Jurnal Mantik
Volume 4, Number 1, May 2020, pp. 511-520 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Data Cluster Mapping Of Global Covid-19 Pandemic Based On Geo-Location

Iskandar Fitri[1], Refly Asmar[2], Albar Rubhasy[3]

[1,2,3.]Information System,
Faculty of Information and Communication Technology, Universitas Nasional, Jl. Sawo Manila, South Jakarta, 12520, Indonesia.

Email: tektel2001@yahoo.com, [2]*reflyasmar@gmail.com*, [3]*albaar.rubhasy@gmail.com*

ARTICLE INFO

Article history:
Received: 04/04/2020
Revised: 20/04/2020
Accepted: 30/05/2020

Keywords:
covid-19,
epidemic,
DBSCAN,
k-means,
silhoutte coefficient,
elbow method

ABSTRACT

The spread of the covid-19 virus pandemic is very fast, where was start of the virus spreading from Wuhan City, Hubei Province, China and suddenly spread out widely to almost around the word. According that kind of pandemic phenomena, this research was conducted to make clustering based on data of latitude and longitude due global spreading of Covid-19 use DBSCAN(Density-Based Spatial Clustering Of Applications With Noise) and K-Meansto find a levelaccurate and suitable in calculating for this pandemic case as the alternative choices for condition analyse in decision making purpose. The algorithm had developed calculate based on characterization from geo-location of the country which is to determine the number quality of cluster use Silhoutte Coefficient and Elbow Methods.Therefore, from calculated results can be analyse similarity of covid-19 spreading pattern refer to clustering in each province or country. From data testing show that DBSCAN method separate the data of noise points with eps=22 and minimum pts=4, and for K-Means method with k = 3. After calculation by use the two methods, finally, can visualize the mapping cluster continent of Asia, Europe and Africa with showing the pattern of increasing covid-19 cases that can began controlled. The other result show cluster for continent of north and South America have increased significant and the Australian Continent cluster gets the lowest case and can controlled.
.

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) declared the *covid-19* became pandemic.In this paper is describe mapping cluster to see and analyse the pattern for spreading of covid-19 pandemic globally.

Cluster analyze is also mention as Unsupervised learning that consist of squeence technic for structure identification in data set without refer label training set that already known of data vector [1].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 featurein case of geo-spatial[2]. The Aim cluster analyze is to find of meaningful observation locationto those similar groups that related to observe a set variable[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. The other method called as *k-means* conduct grouping the object of data as a group's number that determine before[4], 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 [5]. In this research, conduct by use two method of DBSCAN and K-mean for the same input sources data to cluster mapping for pattern spreading of covid-19 in global scale. Use the two methods have purpose to find a level accurate and suitable in calculating for this pandemic case as the alternative choices for condition analyse in decision making purpose. The update data are obtained from official site of*WHO (World Health Organization)*and *worldometers.info/coronavirus/.*, to get similar data from every geo-location based onlatitudes and longitudes.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

2. Research Method

In this case, of this object, the research methodology which is use two-different method for the procedure of calculation are conducted systematically every single step as show in Figure 1. The detail steps are describes in each sub-chapter.

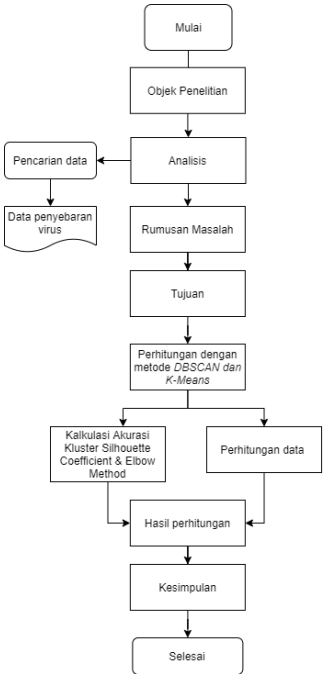

Fig 1. Scheme of Research method

2.1 Collecting of Data

The Data that will test and analyze has attribute such as ; province/city, country,longitude, latitude, date, confirmed number, number of death and number of recovery. All of the data set was collect from website of *WHO* (world Health Organization) and*worldometers.info/coronavirus/.* For data in case of Indonesia are collect from site of*covid19.go.id* and*kemkes.go.id.* The period data collection are from April 1 until April 30, 2020 and all of them tabulate as show in Table 1.

Table 1.
Dataset Example

Country/Region	US	Canada	Netherlands	China	Indonesia
Province/State	Alabama	Alberta	Aruba	Beijing	Jakarta
Lat	32.90210	53.9333	12.5186	40.1824	-6.195779
Long	-86.70589	-116.5765	-70.0358	116.4142	106.84858
Date	15/04/2020	15/04/2020	07/04/2020	15/04/2020	30/04/2020
Confirmed	3734	1732	74	589	4138
Deaths	99	46	0	8	412
Recovered	0	312	14	484	381

2.2 Analyze of Data Cluster

Analyze of data cluster generally is divided two category, first hierarchy cluster and the other one is*non-hierarchy cluster*.From those kind of categories that use in this research is *DBSCAN* method, where it's the most popular use density-based algorithm [7] and for K-Means for a comparative algorithm in calculation of data set purpose. The differences of two methods are show in Table 2.

Table 2.
Comparative between DBSCAN and K-Means Methods.

Algorithm Name	*DBSCAN*	K-Means
Parameter	Size of Neighbourhoods	Cluster Number
Parameter Input	*Eps* and*MinPts*	*k*
Capability to face of	Yes	No

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Algorithm Name	*DBSCAN*	K-Means
Parameter	Size of Neighbourhoods	Cluster Number
Noise		
Geometric/Metric	Distance between nearest points	Distance between points
Use for	Geometric and size of cluster average	General purpose, flat geometric and not tto much clusters.

2.3 DBSCAN Clustering

Density-Based Spatial Clustering of Applications with Noise (*DBSCAN)*is method based on *density based on clustering*[8]. This method is a part of *spatial data mining* where could make spatial big data processing. Exploration of interesting and useful pattern from spatial data set is more difficult than conventional pattern of data numeric because it complexity, relation and auto-correlation among spatial data set [9]. The algorithm program is developed use *Python* language for made of cluster model with library *sklearn* as show in Figure 2. In that scripts was input of Eps value by 22, MinPts by 4 and Metric use *Euclidean Distance*.

```
from sklearn import metrics
from sklearn.cluster import DBSCAN
X2 = latest_data[['Long','Lat']].values
clustering = DBSCAN(eps=22, min_samples=4,metric='euclidean').fit(X2)
predictions1 = clustering.fit_predict(X2)
labels2 = clustering.labels_
```

Fig 2. DBSCAN script use Phyton Language.

2.4 Silhoutte Coefficient

Silhouette Coefficient is use to see the quality and strength of the cluster. This method for measure how far close of relation distance among the data points and distance between the clusters[10]. The script as show in Figure 3.

```
print("Silhouette Coefficient: %0.3f"
% metrics.silhouette_score(X2, labels2))
```

Fig 3. Silhoutte Score

From the calculation results is find the *Silhoutte Score*as can see in Table 4. The best value is"1" and the worse value is "-1". The negative value is a wrong sample that had determined, because differences between the clusters.

2.5 K-Means Clustering

K-Means clustering is a method of non-hierarchy that the first known[11]. K-means is use for measure *Euclidean distance* and interactively determine each record from the sources. In this step was conduct the procedure by select value of k with record initial as cluster center as described in [12].

```
from sklearn.cluster import KMeans
kmeans_1 = KMeans(n_clusters=3)
X = latest_data[['Long','Lat']].values
predictions = kmeans_1.fit_predict(X)
labels2 = clustering.labels_
```

Fig 4. K-Means script use Phyton Language.

The script in Figure 4, made a model of clusters use input parameters of n cluster with determine of value by 3 as cluster, according from calculation results of *Elbow Method* in Figure 6.

2.6 Elbow Method

The fuction of *elbow method* is use for determine the number of cluster from data sets. This basic idea came from unsupervised model that to determine the cluster until total of intra-cluster variation as total of within-cluster variation or total of within-cluster sum of square [13]. The Script in Figure 5, calculate deffrences of centeroid distance average by value from 1 to 10 of optimum cluster and the results as show in Figure 6.

```
K_clusters = range(1,10)
kmeans = [KMeans(n_clusters=i) for i in K_clusters]
Y_axis = latest_data[['Lat']]
X_axis = latest_data[['Long']]
score = [kmeans[i].fit(Y_axis).score(Y_axis) for i in range(len(kmeans))]
plt.plot(score, K_clusters)
plt.xlabel('Rata-Rata jarak ke Centeroid')
plt.ylabel('Jumlah Klaster Optimal')
plt.title('Elbow Curve')
plt.grid()
plt.show()
```

Fig 5. Kalitas klaster dengan *Elbow Method*

3. Results and Discussion

From data processing, results are compare between non-hierarchy clustering results use K-Means method and hierarchy clustering by use DBSCAN method. The notation parameters for *K-Means*use cluster number of *k*and*DBSCAN*method use*eps* and*min_pts*.

3.1 Metode *K-Means*

Implementation in this method can self-determine cluster number of k that wanted. In this step is used *ElbowMethod*which the way by calculated among distance of data average to the centeroid which have the biggest gap will became optimum cluster as show in Figure 6. According Figure 6, found the optimum cluster gap is k = 3, the decided to determine optimum cluster is 3.

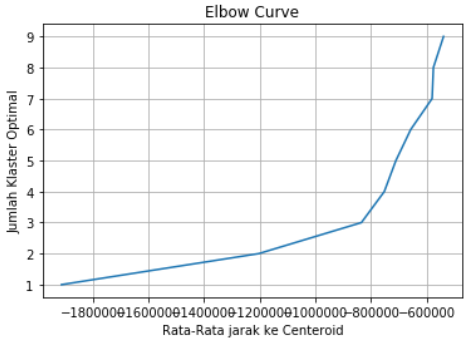

Fig 6. Calculation Results of Quality Cluster use*Elbow Method.*

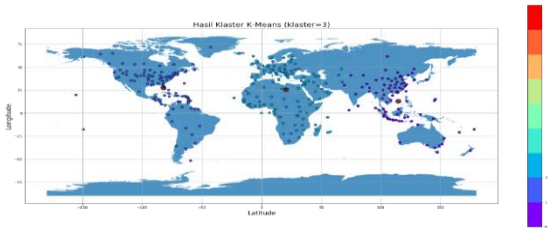

Fig 7. Cluster Results From Calculation use *K-means* Method.

From results as show in Figure 6, we can visualization cluster plot in Figure 7. From Figure 7 show divided became three clusters based on latitude and longitude calculation. Cluster of 0 is at continent of Asia and Australia. Cluster 1 is in position of south and North America continents. The last, for cluster 2 are Europe and Africa continents. All of the cluster show in Table 3.

Table 3.
Cluster of Spreading Poins use by K-Means Method.

Cluster	Spreading Are
0	Asia and Australia Continents
1	North and South America Continents
2	Europe and Africa Continents

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

3.2 Method DBSCAN

Continue calculate use DBSCAN method to identification flux density of areawith *eps* and*min_pts* parameters for determine it optimum value. In this step is use*Silhoutte Coefficient* method. The calculation results is show in Table 4.

Table 4.
Quality of Cluster use Silhoutte Score Method

Eps	Min_Pts	Silhoutte Score	Cluster
12	5	0.383	5
14	4	0.729	6
18	7	0.374	3
19	6	0.395	3
20	5	0.397	3
22	4	0.408	3
23	4	0.380	3

According Table 4, from *Silhoutte Coefficient*calculation with average from minimum and maximum results in each clusters, also use parameters of *eps* and*min_pts* from*DBSCAN*are found that *eps=22* and*min_pts=4* reach the best of*Silhoutte Coefficient*by optimum number cluster which synchronize to cluster divided with*k-means*method. Finally, points of spreading results are show inFigure 8.

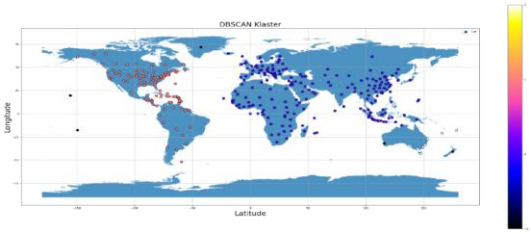

Fig 8. Cluster Results From Calculation use Hasil*DBSCAN* Method.

From Figure 8, cluster 0 is at area of Asia, Europe and Africa continents. For cluster 1 are North and South America continents and cluster 2 is Australia continent as show in Table 5.

Table 5.
Cluster of Spreading Poins use byDBSCAN Method.

Clusters	Spreading Area
0	Asia, Europe and Africa Continents
1	North and South America Continents.
2	Australia Continent.

3.3. Cluster Comparative

In this research, after calculation results use *DBSCAN* and *K-Means* methods had found deference characteristic. These two methods are make form cluster from *Euclidean distance* calculation where data clustering by the same characteristic coordinate of *longitude* and *latitude* as show in Table 6.

Table 6.
Comparative Divided Cluster Number between K-Means and DBSCAN Methods.

Clusters	K-Means	DBSCAN
0	90	207
1	120	117
2	127	8
Noise	-	5

According to Table 6 are found that gaps divided between two methods have much differences. In *DBSCAN* method there is cluster with -1 or noise, where it mean a part of outlier that not include into the three clusters.

Table 7.
Numer of Spreading of Covid-19 in each Clusters use DBSCAN and K-Means Method.

Clusters		0	1	2	*Noise*
DBSCAN	Confirmed	1899733	1293952	6209	2709
	Death	153586	74708	83	43
	Recovered	753845	233657	5212	2335
K-Means	Confirmed	326819	1294634	1581150	-

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Clusters		0	1	2	*Noise*
	Death	10293	74724	143403	-
	Recovered	132452	234244	628353	-

According Table 7, it show that differences between two methods for number of confirmed, death and recovered in cluster 0 and 2. For DBSCAN method in cluster "0" is more cases. It cause that in cluster 0 coverage three continents and for *K-Means* method only covered two continents. At cluster 1, *K-Means* method have just little bit higher of, it cause in *DBSCAN* method have covered North and South Continents with 2 points of noise that not calculated by cluster. For case of cluster 2, its look that K-Means method have number of cases much more because it can covered Europe and Africa continents, meanwhile *DBSCAN* method only covered Australia continent.

3.4. Data Timeline in April 2020

As input data calculation for clustering, take sample data of Covid-19 cases in timeline from April 1-30, 2020 as tabulate as show in Figure 9. The cases have three category that for; confirmed, death and recovery.

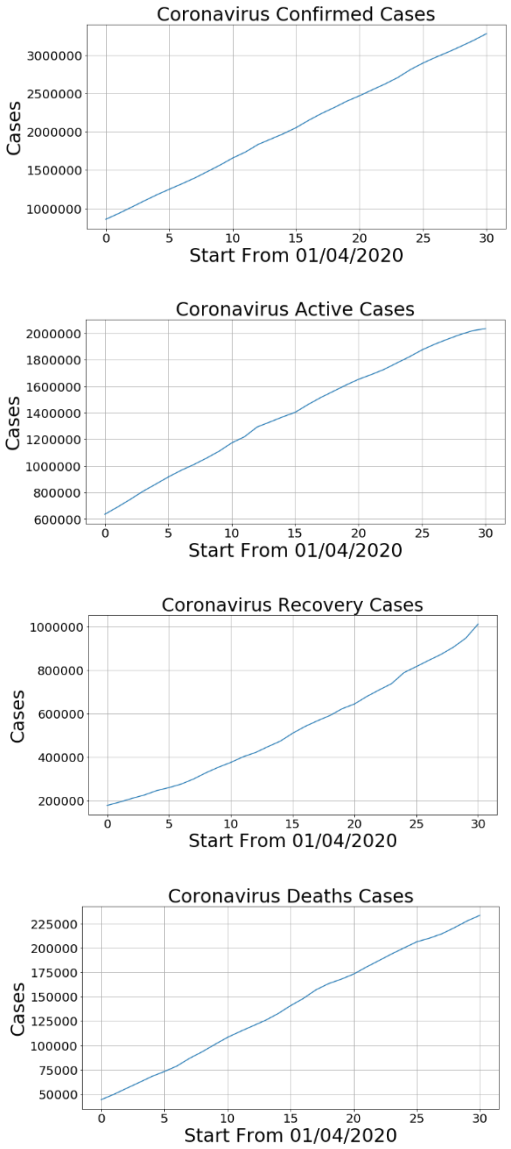

Fig 9. Data Timeline Global Cases of Confirmed, Death and Recovery.

In Figure 9, show in that period the total cases had reach more than 3 million people. For death case have number of 225 thousand people and with recovery case more than 1 million people.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

3.5 Calculate Use by DBSCAN Method

The calculation results by use DBSCAN method visualize in bar chart to show comperative for every cluster from the category of confirm, death and recovery based on the highest number of confirmed cases.

A. Cluster 0 Cases

Fig 10. Number of Confirmed Cluster of"0".

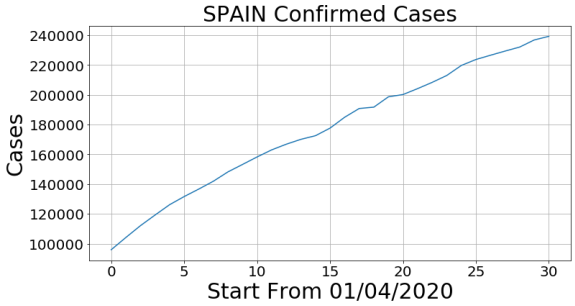

Fig 11. Daily Cases for Spain Country

From Figure 10, cluster 0 is dominant among Europe country with number confirmed cases of above 90.000 people. The biggest case of infected are more than 200.000 people in Spain and for cases of recovery. In cases of death, the highest number of 27.382 people in Italy and at the same time in United Kingdom is find not yet for recovery cases, because the data for UK is not available.

B. Cluster 1 Cases

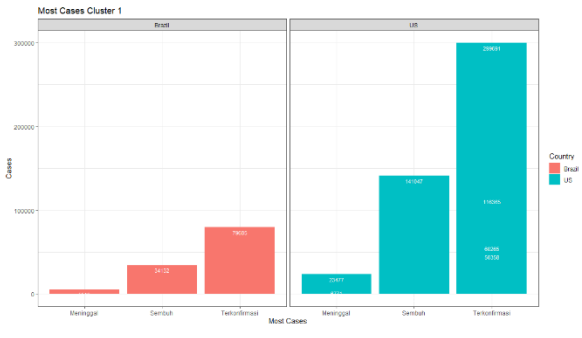

Fig 12. Number of Confirmed Cluster of "1".

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

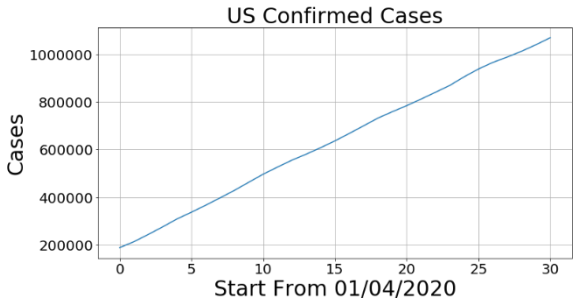

Fig 13. Daily Cases US

In bar chart of Figure 12, for cluster 1 are dominate from America continent with the number of cases is more than 60.000 people, with highest number confirmed of more than 1 million people. New York city took the highest cases of 299.691 people and in the second rank took by Brazil that have confirmed cases of 79.685 people.

C. Cluster 2 Cases

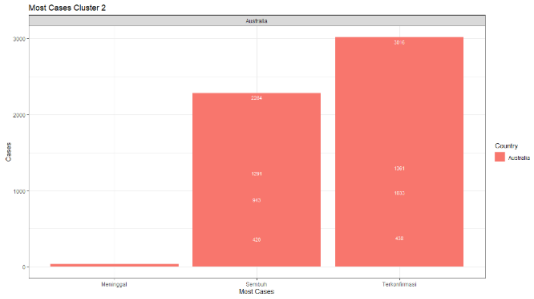

Fig 14. Number of Confirmed Cluster of "2".

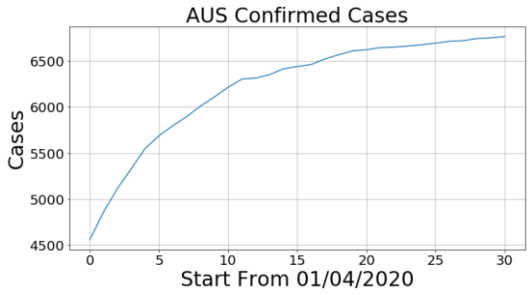

Fig 15. Daily Cases for Australia Country

The Bar chart in Figure 14, the highest number of confirmed cases in Australia continent, more than 6.500 peoplewith the highest cases specific position at New South Wales province by 3.016 people.

3.6 Calculate Use by K-Means Method

In the same scheme scenario of data set is also calculate by use *K-Means* method. The results as show from Figure 16 to Figure 19 for those clusters.

A. Cluster 0 Cases

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

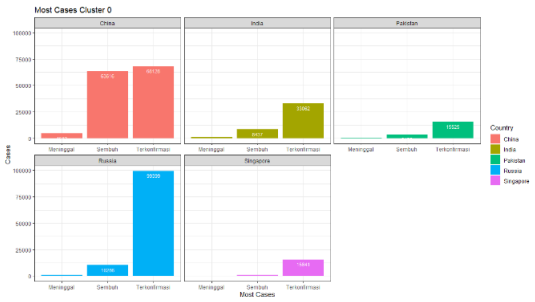

Fig 16. Number of Confirmed Cluster of "0".

Fig 17. Daily Cases for Russia Country

The number people that confirmed infected, for cluster "0" are dominate at around Asia Countries and one country in Europe is Russia by more than 100.000 people. For cases of recovery, the highest number is in China by 63.616 people. It has continued to India by 33.062 people of confirmed cases.

B. Cluster 1 Cases

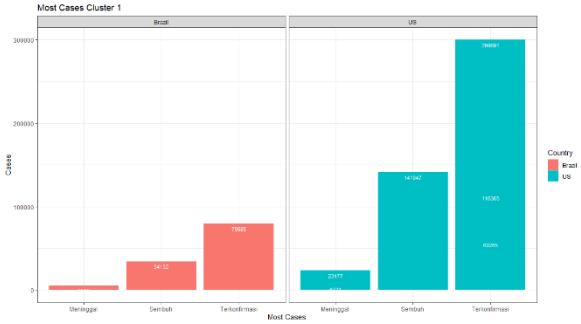

Fig 18. Number of Confirmed Cluster of "1".

In cluster "1" of *K-Means* method, America continent dominate the highest by 300.000 people in US and 79.685 people in Brazil of confirmed cases.

C. Cluster 2 Cases

Fig 19. Number of Confirmed Cluster of "2".

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

The last category of clustering by *K-Means* method is dominate in Europe continent, the highest cases in Spain by 236.899 people with recovery cases by 132.929 people. The highest cases of death was happen in Italy by 27.682 people. At the same time, the recovery cases in UK can be include calculate because the resource data was not available by UK Government.

3.7 Comparative Analyze

The comparative from calculation results of the data use by those two methods based on parameters and data inputs of *Euclidean distance* calculation is find the complexity of both methods as shows in Table 8.

Tabel 8
Kompleksitas

Name Algoritma	Kompleksitas	Run Time
DBSCAN	O(n log n)	0.306s
K-Means	O(kN)	0.108s

The parameters from both methods are determine the results and complex in divided of data clusters. For *DBSCAN*method is determine of *Eps* dan *MinPts*, where the (*n)*is number of digits from data points. Meanwhile, , the *K-Means*method is more faster three times for complexity of data calculation because the input parameters only determine the cluster number of (*k)*that wanted and then calculated distance to centroid linearly. It is different for *DBSCAN*method that has metric complexity between distances of points by (*n log n).*

4. Conclusion

From those calculation results for case of global covid-19 pandemic is find that use between data object with *Euclidean Distance* in *K-Means* method is can be free to determine the cluster based on our scenario, but for *DBSCAN* method must use the parameters of*eps* and*min_pts*to make data clusters. Finally, *DBSCAN*method is more relevant with this pandemic case where the distance or density from each points. However, in case use of *K-Means* has the complexity that faster.

5. References

[1] Bernard Magaret, and DeFreitas. "COMPARATIVE PERFORMANCE ANALYSIS OF CLUSTERING TECHNIQUES IN EDUCATIONAL DATA MINING," International Journal on Computer Science and Information Systems (IADIS). vol. 10, no.2, pp.65-67.
[2] Boeing Geoff, "Clustering to Reduce Spatial Data Set Size," University of California,Berkeley, 2018.
[3] Budi Indra, Rumiarti Deni, "Segmentasi Pelanggan Pada Customer Raltionship Management Di Perusahaan Ritel: Studi Kasus Gramedia Asri Media," Thesis, Universitas Indonesia, Jakarta, 2017.
[4] Darwin Sutawanir, Hajarisman Nusar, and Arsih Nur. "Metode Pengclusteran Berbasis Densitas Menggunakan Algoritma DBSCAN," Universitas Islam Bandung, Indonesia, 2016.
[5] Ester, Martin Kriegel, Hans-Peter Sander, Jörg; Xu, Xiaowei (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9.
[6] Kumaar Arvind, Bholowalia Purnima, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," International Journal of Computer Applications, Vol. 105, Pp. 18, 2014.
[7] Lloyd,S., "Least Squares Quantization in PCM. IEE transactions on information theory," 28(2),pp. 129-137. 1982.
[8] Markatou Marianthi, Alexander H.foss, "kamila: Clustering Mixed-Type Data in R and Hadoop," Journal of Sstatistical Software, vol. 83, pp. 1, Feb. 2018.
[9] Mumtaz K, "An Analysis on Density Based Clustering of Multi Dimensional Spatial Data," Indian Journal of Computer Science and Engineering (IJCSE), vol. 1, no. 1, pp. 8–12, 2010.
[10] Satoto DB, Rochman SME, Khotimah KB, Syakur MA, "Integration K-Means Clustering and Elbow Method For Identification of Best Costumer Profile Cluster," University of Trunojoyo, Madura, 2018.
[11] Savvas .K, Stogiannos Alekos, and Mazis Th, "A Study of Comparative Clustering of EU Countrie using the DBSCAN and K-means Techniques within the Theoretical Framework of Systemic Gepolitical Analysis," Internatonal Journal of Grid and Utility Computing, 2016.
[12] Shekhar S, Zhang P, Huang Y, and Vatsavai RR, "Trends in spatial data mining. Data mining: Next generation challenges and future directions," pp. 357–380, 2003.
[13] Velden de van, D'Enza Iodice, Markos Angelos, "Beyond Tandem Analsis: Joint Dimension Reduction and Clustering in R," Journal of Statistical Software, vol. 91, pp. 1, Oct. 2019.