

## PENGELOMPOKKAN KABUPATEN/KOTA DI JAWA TIMUR BERDASARKAN PENAMBAHAN JUMLAH KASUS POSITIF COVID-19 MENGGUNAKAN METODE *K-MEANS*

Nadya Nisa Dwi Hatmanti

Jurusan Matematika, FMIPA, Universitas Negeri Surabaya

e-mail : nadya.17030214004@mhs.unesa.ac.id

### Abstrak

Covid-19 menyebar dalam waktu yang singkat di berbagai negara di dunia termasuk Indonesia. Jawa Timur adalah provinsi yang memiliki kasus tekonfirmasi Covid-19 yang tinggi. Tingginya kasus Covid-19 di Jawa Timur harus menjadi perhatian lebih dari pemerintah untuk menetapkan kebijakan yang tepat di masing-masing daerah dalam mengurangi penyebaran Covid-19. Karena setiap daerah mempunyai tingkat resiko yang berbeda, sehingga kebijakan yang dilakukan seharusnya berbeda disesuaikan dengan kondisi daerah masing-masing. Dalam artikel ini membahas mengenai pengelompokkan 38 Kabupaten/Kota di Jawa Timur berdasarkan penambahan jumlah kasus positif Covid-19. Data pada penelitian ini merupakan data *time series* karena data positif Covid-19 dicatat setiap hari dari tanggal 19 Maret 2020 sampai 31 Januari 2021. Metode pengelompokkan yang digunakan adalah metode *K-Means*. Diperoleh hasil bahwa cluster yang paling optimal yaitu dengan jumlah  $k=2$ . Nilai koefisien silhouette sebesar 0.75 sehingga *cluster* yang terbentuk dalam kategori *strong cluster*. Dengan jumlah  $k=2$  diperoleh bahwa *cluster* 1 merupakan Kabupaten/Kota dengan penambahan jumlah kasus positif Covid-19 tinggi yaitu Kota Surabaya dan 37 provinsi lainnya masuk dalam *cluster* 2 dengan penambahan jumlah kasus positif Covid-19 yang rendah.

**Kata kunci:** Covid-19, Clustering, Time Series, K-Means

### Abstract

Covid-19 spread in a short time in various countries in the world, including Indonesia. East Java is a province that has a high number of confirmed Covid-19 cases. The high number of Covid-19 cases in East Java should pay more attention to the government to establish the right policies in each region to reduce the spread of Covid-19. Because each region has a different level of risk, the policies carried out should be different according to the conditions of each region. This article discusses the grouping of 38 district/cities in East Java based on the increase in the number of positive cases of Covid-19. The data in this study are time series data because positive data for Covid-19 are recorded every day from 19 March 2020 to 31 January 2021. The method of grouping used is the K-Means method. The results show that the most optimal cluster is the number of  $k = 2$ . The silhouette coefficient is 0.75 so that the cluster formed is in the strong cluster category. With the number of  $k = 2$ , it is found that cluster 1 is a Regency / City with a high increase in the number of positive cases of Covid-19, namely the City of Surabaya and 37 other provinces are included in cluster 2 with the addition of a low number of positive cases of Covid-19.

**Keywords :** Covid-19, Clustering, Time Series, K-Means

## 1. PENDAHULUAN

Covid-19 atau Coronavirus Disease 2019 ialah penyakit yang diakibatkan oleh virus Severe Acute Respiratory Syndrom Coronavirus 2 (SARS-CoV-2). Awalnya Wabah Covid-19 tersebut muncul di Kota Wuhan, Provinsi Hubei, China pada tanggal 31 Desember 2019. Gejala yang paling umum apabila orang terpapar virus ini yaitu demam, batuk kering, dan kelelahan bahkan ada yang tidak mengalami gejala apapun. Untuk sebagian orang juga dapat mengalami

gejala serius seperti kesulitan bernafas, nyeri dada dan kehilangan kemampuan bergerak dan berbicara (World Health Organization, 2020)

Covid-19 menyebar dalam waktu yang singkat di berbagai negara di dunia termasuk Indonesia. Menurut (Gugus Tugas Percepatan Penanganan COVID-19, 2020), orang Indonesia pertama yang terjangkit Covid-19 yaitu di DKI Jakarta pada tanggal 1 Maret 2020. Hingga tanggal 31 Januari 2021, kasus terkonfirmasi Covid-19 di Indonesia telah mencapai 1,078,314. Jawa Timur adalah provinsi

yang memiliki jumlah kasus tinggi yaitu urutan 4 terkait jumlah kasus terkonfirmasi dan urutan pertama terkait jumlah kematian.

Tingginya jumlah kasus Covid-19 di Jawa Timur membuat pemerintah menetapkan kebijakan-kebijakan guna mencegah penyebarluasan virus ini, seperti pemberlakuan PSBB atau Pembatasan Sosial Bersekala Besar. PSBB adalah pembatasan seluruh kegiatan yang berada di tempat umum dan mengkarantinakan diri dirumah masing-masing. Namun, kebijakan ini tidak efektif diterapkan karena dapat menyebabkan industri dan pekerjaan masyarakat tersendat. Sehingga, pemerintah mengeluarkan kebijakan baru yang diberi nama “*new normal*”. Kebijakan ini mengharuskan masyarakat melakukan kegiatan dengan menjalankan protokol kesehatan yaitu berkegiatan dengan memakai masker, tidak berjabat tangan, tetap menjaga jarak, dan sebagainya.

Berdasarkan latar belakang tersebut, pada penelitian ini akan dilakukan pengelompokan Kabupaten/Kota di Jawa Timur yang didasarkan pada penambahan jumlah kasus Covid-19. Data yang digunakan merupakan data yang diperoleh berdasarkan urutan waktu atau data *time series* penambahan jumlah tekonfirmasi Covid-19 di 38 Kabupaten/Kota di Jawa Timur dari tanggal 19 Maret 2020 sampai 31 Januari 2021. Penelitian ini dapat dijadikan informasi bagi pemerintah untuk mengetahui daerah di Jawa Timur yang harus diperketat dalam penanganan kasus Covid-19. Sehingga, dapat dijadikan salah satu solusi untuk menekan angka kasus Covid-19

*Clustering* merupakan pengelompokan yang didasarkan pada karakteristik tertentu. Pengelompokan dilakukan dengan mengelompokkan objek yang memiliki karakteristik yang sama ataupun mirip kedalam satu kelompok. Metode pengelompokan atau *clustering* pada penelitian ini menggunakan metode *K-Means*. *K-Means* merupakan metode pengelompokan non Hierarki yang melakukan pengelompokan dengan cara partisi. Apabila dibandingkan dengan metode clustering lainnya, metode *K-Means* merupakan metode yang cukup baik untuk diterapkan pada semua jenis data (M.Z. Rodriguez et al., 2019). Penelitian sebelumnya terkait metode *K-Means* telah dilakukan oleh (M. Habibi et al., 2019) yang digunakan untuk melakukan pengelompokan karakteristik pengguna instagram dan (A. D. Munthe, 2019) yang diterapkan pada data *time series* nilai produksi padi, selanjutnya penelitian mengenai pengelompokan negara yang didasarkan pada kasus Covid-19 juga telah dilakukan oleh (Vasilios Zarikas, Stavros G. Pouloupoulos, Zoe Gareiou, 2020).

## 2. KAJIAN TEORI

### *CoronaVirus Disease 2019 (COVID-19)*

*CoronaVirus Disease 2019* atau dikenal dengan sebutan Covid-19 merupakan penyakit yang menyerang

sistem pernafasan yang dapat menular dari orang ke orang. Covid-19 pertama kali diketahui di Wuhan, Republik Rakyat China pada tanggal 31 Desember 2019 dan menyebar dalam waktu singkat di berbagai daerah bahkan negara lain. Dan pada tanggal 11 Maret 2020, World Health Organization (WHO) melaporkan bahwa Covid-19 telah menjadi pandemi di seluruh dunia. Apabila orang terkena Covid-19, tidak langsung menunjukkan gejala namun gejala tersebut muncul 5-6 hari atau bahkan dapat sampai 1-14 hari. Gejala yang dialami biasanya adalah sakit tenggorokkan, flu, sesak nafas, batuk, lesu dan letih bahkan pada beberapa orang juga dapat mengalami pneumonia. Gejala tersebut muncul secara bertahap. Pada orang yang memiliki masalah kesehatan seperti hipertensi, jantung, dan kencing manis atau diabetes beresiko mengalami sakit yang lebih serius bahkan dapat menyebabkan kematian (World Health Organization, 2020).

### *Cluster Analysis*

*Cluste Analysis* adalah salah satu analisis multivariate yang bertujuan untuk pengelompokan objek atau data observasi menjadi beberapa kelompok. Objek yang mempunyai kesamaan atau kemiripan akan dijadikan 1 kelompok dan memiliki perbedaan dengan anggota kelompok lain (Rencher, 2002). *Cluster* yang baik mempunyai ciri-ciri kesamaan (homogenitas) besar di dalam satu kelompok dan perbedaan (heterogenitas) besar dengan kelompok lain. *Cluster Analysis* dibedakan menjadi 2 yaitu metode Hierarki dan Non Hierarki. Metode hierarki melakukan proses pengelompokan dengan mengelompokkan objek yang mempunyai karakteristik paling mirip atau yang mempunyai jarak terdekat, kemudian dilanjutkan dengan objek yang paling mirip kedua, dan begitu pula seterusnya sampai membentuk suatu pohon hierarki atau tingkatan antar objek yang mempunyai kemiripan sampai tidak memiliki kemiripan. Untuk membantu memperjelas proses *cluster analysis* dengan metode hierarki tersebut biasanya dipresentasikan menggunakan dendogram. Berbeda dengan metode non hirarki, *cluster analysis* pada metode non hirarki diawali dengan menentukan jumlah cluster, selanjutnya melakukan pengelompokan tanpa mengikuti proses tingkatan atau hierarki. Metode ini merupakan Metode *K-Means*.

### *Metode K-Means*

Metode *K-Means* ialah salah satu metode pengelompokan Hierarki yang mempartisi objek ke dalam beberapa kelompok. Objek yang memiliki persamaan akan dikelompokkan dalam 1 kelompok yang sama. Apabila dibandingkan dengan metode Hierarki, metode *K-Means* banyak digunakan oleh peneliti karena dapat mengelompokkan data yang lebih besar dengan cepat dan metode yang simple . Pengelompokan dengan metode *K-*

*Means* dilakukan dengan tahapan sebagai berikut (Johnson, R., & Wichern, 2007) :

1. Menentukan jumlah *k* (*cluster*) dengan mempertimbangkan teori yang relevan sehingga diperoleh banyaknya kelompok yang disepakati untuk dibentuk.
2. Menentukan sebanyak *k* sebagai titik pusat (*centroid*) awal *cluster* secara sebarang. Penentuan titik pusat *cluster* untuk selanjutnya menggunakan persamaan sebagai berikut:

$$y_i = \frac{\sum_{i=1}^n x_i}{n} \tag{1}$$

Dengan  $y_i$  merupakan pusat *cluster*,  $x_i$  adalah objek ke  $i$ , dan  $n$  merupakan jumlah objek sebagai anggota *cluster*.

3. Menghitung jarak setiap objek terhadap titik pusat *cluster*. Hal ini bertujuan untuk menentukan suatu objek masuk ke dalam *cluster* yang mana. Perhitungan jarak tersebut menggunakan *Euclidean distance* yang dirumuskan dengan persamaan sebagai berikut :

$$d = \sqrt{(x_i - y_i)^2 + \dots + (x_n - y_n)^2} \tag{2}$$

Dimana  $y_i$  merupakan pusat *cluster*,  $x_i$  adalah objek ke  $i$ , dan  $n$  merupakan jumlah objek sebagai anggota *cluster*

4. Setiap objek dinyatakan masuk sebagai anggota *cluster* tersebut apabila mempunyai jarak yang paling dekat dengan pusat *cluster*. Selanjutnya menentukan pusat *cluster* yang baru dengan persamaan (1)
5. Mengulangi tahapan ke 3 apabila pusat *cluster* yang baru berbeda.

Algoritma *K-Means* dinyatakan selesai apabila anggota cluster sebelumnya sama dengan anggota cluster selama pengulangan. Jika belum dilakukan pengulangan sampai hasilnya sama.

**Koefisien Silhouette**

Koefisien *silhouette* adalah metode yang digunakan untuk mengevaluasi *cluster* untuk mengetahui ketepatan pengelompokkan. Untuk menghitung Nilai Koefisien *silhouette* dirumuskan sebagai berikut :

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

Dengan  $S(i)$  adalah Koefisien *silhouette*,  $a(i)$  adalah rata-rata objek ke- $i$  dengan seluruh objek pada *cluster* yang sama, dan  $b(i)$  adalah rata-rata objek ke- $i$  dengan seluruh objek pada *cluster* yang berbeda. Terdapat kriteria koefisien Silhoutte yang ditetapkan oleh (Kaufman, 1990) sebagai berikut :

Tabel 1. Kriteria Koefisien *silhouette*

Nilai Koefisien Silhoutte	Kriteria Cluster
0,71-1,00	Strong

Nilai Koefisien Silhoutte	Kriteria Cluster
0,51-0,70	Good
0,26-0,50	Weak
0,00-0,25	Bad

**Metode Elbow**

Dalam menentukan banyaknya kelompok yang optimal salah satu metode yang dapat digunakan adalah Metode *Elbow*. Untuk menentukan jumlah *cluster* berikut disajikan algoritma *Elbow* yang didasarkan pada sum of square error (SSE) sebagai berikut :

$$SSE = \sum_{k=1}^k \sum_{x_i \in S_k} \|X_i - C_k\|^2 \tag{4}$$

Dimana  $k$  merupakan jumlah kelompok yang digunakan pada algoritma *k-means*,  $X_i$  merupakan banyaknya data, dan  $C_k$  ialah jumlah clusteri pada cluster ke  $k$

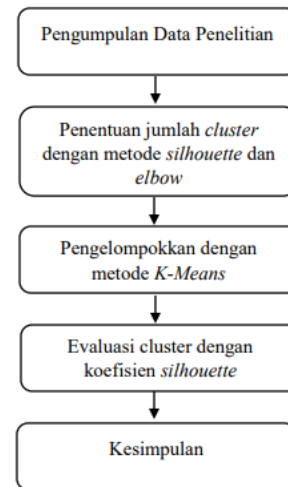
**3. METODE**

**Data Penelitian**

Pada penelitian ini data yang digunakan adalah data penambahan jumlah kasus terkonfirmasi Covid-19 di 38 Kabupaten/Kota di Jawa Timur. Penelitian ini bertujuan untuk mengelompokkan data *time series*, sehingga diambil data jumlah kasus terkonfirmasi Covid-19 setiap harinya dari tanggal 19 Maret 2020 sampai 31 Januari 2021. Data tersebut diperoleh dari website resmi pemerintah provinsi Jawa Timur yaitu <http://infocovid19.jatimprov.go.id/>.

**Diagram Alir Penelitian**

Untuk mempermudah pemahaman alur dalam penelitian ini, diberikan diagram alir sebagai berikut:



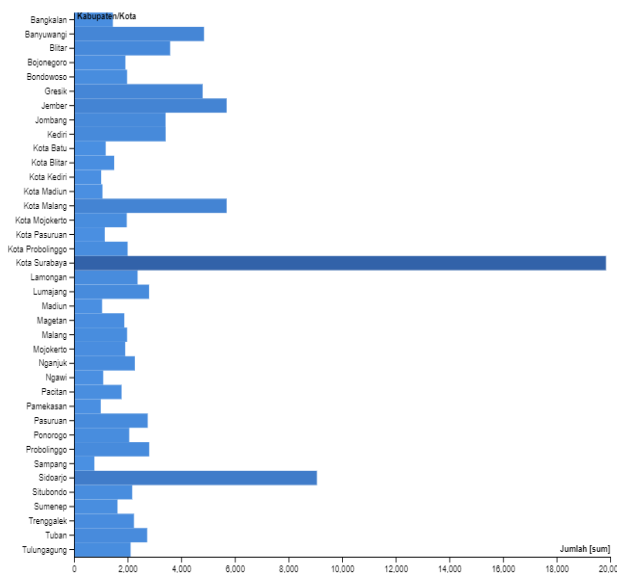
Gambar 1. Diagram Alir Penelitian

Berdasarkan Gambar 1, tahapan penelitian ini yaitu pertama pengumpulan data penambahan jumlah kasus terkonfirmasi Covid-19 pada masing-masing kabupaten/kota di Jawa Timur dari tanggal 19 Maret 2020 sampai dengan 31 Januari 2021. Selanjutnya, sebelum

dilakukan pengelompokan maka menentukan jumlah cluster terlebih dahulu menggunakan metode silhouette dan elbow. Kemudian dilakukan pengelompokan menggunakan metode *K-Means* dengan tahapan seperti yang sudah dijelaskan pada kajian teori. Setelah didapatkan hasil pengelompokan, maka akan dilakukan *cluster validation* dengan koefisien *silhouette*. Nilai koefisien *silhouette* berada pada rentang -1 sampai 1, apabila bernilai 1 maka data sudah dikelompokkan di cluster yang tepat dan begitu sebaliknya.

**4. PEMBAHASAN**

Pada bagian ini akan dijelaskan mengenai pengelompokan 38 Kabupaten/Kota di Jawa Timur berdasarkan karakteristik penambahan jumlah kasus terkonfirmasi Covid-19 dengan data berupa *time series*. Untuk mendapatkan petunjuk awal mengenai karakteristik penyebaran Covid-19 di Kabupaten/Kota di Jawa Timur, Gambar 2 menampilkan grafik jumlah kumulatif kasus terkonfirmasi Covid-19 sampai tanggal 31 Januari 2021 sebagai berikut :

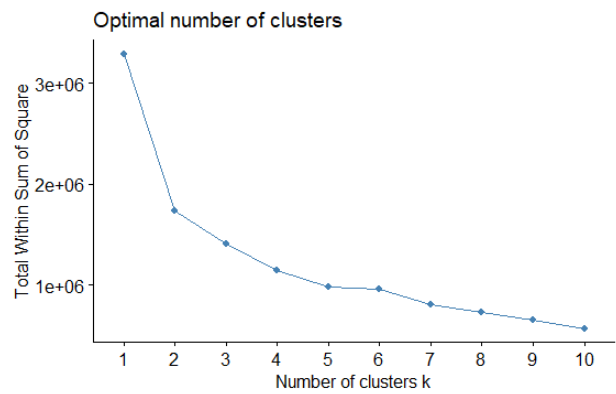


Gambar 2. Grafik Kasus Positif Covid-19 di Jawa Timur

Dari gambar 2 dapat diketahui bahwa Kota Surabaya merupakan daerah di Jawa Timur dengan jumlah kasus positif Covid-19 terbesar dan disusul oleh daerah lainnya yaitu kabupaten Jember dan Kota Malang. Untuk daerah yang mempunyai jumlah kasus positif Covid-19 terkecil yaitu Kabupaten Sampang.

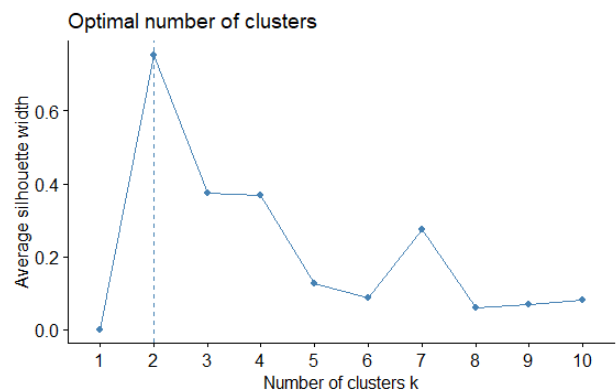
**Penentuan Jumlah Cluster**

Dalam menentukan jumlah kelompok, pada penelitian ini akan digunakan dua metode yaitu *silhouette* dan *elbow*. Diperoleh hasil sebagai berikut :



Gambar 3. Grafik Jumlah Cluster Dengan Metode *elbow*

Berdasarkan Gambar 3, dengan metode *elbow* diperoleh hasil bahwa cluster optimum yang dapat dibentuk k=2, karena nilai k yang mengalami penurunan secara drastis dan membentuk siku berada di k=2. Untuk cluster optimal yang kedua yaitu pada k=3.



Gambar 4. Grafik Jumlah Cluster Dengan Metode *Silhouette*

Berdasarkan Gambar 4 dengan Metode *Silhouette* garis putus-putus menunjukkan bahwa cluster optimum yaitu k=2.

Dari hasil yang diperoleh dari metode *silhouette* dan *elbow* dapat disimpulkan bahwa pada penelitian ini akan digunakan k=2 dan sebagai pembanding juga dilakukan pengelompokan dengan k=3.

**Clustering dengan K-Means**

Setelah didapatkan jumlah kelompok, selanjutnya pada bagian ini akan dilakukan pengelompokan dengan metode *K-Means*. Dengan bantuan R studio pengelompokan *K-Means* dengan jumlah k=2 diperoleh hasil sebagai berikut :

Tabel 2. Hasil Clustering dengan k=2

Cluster	Anggota
1	Kota Surabaya

<i>Cluster</i>	<i>Anggota</i>
2	Banyuwangi, Lamongan, Bangkalan, Gresik, Probolinggo, Jombang, Kota Malang, Pasuruan, Bojonegoro, Magetan, Jember, Bondowoso, Madiun, Nganjuk, Kota Mojokerto, Mojokerto, Blitar, Kota Kediri, Kediri, Pamekasan, Kota Batu, Sampang, Kota Pasuruan, Situbondo, Ngawi, Pacitan, Malang, Ponorogo, Lamongan, Kota Probolinggo, Sumenep, Lumajang, Trenggalek, Sidoarjo, Tuban, Kota Blitar, Tulungagung

Berdasarkan karakteristik data penambahan jumlah kasus positif Covid-19, dapat disimpulkan bahwa *cluster 1* merupakan Kabupaten/Kota dengan penambahan jumlah kasus positif Covid-19 tinggi dan *cluster 2* merupakan Kabupaten/Kota dengan penambahan jumlah kasus positif Covid-19 cenderung rendah.

Selanjutnya akan dilakukan clustering dengan jumlah  $k=3$  diperoleh hasil sebagai berikut :

Tabel 3. Hasil *Clustering* dengan  $k=3$

<i>Cluster</i>	<i>Anggota</i>
1	Kota Surabaya
2	Banyuwangi, Gresik, Jember, Kediri, Kota Malang, dan Mojokerto
3	Lamongan, Bangkalan, Kota Probolinggo, Jombang, Pasuruan, Bojonegoro, Magetan, Bondowoso, Madiun, Nganjuk, Kota Mojokerto, Blitar, Kota Kediri, Pamekasan, Kota Batu, Sampang, Kota Pasuruan, Situbondo, Ngawi, Pacitan, Malang, Ponorogo, Lamongan, Probolinggo, Sumenep, Lumajang, Trenggalek, Sidoarjo, Tuban, Kota Blitar, Tulungagung

Berdasarkan karakteristik data penambahan jumlah kasus positif Covid-19, dapat disimpulkan bahwa *cluster 1* merupakan Kabupaten/Kota dengan penambahan jumlah kasus positif Covid-19 tinggi, *cluster 2* penambahan kasus sedang, dan *cluster 3* penambahan kasus rendah.

**Cluster Validation**

Setelah dilakukan pengelompokkan, pada bagian ini akan dilakukan validasi *cluster* dengan koefisien silhouette. Hal ini bertujuan untuk mengetahui *cluster* yang paling optimal. Dengan bantuan R studio diperoleh hasil sebagai berikut :

Tabel 4. Nilai Validasi

<i>Jumlah k</i>	<i>Koefisien Silhouette</i>
2	0.75
3	0.37

Dari nilai koefisien silhouette dapat disimpulkan bahwa *cluster* yang paling optimal untuk mengelompokkan 38 Kabupaten/Kota di Jawa Timur yang didasarkan pada penambahan kasus konfirmasi Covid-19 yaitu dengan jumlah  $k=2$ . Dengan nilai koefisien silhouette sebesar 0.75, menurut kriteria koefisien silhouette pada Tabel 1 *cluster* dikategorikan *Strong cluster*.

**5. PENUTUP**

**Simpulan**

Berdasarkan hasil analisis, dapat disimpulkan bahwa *cluster* yang paling optimal untuk mengelompokkan 38 Kabupaten/Kota di Jawa Timur berdasarkan penambahan kasus konfirmasi Covid-19 yaitu dengan jumlah  $k=2$ . Nilai koefisien silhouette sebesar 0.75 sehingga *cluster* yang terbentuk dalam kategori *strong cluster*. Dengan jumlah  $k=2$  diperoleh bahwa *cluster 1* merupakan Kabupaten/Kota dengan penambahan jumlah kasus positif Covid-19 tinggi yaitu Kota Surabaya dan 37 provinsi lainnya masuk dalam *cluster 2* dengan penambahan jumlah kasus positif Covid-19 yang rendah.

**Saran**

Dalam penelitian selanjutnya, metode *clustering* yang digunakan dapat lebih dari 1. Hal ini bertujuan untuk membandingkan metode *clustering* untuk menghasilkan *cluster* yang paling optimal.

**DAFTAR PUSTAKA**

A. D. Munthe. (2019). Penerapan clustering time series untuk menggerombolkan provinsi di Indonesia berdasarkan nilai produksi padi. *Litbang Sukowati*, (2), 1–11.

Gugus Tugas Percepatan Penanganan COVID-19. (2020). No T. Retrieved from <https://covid19.go.id/>

Johnson, R., & Wichern, D. (2007). *Clustering, Distance Methods, And Ordination, In R. Applied Multivariate Statistical Analysis*. Upper Saddle River : Pearson Prentice Hall.

Kaufman, L. & R. (1990). *Finding Groups in Data An Introduction to Cluster Analysis*. New Jersey: John Wiley & Sons Inc Publication.

M. Habibi et al. (2019). Clustering User Characteristics Based on the influence of Hastags on the Instagram Platform. *Indones. J. Comput. Cybern. Syst.*, 13 no 4, 399–408.

M.Z. Rodriguez et al. (2019). *Clustering Algorithms: A comparative approach*. 14 no 1, 1–37.

- Rencher, A. C. (2002). *Methods of Multivariate Analysis Second Edition*. Canada: A John Wiley & Sons, Inc. Publication.
- Vasilios Zarikas, Stavros G. Pouloupoulos, Zoe Gareiou, E. Z. (2020). Clustering analysis of countries using the Covid-19 cases dataset. *Pre-Proof*.
- World Health Organization. (2020). Media Statement: Knowing the risks for COVID-19.