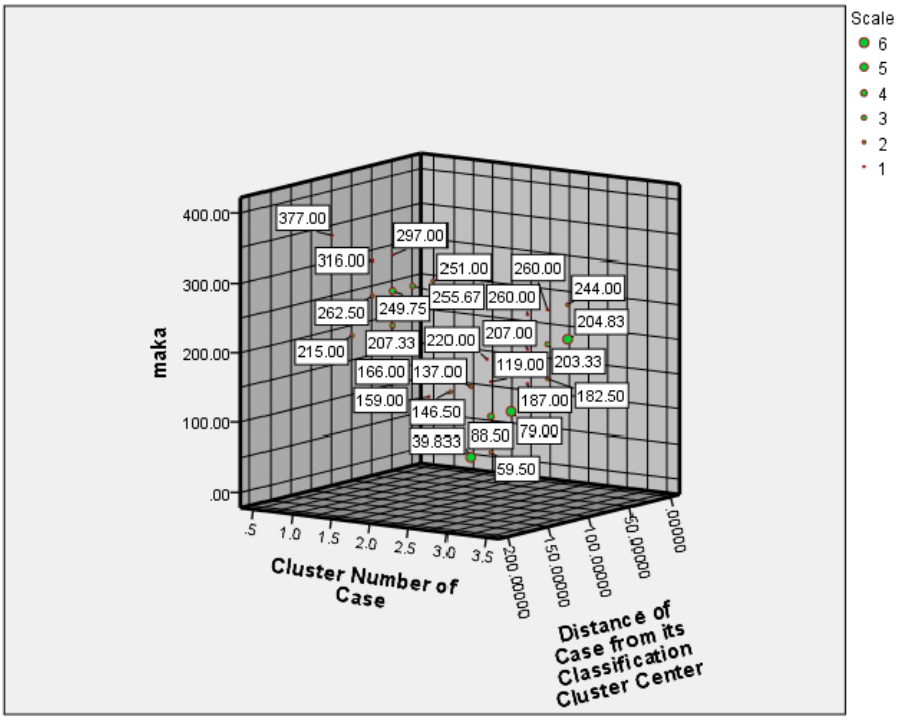
	Cluster		
	1	2	3
Riyad	299.32	114.56	240.22
Maka	254.47	87.88	212.56
Eastern	239.79	62.76	181.22
Jazan	68.68	46.56	85.11
Madina	61.58	28.48	60.56
Asir	134.53	43.84	124.78
alqasim	61.47	39.20	69.56
najran	39.53	23.60	39.50
tabuk	20.00	11.04	23.94
northern	21.63	9.44	19.39
Aljouf	7.05	8.12	6.89
Hail	33.47	16.68	45.83
Albahah	23.63	8.40	23.94

Table 5. Analysis of variance (ANOVA) table.

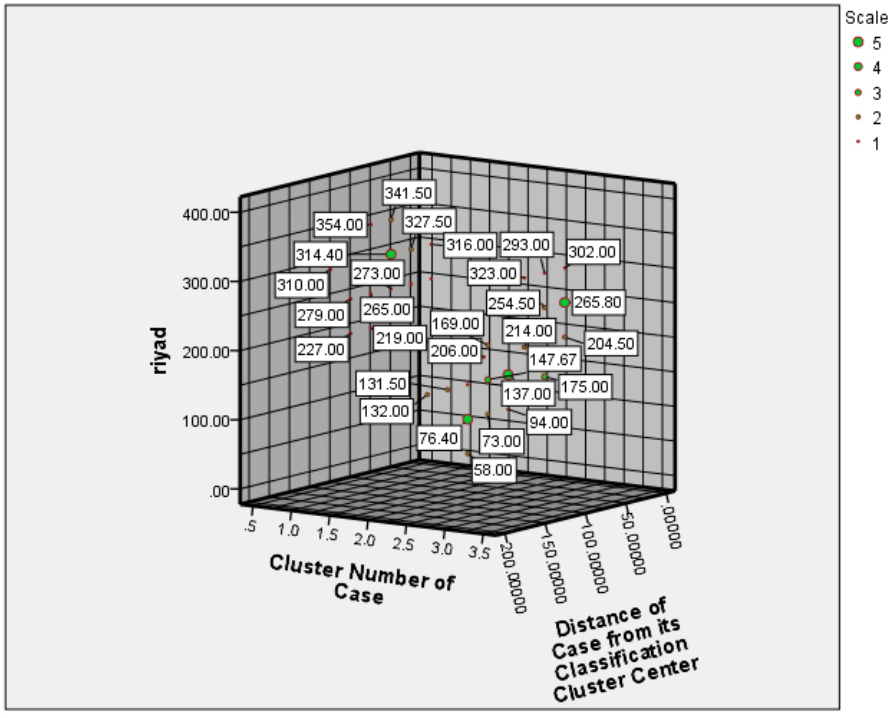
ANOVA							
	Cluster		Error				
	Mean Square	Df	Mean Square	Df	F	Significant	Decision
Riyad	197695.255	2	1649.074	59	119.883	.0000000000000	Reject H_0
Maka	167569.573	2	1449.760	59	115.584	.0000000000000	Reject H_0
Eastern	180438.045	2	1757.200	59	102.685	.0000000000000	Reject H_0
Jazan	8012.696	2	541.865	59	14.787	.0000062305572	Reject H_0
Madina	7933.213	2	179.107	59	44.293	.0000000000018	Reject H_0
Asir	55538.880	2	1247.410	59	44.523	.0000000000016	Reject H_0
Alqasim	5424.603	2	783.512	59	6.923	.0019907085400	Reject H_0
Najran	1889.091	2	179.885	59	10.502	.0001254453920	Reject H_0
Tabuk	954.766	2	53.185	59	17.952	.0000008133616	Reject H_0
Northern	942.442	2	74.489	59	12.652	.0000267665508	Reject H_0
Aljouf	9.938	2	17.820	59	.558	.5755026120000	Accept H_0
Hail	4586.049	2	196.384	59	23.352	.0000000338327	Reject H_0
Albahah	1765.866	2	61.413	59	28.754	.0000000019171	Reject H_0

Table 6. Arranging the governorate sequentially according to their impact on the spread of the disease.

No.	Governorate	Significant
1-	(Riyad, Maka, Eastern)	0.0000000
2-	Asir	0.0000000000016
3-	Madina	0.0000000000018
4-	Albahah	0.0000000019171
5-	Hail	0.0000000338327
6-	Tabuk	0.0000008133616
7-	Jazan	0.0000062305572
8-	Northern	0.0000267665508
9-	Najran	0.0001254453920
10-	Alqasim	0.0019907085400

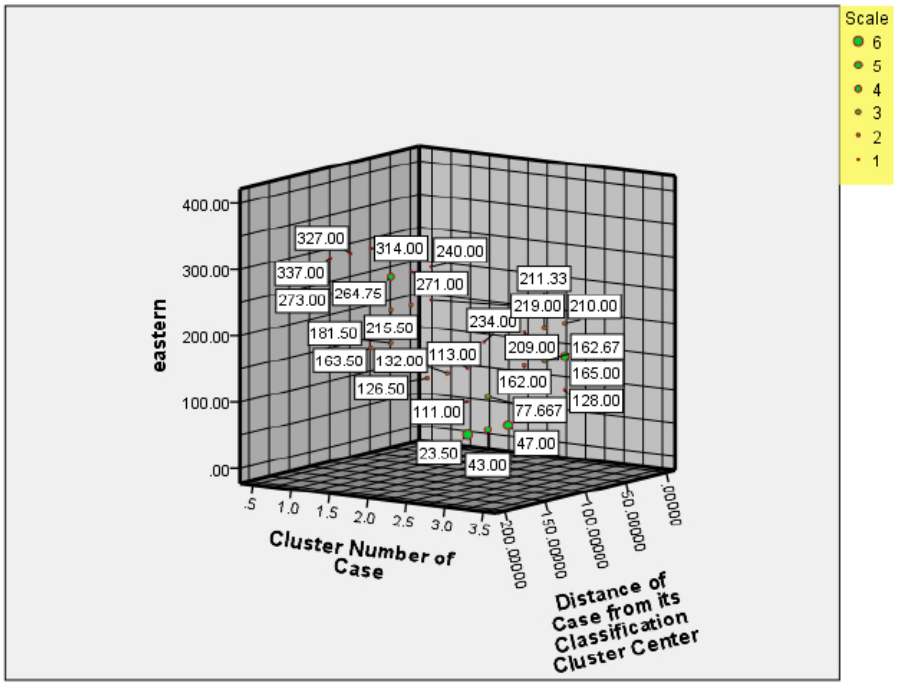


(a)



(b)

Fig. 2. Governorate observations are distributed over the clusters as shown in pictures (a), (b), and (c).



(c)

Fig. 2. Continued.

Conclusion

The optimal number of clusters is three, as shown by creftb2,tb4 that means that the governorates in Saudi Arabia were divided into three groups based on the similarity in data characteristics (and their contribution to the spread of the disease). While Table 3 shows a very excellent result from the clustering process with four iterations, we saw a poor outcome when we tested the data with a different number of clusters; hence, the optimal number of clusters is three. Table 5 shows that we have rejected the H_0 and the results of applying ANOVA to analyze the difference between the means of the clusters and determine whether these differences are statistically significant are shown, since the significance for each governorate was smaller than the F = (0.01) excepted Aljouf governorate, indicating differences between governorates in the spread of the disease. Finally, we have concluded from this study that the more affected governorates (Riyad, Maka, and Eastern) since the significance was = 0 as shown in Table 6 and because of the abundance of trade and the influx of people from all over the world.

Based on this conclusion, we expect that these governorates will be the basis for the spread of any pandemic in the future.

Acknowledgment

We appreciate the efforts of everyone who contributed even a little to this work.

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for republication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University of Baghdad.

Authors' contribution statement

R M, W R. And I H designed the study. I H collected and refined the data, drew the algorithm and tables, R. analyzed the data and obtained the final results, and R and W wrote the research text.

References

1. Kalpit GS, Atul P. Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. International Int J Curr Innov Res. 2017;13(5):899–906.
2. Devi YB, Dayang NAJ, Shahliza AH, Fransiskus A. Natural Language Processing For Requirement Elicitation in University Using Kmeans And Maenshift Algorithm. Baghdad Sci J. 2024;21(2 Special Issue):0561–0567. https://doi.org/10.21123/bsj.2024.9675.
3. Shrook ASA, Bahaa ARQ, Ashraf MS. Using the Hierarchical Cluster Analysis and Fuzzy Cluster Analysis Methods for Classification of Some Hospitals in Basra. Baghdad Sci J. 2021;18(4):1212–1217. http://dx.doi.org/10.21123/bsj.2021.18.4.1212.
4. Zeynel C, Figen Y. Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster structures. Journal of Agricultural Informatics. 2015;6(3):13–23. https://doi.org/10.17700/jai.2015.6.3.196.
5. YanPing Z, XiaoLai Z. K-means Clustering Algorithm and its Improvement Research. J Phys Conf Ser. 2021;1873:1–5. http://dx.doi.org/10.1088/1742-6596/1873/1/012074.
6. Yadgar SA. A new approach to the Fuzzy c-means Clustering Algorithm by Automatic Weights and Local Clustering. Passer J Basic Appl Sci. 2021;3(1):95–101. https://doi.org/10.24271/psr.18.
7. Kristina PS, Miin-Shen Y. Unsupervised K-Means Clustering Algorithm. IEEE Access. 2020;8:80716–80720. http://dx.doi.org/10.1109/ACCESS.2020.2988796.
8. Rand MF, Iden HA. Turbid of Water By Using Fuzzy C- Means and Hard K- Means. Baghdad Sci J. 2020;17(3):988–993. https://doi.org/10.21123/bsj.2020.17.3(Suppl.).0988.
9. Youguo L, Haiyan W. A Clustering Method Based on K-Means Algorithm. Phys Procedia. 2012;25:1104–1109. http://dx.doi.org/10.1016/j.phpro.2012.03.206.
10. Sonia Y, Sachin S. Study Of Existing Methods & Techniques Of K-Means Clustering. Educ Adm.: Theory Pract. 2024;30(4):1806–1813. https://doi.org/10.53555/kuey.v30i4.1755.

تقنية العنقدة المعتمدة على طريقة متوسطات K الحادة لتحديد المحافظة الأكثر تأثيرا في انتشار كوفيد-19في المملكة العربية السعودية

رند مهند فوزي[1]، ورود رياض عبدالحسين[2]، ايدن حسن الكناني[3]

[1] *قسم الرياضيات ، كلية التربية للعلوم الصرفة ابن الهيثم،جامعة بغداد، بغداد، العراق.*

[2] *قسم الرياضيات و تطبيقات الحاسوب،كلية العلوم ، جامعة النهرين ، بغداد ، العراق.*

[3] *قسم الرياضيات ، كلية العلوم للبنات ،جامعة بغداد، بغداد ، العراق.*

المستخلص

تعد المملكة العربية السعودية مكان تجمع اغلب جنسيات العالم الاسلامي فعند انتشار مرض معين سيكون من المهم معرفة اي محافظة ذات التأثير الاكبر في انتشار المرض لاتخاذ الاحتياطات اللازمة للحد من انتشاره و هذا هو الهدف من هذه الدراسة. كوفيد-19، أحدث الجائحة سببها فيروس كورونا SRS-COV-2 يُعرف بجائحة كورونا. ومن أجل تحديد المحافظة السعودية التي كان لها التأثير الأكبر على انتشار الوباء، تم جمع البيانات الفعلية لثلاث عشرة محافظة على مدار شهرين (يوليو وأغسطس). وتم تحليل البيانات باستخدام التحليل العنقودي. تم تقسيم المحافظات السعودية إلى عناقيد (مجموعات) باستخدام تقنية التجميع Hard K-Means (H.KM) ، و تمثل هذه المراكز الخصائص الرئيسية لكل عنقود (مجموعة) وتم حساب العدد الأمثل للعناقيد (المجموعات) من خلال تطبيق طريقة صحة العنقدة لتحديد المجموعة التي لها التأثير الأكبر على انتشار الوباء. نستخدم تحليل التباين (جدول ANOVA) لتحديد المحافظة التي لها التأثير الأكبر على انتشار المرض من خلال معرفة التباين داخل كل عنقود (مجموعة)و بين العناقيد (المجموعات) و الهدف م تحليل ANOFA هو تحديد ما اذا كانت هناك فروق ذات دلالة إحصائية بين المحافظات (العناقيد) من حيث انتشار المرض . وتشير خلاصة الدراسة إلى أن المحافظات (مكة والرياض والشرقية) كان لها التأثير الأكبر في انتشار وباء كورونا.

الكلمات المفتاحية: جدول انوفا، العنقدة، كورونا، متوسطات K الحادة، صحة العنقدة.