LAPORAN
PENELITIAN TERAPAN
UNIVERSITAS LAMPUNG

IMPLEMENTASI METODE CLUSTER NONHIERARKI PADA PEMETAAN SEBARAN DATA COVID-19 DI INDONESIA

KATEGORI
Penelitian Terapan

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
2021

HALAMAN PENGESAHAN

Judul Penelitian	:	Implementasi metode cluster nonhierarki pada pemetaan sebvaran data COVID-19 di Indonesia tahun 2020
Manfaat sosial ekonomi	:	Menjadi bahan pertimbangan dalam mengambil kebijakan Penanggulangan Covid-19 di Indonesia
Jenis penelitian	:	Penelitian terapan
Ketua Peneliti		
a. Nama Lengkap	:	Dr. Ir. Netti Herawati, M.Sc.
b. NIDN	:	0025016503
c. SINTA ID	:	6169478
d. Jabatan fungsional	:	Lektor Kepala
e. Program studi	:	Matematika
f. Nomo HP/email	:	081273809624/netti.herawati@fmipa.unila.ac.id
Anggota Peneliti (1)		
a. Nama Lengkap	:	Dr. Khoirin Nisa, M.Si.
b. NIDN	:	0026077401
c. SINTA ID	:	6050683
d. Jabatan fungsional	:	Lektor
e. Program studi	:	Matematika
f. Nomo HP/email	:	081379846402/ nisa.mahfudh@gmail.com
Anggota Peneliti (2)		
a. Nama Lengkap	:	Subian Saidi, M.Si.
b. NIDN	:	0021088004
c. SINTA ID	:	6681591
d. Jabatan fungsional	:	Asisten Ahli
e. Program studi	:	Matematika
f. Nomo HP/email	:	081366351121/ subian.saidi@fmipa.unila.ac.id
Jumlah mahasiswa yang terlibat	:	2
Jumlah alumni yang terlibat	:	1
Jumlah staf yang terlibat	:	1
Lokasi kegiatan	:	Lab. Matematika dan Statistika Terapan FMIPA Unila
Lama kegiatan	:	6 (enam) bulan
Biaya penelitian	:	Rp. 35.000.000;
Sumber dana	:	BLU Unila

Bandar Lampung, 21 September 2021

Mengetahui,
Dekan FMIPA Unila

Dr. Suripto Dwi Yuwono, M.Sc.
19740705200003100[illegible]

Ketua Peneliti,

Dr. Netti Herawati, M.Sc.
196501251990032001

Menyetujui,
Ketua LPPM Universitas Lampung

Dr. Lusmeilia Afriani, D.E.A
NIP. 196505101993032008

IDENTITAS DAN URAIAN UMUM

1. Judul Penelitian : Implementasi metode cluster nonhierarki pada pemetaan sebaran data COVID-I9 di Indonesia tahun 2020

2. Tim Peneliti :

No	Nama	Jabatan	Bidang Keahlian	Program Studi	Alokasi Waktu (jam/ minggu)
1	Dr. Netti Herawati	Ketua	Statistika Terapan	Matematika	48
2	Dr. Khoirin Nisa	Anggota 1	Statistika	Matematika	30
3	Subian Saidi, S.Si., M.Si	Anggota 2	Matematika	Matematika	20

3. Objek Penelitian (Jenis Material yang akan diteliti dan segi penelitian):
   Data covid-19 di Indonesia

4. Masa Pelaksanaan
   Mulai : Bulan 27 April Tahun 2021
   Berakhir : 21 September Tahun 2021

5. Usulan Biaya : Rp.35.000.000,00

6. Lokasi Penelitian : Lab Matematika dan Statistika Terapan FMIPA Unila

7. Kontributor mendasar dari hasil penelitian ini adalah menambah khasanah keilmuan statistika tentang efisiensi metode dalam menangani pencilan pada data kesehatan untuk dilakukan pengklasteran

8. Sasaran jurnal: scopus

DAFTAR ISI

DAFTAR ISI

RINGKASAN

IMPLEMENTASI METODE CLUSTER NONHIERARKI PADA PEMETAAN SEBARAN DATA COVID-19 DI INDONESIA TAHUN 2020

Coronavirus (COVID-19) adalah suatu penyakit menular yang ditemukan pada tahun 2019. Kasus virus ini ditemukan pertama kali di provinsi Wuhan dengan gejala antara lain batuk, demam, letih, sesak nafas, dan mengalami penurunan nafsu makan. Saat ini COVID-19 telah tersebar ke lebih dari 190 negara salah satunya yaitu negara Indonesia. Penyebaran COVID-19 yang merata di seluruh provinsi di Indonesia, merupakan penyebaran yang cukup cepat dan berdampak negatif. Luasnya wilayah Indonesia memungkinkan diperlukannya pengklasteran provinsi di Indonesia berdasarkan penyebaran COVID-19. Pengklasteran ini akan menghasilkan titik-titik pusat penyebaran kasus COVID-19 yang nantinya dapat dievaluasi dan dijadikan suatu informasi. Salah satu cabang dari ilmu statistika yang membahas tentang metode pengklasteran tersebut adalah analisis klaster. Analisis klaster merupakan suatu teknik analisis multivariat yang berguna untuk mengelompokkan data observasi ataupun variabel-variabel ke dalam klaster sedemikian rupa sehingga masing-masing klaster bersifat homogen sesuai dengan faktor yang digunakan untuk melakukan pengklasteran.

Penelitian ini memiliki tujuan melakukan pengklasteran provinsi berdasarkan kasus positif, kasus meninggal, dan kasus sembuh COVID-19 di Indonesia. Dari hasil tersebut terlihat bahwa metode trimmed k-means mampu memisahkan outlier pada data dan memberikan cluster yang optimal. Hasil ini konsisten dengan penelitian sebelumnya yang telah melaporkan metode pengelompokan k-means yang dipangkas kuat untuk outlier [16-17,24-25, 31]. Terkait data pandemi Covid-19 di Indonesia, metode trimmed k-means clustering membentuk 3 cluster dan memisahkan outlier dari 3 cluster tersebut sehingga membentuk cluster tersendiri. Klaster 1 terdiri dari 14 provinsi dan Klaster 2 dan 3 masing-masing terdiri dari 10 dan 6 provinsi. Di satu sisi, provinsi outlier yaitu DKI Jakarta, Jawa Barat, Jawa Tengah, dan Jawa Timur membentuk klaster tersendiri di luar 3 klaster di atas. Hal ini membuktikan bahwa metode trimmed k-means clustering bekerja sangat baik dalam mengelompokkan data yang mengandung outlier dan dapat digunakan pada jenis data yang sejenis.

Kata kunci: covid-19, klaster, *trimmed k-means*

BAB 1. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Coronavirus (COVID-19) adalah suatu penyakit menular yang ditemukan pada tahun 2019. Orang-orang yang terinfeksi virus ini akan mengalami penyakit pernapasan dari kategori ringan hingga menengah dan dapat sembuh tanpa harus ada perawatan khusus. Kasus virus ini ditemukan pertama kali di provinsi Wuhan dengan gejala antara lain batuk, demam, letih, sesak nafas, dan mengalami penurunan nafsu makan [1].

Saat ini COVID-19 telah tersebar ke lebih dari 190 negara salah satunya yaitu negara Indonesia. Penyebaran COVID-19 yang merata di seluruh provinsi di Indonesia, merupakan penyeberan yang cukup cepat dan berdampak negatif. Luasnya wilayah Indonesia memungkinkan diperlukannya pengklasteran provinsi di Indonesia berdasarkan penyebaran COVID-19. Pengklasteran ini akan menghasilkan titik-titik pusat penyebaran kasus COVID-19 yang nantinya dapat dievaluasi dan dijadikan suatu informasi. Salah satu cabang dari ilmu statistika yang membahas tentang metode pengklasteran tersebut adalah analisis klaster [2].

Analisis klaster merupakan suatu teknik analisis multivariat yang berguna untuk mengelompokkan data observasi ataupun variabel-variabel ke dalam klaster sedemikian rupa sehingga masing-masing klaster bersifat homogen sesuai dengan faktor yang digunakan untuk melakukan pengklasteran [3]. Analisis klaster dibagi menjadi dua metode yaitu metode hierarki (*hierarchical clustering methods*) dan metode non-hierarki (*non-hierarchical clustering methods*). Perbedaan antara kedua metode tersebut terletak pada penentuan jumlah klaster yang akan dihasilkan. Metode hierarki digunakan apabila jumlah klaster yang diinginkan belum diketahui,

sedangkan metode non-hierarki digunakan apabila jumlah klaster yang diinginkan telah ditentukan sebelumnya. Dalam penelitian ini akan digunakan metode non-hierarki. Metode non-hierarki yang akan digunakan dalam penelitian ini yaitu metode *trimmed k-means*. Kelebihan metode *trimmed k-means* terletak pada tingkat keakuratan yang tinggi dan waktu komputasi yang cepat [3,4] kelebihan metode *trimmed k-means* yaitu kekar terhadap pencilan.

Hal-hal tersebut yang kemudian menjadi dasar untuk mengelompokkan provinsi berdasarkan kasus positif, kasus meninggal, dan kasus sembuh COVID-19 di Indonesia. Dengan pengklasteran ini diharapkan dapat membantu pemeritah dalam pengklasteran wilayah berdasarkan zona sehingga pemerintah dapat lebih mudah dalam menangani kasus COVID -19 di Indonesia.

1.2 Tujuan Khusus

Tujuan khusus penelitian ini adalah pengklasteran provinsi di Indonesia berdasarkan kasus positif, kasus meninggal, dan kasus sembuh COVID-19 di Indonesia menggunakan metode *trimmed k-means*.

1.3 Keutamaan Penelitian

Keutamaan penelitian ini adalah sebagai berikut:

1. Mengetahui klaster provinsi di Indonesia berdasarkan zona.
2. Mengetahui perubahan kelompok provinsi pada 5 bulan pertama dan 5 bulan terakhir.
3. Membantu pemerintah dalam pengklasteran wilayah berdasarkan zona.
4. Memudahkan pemerintah dalam menangani kasus COVID-19 di Indonesia.

1.4 Target Temuan Penelitian

Target temuan utama dari penelitian ini adalah diperoleh klaster provinsi di Indonesia pada kasus positif, kasus meninggal, dan kasus sembuh COVID-19 di Indonesia dari maret 2020-Juli 2021 sehingga dapat digunakan dalam penentuan zona.

Roadmap Penelitian

BAB 2. TINJAUAN PUSTAKA

2.1 Coronavirus (COVID-19)

Coronavirus (COVID-19) adalah suatu penyakit menular yang ditemukan pada tahun 2019. Orang-orang yang terinfeksi virus ini akan mengalami penyakit pernapasan dari kategori ringan hingga menengah dan dapat sembuh tanpa harus ada perawatan khusus. Kasus virus ini ditemukan pertama kali di provinsi Wuhan dan beberapa gejala yang dialami apabila terinfeksi virus ini antara lain batuk, demam, letih, sesak nafas, dan mengalami penurunan nafsu makan [1]. Secara umum virus ini dapat menular melalui droplet atau cairan tubuh yang dikeluarkan selama bersin dan batuk [5].

2.2 Analisis Klaster

Analisis klaster merupakan suatu teknik analisis multivariat yang bertujuan untuk mengklasterkan data observasi ataupun variabel-variabel ke dalam klaster sedemikian rupa sehingga masing-masing klaster bersifat homogen sesuai dengan faktor yang digunakan untuk melakukan pengklasteran. Data mengenai ukuran kesamaan tersebut dapat dianalisis dengan analisis klaster sehingga dapat ditentukan siapa yang masuk klaster mana [2,3,4,6].

2.3 Ukuran Kemiripan Objek (jarak)

Dalam pengklasteran digunakan suatu ukuran yang dapat menerangkan keserupaan atau kedekatan antardata, yaitu ukuran jarak atau similaritas. Ukuran jarak yang biasa digunakan dalam analisis klaster yaitu jarak *Euclidean*. Rumus jarak *euclidean* dinyatakan sebagai berikut:

$$d_{ij} = d(x_i, x_j) = \sum_{k=1}^{p} \|x_{ik} - x_{jk}\| = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \quad (2.1)$$

dimana:

d_{ik} = jarak objek ke-*i* dengan objek ke-*j*

p = jumlah variabel

x_{ik} = data dari objek ke-i pada variabel ke-k

x_{jk} = data dari objek ke-j pada variabel ke-k

2.4 Metode Non-hirarki

Metode non-hirarki dimulai dengan menentukan terlebih dahulu jumlah klaster yang diinginkan. Setelah jumlah klaster diketahui, kemudian proses klaster dilakukan tanpa mengikuti proses hirarki [7]

2.4.1 Metode *Trimmed k-means*

K-means yang dipangkas diperkenalkan oleh [16]. Dengan menggunakan metode ini, seseorang diperbolehkan untuk menghilangkan proporsi tertentu dari kemungkinan outlier ketika mengelompokkan hasil [24,25]; proporsi yang akan dihilangkan adalah besarnya data rate yang akan dipangkas dalam metode ini. Metode trimmed k-means digunakan untuk mengatasi outlier yang terdapat pada suatu cluster data yang akan dikelompokkan. Konsep utama dari metode trimmed k-means adalah membentuk k cluster baru dengan cara menghilangkan, atau memangkas outlier yang terdapat pada data. Metode tersebut termasuk dalam kelas prosedur berdasarkan "pemangkasan tidak memihak" (ditentukan oleh data) dengan tujuan membuat teknik pengelompokan hierarki klasik yaitu k-means, lebih kuat. Selanjutnya, metode k mean umum didasarkan pada minimalisasi perbedaan antara variabel acak (atau sampel variabel acak ini) dan himpunan dengan k poin yang diukur, menggunakan fungsi penalti Φ [16].

Misal α berada pada interval terbuka yaitu $\alpha \in (0,1)$, k adalah nagka real. Dan Φ adlah fungsi penalti. Untuk setiap set A pada $P(A) \geq 1-\alpha$ dan setiap k-set $M = \{m_1, m_2, \ldots, m_k\}$ dalam R^p, lmaka variasi dari *M* pada *A:*

$$V_{\Phi}^{A}(M) = \frac{1}{P(A)} \int_{A} \Phi\left(\inf_{i\,=\,1,\ldots,k} \|X - m_i\| \right) dP$$

Nilai $V_{\Phi}^{A}(M)$ mengukur berapa baik M mewakili peluang mewakili massa probabilitas dari P ang ditentukan pada*A*, an tugas kita adalah memilih sekumpulan massa probabilitas yang diberikan sedemikian rupa sehingga variasinya diminimalkan. Hal ini dilakukan dengan meminimalkan $V_{\Phi}^{A}(M)$ pada *A* dan *M* dengan cara:

1. Cari nilai variasi-*k*- pada A, $V_{\mathrm{k},\Phi}^{A}(M)$, dengan meminimumkan *M*:

$$V_{\mathrm{k},\Phi}^{A}(M) = \inf_{M \subset R^{p}, \#M \neq k} \; V_{\Phi}^{A}(M);$$

2. CAri variasi trimmed *k*-variation, $V_{k,\Phi,\alpha}$, dengan meminimumkan *A*:

$$V_{k,\Phi,\alpha} \coloneqq V_{k,\Phi,\alpha}(X) \coloneqq V_{k,\Phi,\alpha}(P_X) \coloneqq \inf_{A \in \beta^{p}, P(A) \geq 1-\alpha} \; V_{\mathrm{k},\Phi}^{A}.$$

Kita ingin trimmed set A_0 dan k-set $M_0 = \{m_1^0, m_2^0, \ldots m_k^0\}$ $dengan\ kondisi$ $V_{\Phi}^{A_0}(M_0) = V_{k,\Phi,\alpha}$ [16].

Principal Component Analysis (PCA) adalah teknik statistik multivariat yang bertujuan untuk mengurangi dimensi data untuk mendapatkan variabel baru yang tidak berkorelasi dan mempertahankan sebagian besar informasi yang terkandung dalam variabel asli. Variabel yang dihasilkan merupakan kombinasi linear dari variabel asli dan disebut sebagai komponen utama (PC). Jumlah kuadrat koefisien dalam kombinasi linier sama dengan satu, dan PC-nya ortogonal.

Misal $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ adalah variable acak dari normal multivariate dengan nilai tengah vector $\boldsymbol{\mu} = (\mu, \mu, \ldots, \mu)$ dan matrik kovarian $\boldsymbol{\Sigma}$ serta *k* independent eigenvectors, $\boldsymbol{a}_k$, $k = 1,2, \ldots, k$. Principal component dapat ditulis sebagai::

$$\boldsymbol{Y = A(X - \mu)}$$

denagn **A** adalahmatriks *p*x *p* dari variable $\boldsymbol{X}$ pada *n*-element variable y. Vektor kolom **A** adlaah **Σ**, i.e. $\boldsymbol{A} = [\boldsymbol{a_1} \mid \boldsymbol{a_2} \mid \ldots \mid \boldsymbol{a_k}]$. Maka i-th principal component adalah:

$$y_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{ki}X_k = \boldsymbol{a_i^T X}$$

2.5 *Silhoutte Index*

Validasi jumlah klaster dilakukan dengan menggunakan *cluster validity index* atau indeks validitas. Indeks validitas berfungsi mengukur derajat kekompakan dan separasi struktur data pada seluruh klaster dan menemukan jumlah klaster optimal yang kompak dan terpisah dari klaster yang lain [10].

Metode validasi *silhouette index* merupakan salah satu ukuran validasi yang berbasis kriteria internal. *Silhouette index* akan mengevaluasi penempatan setiap objek dalam setiap klaster dengan membandingkan jarak rata-rata objek dalam satu klaster dan jarak antara objek dengan klaster yang berbeda [13]. Cara menghitung koefisien *silhouette* yang didefinisikan sebagai rata-rata **s(i)** yaitu:

$$\boldsymbol{SC} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{i}) \tag{2.6}$$

dengan

$$\boldsymbol{s}(\boldsymbol{i}) = \frac{\boldsymbol{b}(\boldsymbol{i})-\boldsymbol{a}(\boldsymbol{i})}{max(\boldsymbol{a}(\boldsymbol{i}),\boldsymbol{b}(\boldsymbol{i}))}; \quad \boldsymbol{b}(\boldsymbol{i}) = \min \boldsymbol{d}(\boldsymbol{i},\boldsymbol{C}); \quad \boldsymbol{a}(\boldsymbol{i}) = \frac{1}{|\boldsymbol{A}|-1}\sum_{j\in A, \in j\neq i} \boldsymbol{d}(\boldsymbol{i},\boldsymbol{j})$$

Hasil perhitungan nilai koefisien *silhouette index* berada pada *range* -1 sampai 1. Semakin besar nilai koefisien *silhouette* akan semakin baik kualitas suatu kelompok.

BAB 3. METODE

3.1 Tempat dan Waktu Penelitian

Penelitian ini dilaksanakan di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung. Waktu pelaksanaan penelitian ini dilakukan dari bulan Februari sampai Juli 2021.

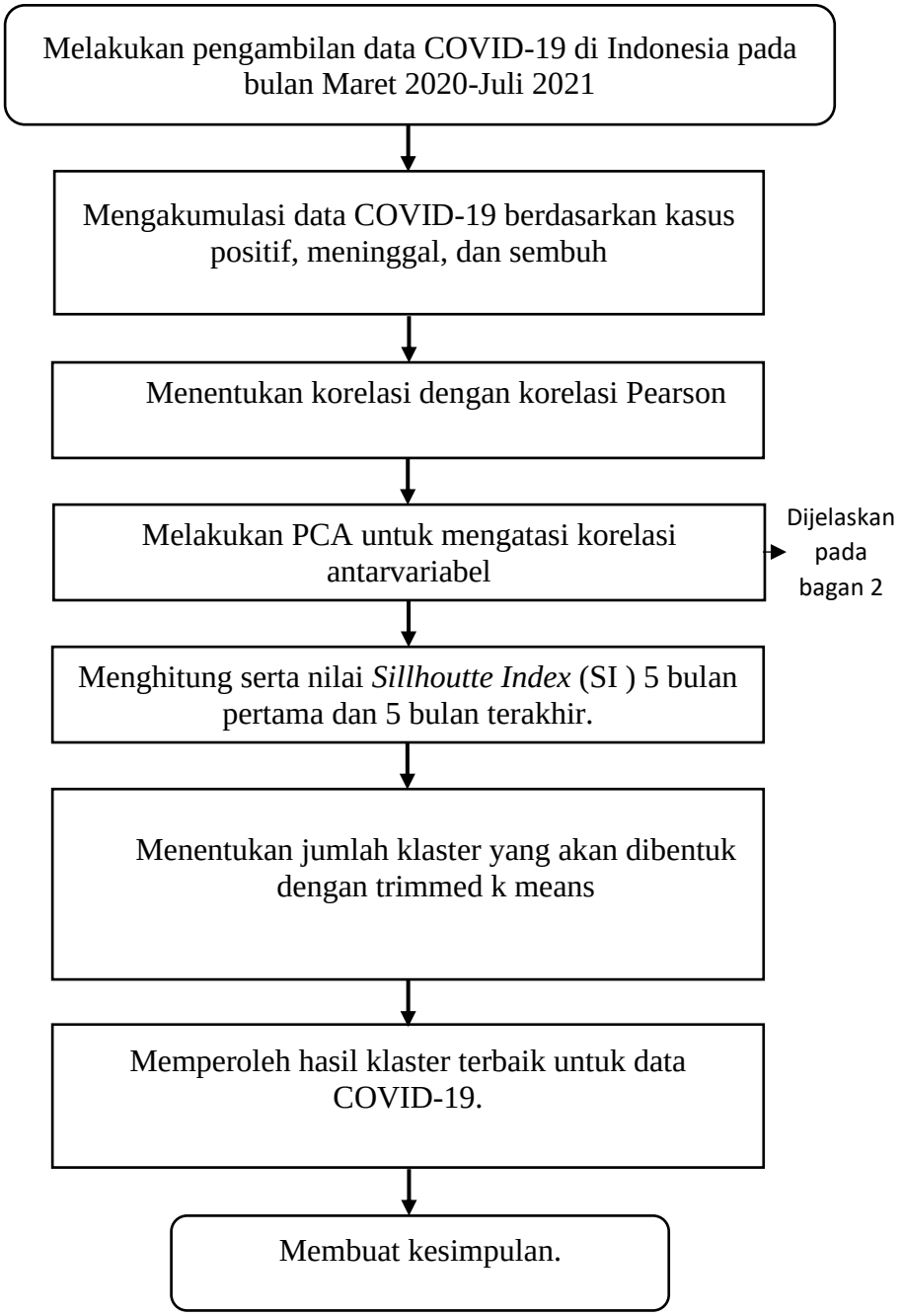
3.2 Data Penelitian

Data yang digunakan dalam penelitian ini adalah data COVID-19 Indonesia dari bulan Maret 2020 sd Juli 2021 yang diperoleh dari laporan analisis data COVID-19 yang disusun oleh Tim Pakar Satuan Tugas Penanganan COVID-19 dengan data bersumber dari Kementrian Kesahatan Republik Indonesia. Dengan objek 34 provinsi dan 3 variabel yaitu angka positif COVID-19, angka kematian akibat COVID-19, dan angka sembuh dari COVID-19.

3.3 Tahapan Penelitian

Tahapan penelitian ini dilakukan dengan langkah-langkah yang dijelaskan melalui bagan alir sebagai berikut

Bagan 1. Tahapan umum analisis klaster



IV.HASIL DAN PEMBAHASAN

4.1 Deskriptif analisis data Covid-19 di Indonesia

Berikut adalah hasil analisis dereskriptif dari data covid-19 di 34 provinsi di Indonesia yang diambil dari 23 Maret 2020- 23 Juli 2021.

Table 1. Summary of COVID-19 cases of 34 Province in Indonesia

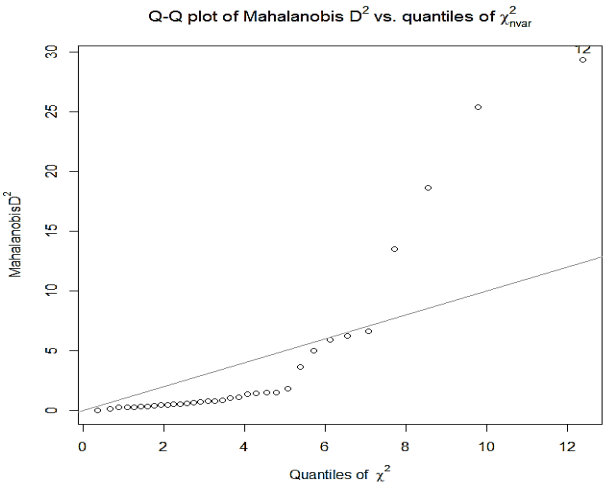
Descriptive Statistics	Confirmed Cases	Death Cases	Recovered Cases	MIR
Minimum	7103	149	5930	0.897
1st Quartile	17790	314	13393	1.710
Median	31069	775	22580	2.250
Mean	90655	2363	71084	2.531
3rd Quartile	73331	1706	57341	2.766
Maximum	778521	17512	678992	6.568
Standard deviation	164523.3	4292.167	136554.595	1.180

Bila kita lihat dari nilai simpangan baku yang tinggi, kita bias mengatakan bahwa kemungkinan adanya outliers pada data sangat besar. Untuk meyakinkan hal tersebut kita melakukan pendeteksian keberadaan outliers dnegan menggunakan Malahanobis squared distance method.

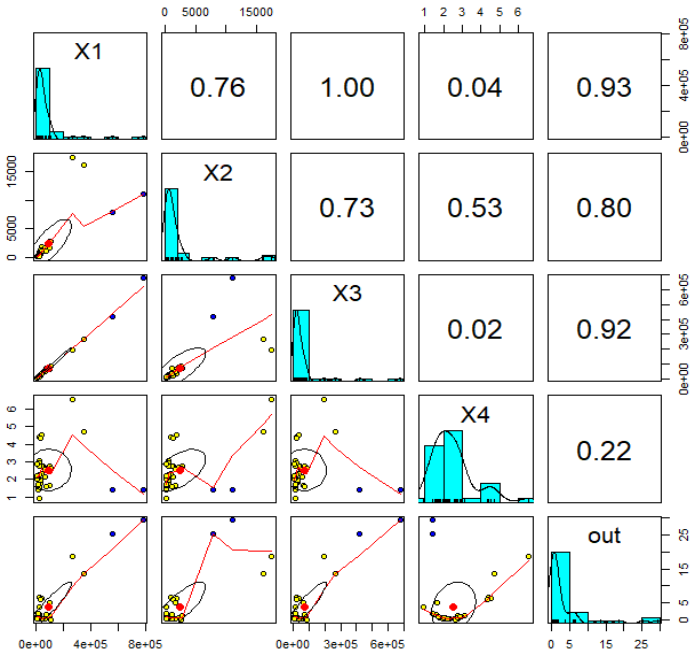
4.2 Pendeteksian Pencilan

Pendeteksian pencilan yaitu metode jarak kuadrat Mahalanobis. Pengamatan ke- i teridentifikasi pencilan jika $d_{MD}^2(i) > \chi_{p,1-\alpha}^2$. Berdasarkan kasus ini, p merupakan banyaknya variabel yang diteliti yaitu 3 variabel dan nilai α yang digunakan sebesar 5%. Maka diperoleh nilai $\chi_p^2 = 3.(1 - 0.05)$ sebesar 7,8150. Berdasarkan pendeteksian pencilan dengan membandingkan hasil jarak kuadrat Mahalanobis tiap objek dan nilai$\chi_{4,0,95}^2$, diketahui bahwa terdapat 4 provinsi yang merupakan

pencilan, yaitu DKI Jakarta, Jawa Barat, Jawa Tengah, dan Jawa Timur. Pada kasus ini data pencilan tetap disertakan dalam analisis klaster yang digunakan. Untuk Boxplot data COVID-19 di Indonesia dapat dilihat pada Gambar 1 sebagai berikut.



Gambar 1. Q-Q plot of Mahalanobis squared distance vs. chi-squared quantiles



Gambar 2. Pair panels plot of Covid-19 Data in Indonesia

Dari hasil uji Malahanonis squared distances kita dapat 4 outlier karena ada 4 titik yang ebrada diatas garis wilayah data. Berikut adalah list dari outliers dari provinsi yang mengalami penyebaran covid-19 lebih tinggi dari provinsi lain (Table 2).

Table 2. List of outliers in the Covid-19 Data in Indonesia

Province	Confirmed Cases	Death Cases	Recovered Cases	MIR
DKI Jakarta	778521	11131	678992	1.430
West Java	556181	7917	421977	1.423
Central Java	343210	16195	267511	4.719
East Java	266638	17512	194233	6.568

Kita juga perlu melakukan pendekteksian korelasi antarvariabel agar asumsi metode terpenuhi. Pada Tabel 3 kita lihat bahwa nilai pPearson korelasi dari variable ppositif, jumlah kematian, jumlah smebuh dan mortality incidence ratio cukup tinggi. Ini emnandakan ada korelasi anta variable.

Table 3. Correlation matrix between variables

	Confirmed Cases	Death Cases	Recovered Cases	MIR
Confirmed Cases	1.000	0.758	0.997	0.036
Death Cases	0.758	1.000	0.732	0.533
Recovered Cases	0.997	0.732	1.000	0.015
MIR	0.015	0.533	0.015	1.000

Untuk mengatasi maslaah korelasi, digunakan Principal Component Analysis sehingga korelasi antarvariabel menjadi rendah. Tabel 4 menyajikan hasil analisis dengan PCa untuk menurunkan nilai korelasi antarvariabel.

Table 4. Robust Principal Component Coefficients

Variables	PC1	PC2	PC3	PC4
Confirmed Cases	-0.561	-0.253	-0.099	0.781
Death Casess	-0.571	0.153	-0.671	-0.446
Recovered Cases	-0.558	-0.257	0.681	-0.398
MIR	-0.215	0.920	0.274	0.179

Dari Tabel 4 dapat dilihat bahwa maslaah korelasi antarvariabel telah teratasi dnegan menurunnya nilai korelasi. Selanjutnya kita bias melakukan analisis kluster dengan menggunakan trimmed k-means.

4.3 Analisis Klaster Menggunakan Metode *trimmed k-means*

Untuk menggunakan metode trimmed k-means, pertama kita cek dulu dengan silhouette index dan separation index jumlah klister yang ada dengan mengunakan program R. Dari hasil analisis dengan Silhoutte indek kita daopat bahwa ada 3 cluster yang terbentuk karena nilai kurva mencapai maksimum pada k=3 (gambar 4)

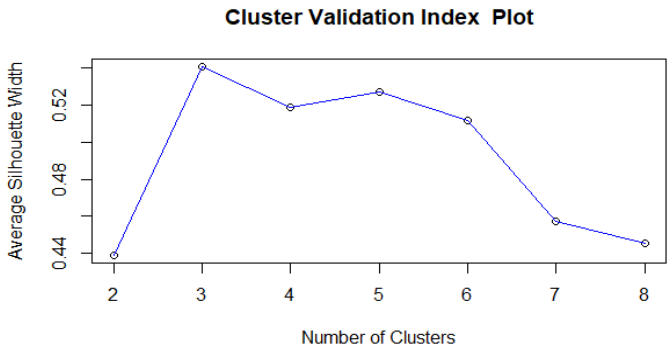


Figure 4. Plot of average silhouette index for *k*=2,…, 8

Selanjutnya adalah melakukan pemisahan sdengan menggunakan separation indeks sepeti tertera pada Gambar 5.

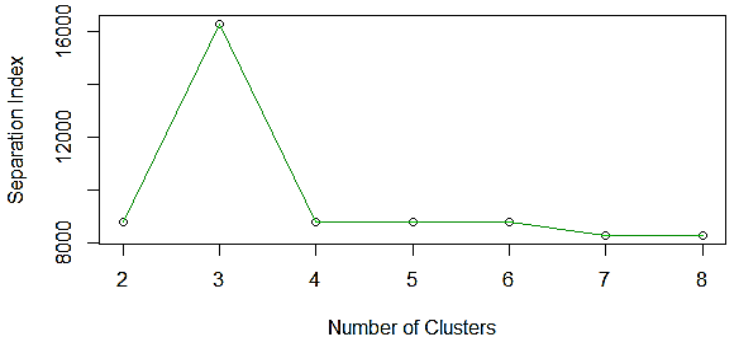


Figure 5. Plot of separation index for k=2, …, 8.

Kemudian diterapkan metode trimmed k-means dari hasil pemisahan tersebut untuk memeastikan berapa jumlah kluster yang seharusnya. HAsil analisis kluster dengan trimmed k-means dapat dilihat pada Gambar 6.

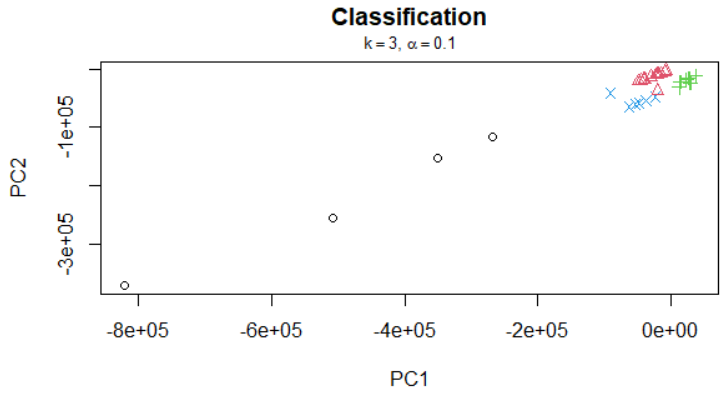


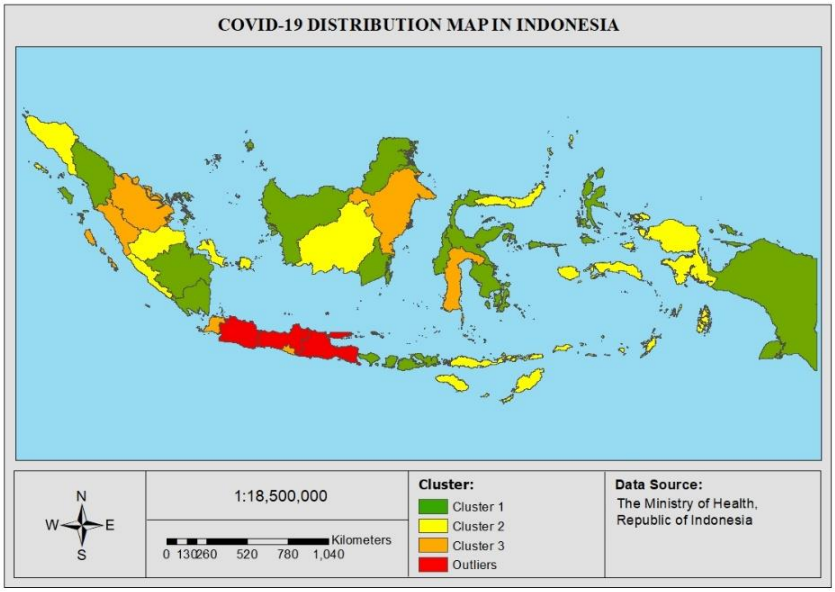
Figure 6. Trimmed *k*-means classification plot

Dari Gambar 6 dapat kita lihat bahwa trimmed k-means juga menghasilkan jumlah kluster sebanyajk 3 (k=3). Hal ini sesuai dnegan Silhoute indeks. Dengan menggabungkan seluruh data dengan jumlah kluster yang ada maka kita dpat 4 provinsi yang tidka bias dgabungkan ke dalam ketiga kluster karena keempat provinsi ini merupakan outliers. Oleh karena itu, ke empat provinsi tersbeut membentuk kluster sendiri. Deskripsi jumlah kluster, provinsi yang etrmasuk dalam kluster dan outliers bias dilihat pada Tabel 6.

Table 6. Anggota kluster

Cluster	Cluster size	Province
1	14	North Sumatera, South Sumatera, Lampung, Riau Island, Bali, West Nusa Tenggara, West Kalimantan, South Kalimantan, North Kalimantan, Central Sulawesi, South East Sulawesi, West Sulawesi, Maluku Utara, Papua
2	10	Aceh, Jambi, Bengkulu, Bangka Belitung Island, East Nusa Tenggara, Central Kalimantan, North Sulawesi, Gorontalo, Maluku, West Papua
3	6	West Sumatera, Riau, Banten, DI Yogyakarta, East Kalimantan, South Sulawesi
outliers	4	DKI Jakarta, West Java, Central Java, and East Java

Bila kita gambarkan dlaam map secara keseluruhan maka dpat dilihat pada gambar 7 berikut:



Gambar 6. Covid-19 cluster result map in Indonesia between March 2020 and July 2021

Selanjutnya, untuk mengetahui karakteristik dari setiap kluster, dilakukan analisis lanjutan untuk mendapatkan cluster centroid yaitu nilai rata-rata dari setiap variabel pada setiap kluster ($\overline{\boldsymbol{X}}_{(x)k}$). Berikutnya, kita bandingkan nilai rata-rata ini dengan nilai rata-rata populasi. ($\boldsymbol{\mu}_{(x)}$) Bila $\overline{\boldsymbol{X}}_{(x)k} \leq \boldsymbol{\mu}_{(x)}$, maka dikategorikan sebagai "low", sedangkan bila $\overline{\boldsymbol{X}}_{(x)k} \geq \boldsymbol{\mu}_{(x)}$, dikategorikan "high" dan dikategorikan "very high" bila $\overline{\boldsymbol{X}}_{(x)k} \geq \boldsymbol{\mu}_{(x)}$ dibandingkan dengan rata-rata. hasil dapat dilihat pada Tabel 7 dan 8.

Table 7. Cluster Centroid

Cluster	Confirmed Cases	Death Cases	Recovered Cases	MIR
1	28708.8	787.9	21238.2	2.58
2	20519.3	453.7	15341.4	2.24
3	88434.7	2004.8	67230.7	2.24
Outliers	371262.6	14269.5	341566.9	3.54
Data $(\boldsymbol{\mu}_{(x)})$	26391.3	648.9	20053.9	2.40

Table 8. Cluster Characteristics

Cluster	Confirmed Cases	Death Cases	Recovered Cases	MIR
1	High	High	High	High
2	Low	Low	Low	Low
3	High	High	High	Low

Outliers	Very High	Very High	Very High	Very High

Dari hasil klasterisasi data Covid-19 di Indonesia antara Maret 2020 hingga Juli 2021, terlihat penyebaran pandemi virus corona 2 (SARS-Cov-2) bervariasi antar provinsi. Dari 34 provinsi di Indonesia, empat provinsi berdasarkan metode Mahalanobis squared distance terdeteksi sebagai outlier yang berarti kasus terkonfirmasi, meninggal, sembuh dan MIR di keempat provinsi tersebut jauh lebih tinggi dibandingkan provinsi lainnya. Untuk mengatasi outlier tersebut diterapkan metode trimmed k-means clustering. Dari hasil tersebut terlihat bahwa metode trimmed k-means mampu memisahkan outlier pada data dan memberikan cluster yang optimal. Hasil ini konsisten dengan penelitian sebelumnya yang telah melaporkan metode pengelompokan k-means yang dipangkas kuat untuk outlier [16-17,24-25, 31]

Terkait data pandemi Covid-19 di Indonesia, metode trimmed k-means clustering membentuk 3 cluster dan memisahkan outlier dari 3 cluster tersebut sehingga membentuk cluster tersendiri. Klaster 1 terdiri dari 14 provinsi dan Klaster 2 dan 3 masing-masing terdiri dari 10 dan 6 provinsi. Di satu sisi, provinsi outlier yaitu DKI Jakarta, Jawa Barat, Jawa Tengah, dan Jawa Timur membentuk klaster tersendiri di luar 3 klaster di atas. Hal ini membuktikan bahwa metode trimmed k-means clustering bekerja sangat baik dalam mengelompokkan data yang mengandung outlier dan dapat digunakan pada jenis data yang sejenis.

V. KESIMPULAN

Ditemukan tiga klaster optimal pandemi Covid-19 di Indonesia antara Maret 2020 hingga Juli 2021, berdasarkan indeks separasi maksimum menggunakan metode trimmed k-means clustering. Klaster 1 terdiri dari 14 provinsi dan Klaster 2 dan 3 masing-masing terdiri dari 10 dan 6 provinsi. Sedangkan empat provinsi yaitu DKI Jakarta, Jawa Barat, Jawa Tengah, dan Jawa Timur merupakan outlier dan membentuk klaster tersendiri di luar 3 klaster di atas. Hal ini menunjukkan bahwa keempat provinsi tersebut memiliki kasus terkonfirmasi, kematian sembuh, rasio kematian terhadap insiden (MIR) yang lebih tinggi dibandingkan provinsi lain di Indonesia.

DAFTAR PUSTAKA

1. Livana, Ph.; Suwoso, R.H.; Febrianto, T.; Kushindarto, D.; Aziz, F. Dampak Pandemi Covid-19 Bagi Perekonomian Masyarakat Desa. *Indones. J. Nurs. Health Sci.* **2020**, *1*, 37–48.
2. Fahrika, A.I.; Roy, J. Dampak pandemi covid 19 terhadap perkembangan makro ekonomi di indonesia dan respon kebijakan yang ditempuh. *INOVASI* **2020**, *16*, 206–213, doi:10.29264/jinv.v16i2.8255.
3. Rohmah, S.N. Adakah Peluang Bisnis Di Tengah Kelesuan Perekonomian Akibat Pandemi Corona? *ADALAH* **2020**, *4*, 63–74.
4. Muhyiddin, M. Covid-19, New Normal, Dan Perencanaan Pembangunan Di Indonesia. *J. Perenc. Pembang. Indones. J. Dev. Plan.* **2020**, *4*, 240–252, doi:10.36574/jpp.v4i2.118.
5. Saidi, S.; Herawati, N.; Nisa, K. Modeling with Generalized Linear Model on Covid-19: Cases in Indonesia. *Int. J. Electron. Commun. Syst.* **2021**, *1*, 25–32.
6. Setiawan, S.S., Netti Herawati, Khoirin Nisa, Eri Nonparametric Modeling Using Kernel Method for the Estimation of the Covid-19 Data in Indonesia During 2020. *Int. J. Math. Trends Technol. IJMTT*.
7. Abdullah, D.; Susilo, S.; Ahmar, A.S.; Rusli, R.; Hidayat, R. The Application of K-Means Clustering for Province Clustering in Indonesia of the Risk of the COVID-19 Pandemic Based on COVID-19 Data. *Qual. Quant.* **2021**, 1–9, doi:10.1007/s11135-021-01176-w.
8. Indraputra, R.A.; Fitriana, R. K-Means Clustering Data COVID-19. *J. Tek. Ind.* **2020**, *10*, 275–282, doi:10.25105/jti.v10i3.8428.
9. Virgantari, F.; Faridhan, Y.E. K-Means Clustering of COVID-19 Cases in Indonesia's Provinces. **2020**, 7.
10. Vahabi, N.; Salehi, M.; Duarte, J.D.; Mollalo, A.; Michailidis, G. County-Level Longitudinal Clustering of COVID-19 Mortality to Incidence Ratio in the United States. *Sci. Rep.* **2021**, *11*, 3088, doi:10.1038/s41598-021-82384-0.
11. Rojas, F.; Valenzuela, O.; Rojas, I. Estimation of COVID-19 Dynamics in the Different States of the United States Using Time-Series Clustering. *medRxiv* **2020**, 2020.06.29.20142364, doi:10.1101/2020.06.29.20142364.
12. Maugeri, A.; Barchitta, M.; Basile, G.; Agodi, A. Applying a Hierarchical Clustering on Principal Components Approach to Identify Different Patterns of the SARS-CoV-2 Epidemic across Italian Regions. *Sci. Rep.* **2021**, *11*, 7082, doi:10.1038/s41598-021-86703-3.
13. Kumar, S. Use of Cluster Analysis to Monitor Novel Coronavirus-19 Infections in Maharashtra, India. *Indian J. Med. Sci.* **2020**, *72*, 44–48, doi:10.25259/IJMS_68_2020.
14. Choi, Y.-J.; Park, M.-J.; Park, S.J.; Hong, D.; Lee, S.; Lee, K.-S.; Moon, S.; Cho, J.; Jang, Y.; Lee, D.; et al. Types of COVID-19 Clusters and Their Relationship with Social Distancing in the Seoul Metropolitan Area, South Korea. *Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis.* **2021**, *106*, 363–369, doi:10.1016/j.ijid.2021.02.058.
15. Gallegos, M.T.; Ritter, G. A Robust Method for Cluster Analysis. *Ann. Stat.* **2005**, *33*, 347–380, doi:10.1214/009053604000000940.
16. Cuesta-Albertos, J.A.; Gordaliza, A.; Matran, C. Trimmed K-Means: An Attempt to Robustify Quantizers. *Ann. Stat.* **1997**, *25*, 553–576.
17. Garcia-Escudero, L.A.; Gordaliza, A. Robustness Properties of k Means and Trimmed k Means. *J. Am. Stat. Assoc.* **1999**, *94*, 956–969, doi:10.2307/2670010.

18. Larasati, S.D.A.; Nisa, K.; Herawati, N. Robust Principal Component Trimmed Clustering of Indonesian Provinces Based on Human Development Index Indicators. *J. Phys. Conf. Ser.* **2021**, *1751*, 012021, doi:10.1088/1742-6596/1751/1/012021.
19. Meng, S.; Fu, Y.; Liu, T.; Li, Y. Principal Component Analysis for Clustering Temporomandibular Joint Data. In Proceedings of the 2015 8th International Symposium on Computational Intelligence and Design (ISCID); December 2015; Vol. 1, pp. 422–425.
20. Rahman, A.S.; Rahman, A. Application of Principal Component Analysis and Cluster Analysis in Regional Flood Frequency Analysis: A Case Study in New South Wales, Australia. *Water* **2020**, *12*, 781, doi:10.3390/w12030781.
21. Untoro, M.C.; Anggraini, L.; Andini, M.; Retnosari, H.; Nasrulloh, M.A. Penerapan metode k-means clustering data COVID-19 di Provinsi Jakarta. *Teknol. J. Ilm. Sist. Inf.* **2021**, *11*, 59–68, doi:10.26594/teknologi.v11i2.2323.
22. Utomo, W. The Comparison of K-Means and k-Medoids Algorithms for Clustering the Spread of the Covid-19 Outbreak in Indonesia. *Ilk. J. Ilm.* **2021**, *13*, 31–35, doi:10.33096/ilkom.v13i1.763.31-35.
23. Hutagalung, J.; Ginantra, N.L.W.S.R.; Bhawika, G.W.; Parwita, W.G.S.; Wanto, A.; Panjaitan, P.D. COVID-19 Cases and Deaths in Southeast Asia Clustering Using K-Means Algorithm. *J. Phys. Conf. Ser.* **2021**, *1783*, 012027, doi:10.1088/1742-6596/1783/1/012027.
24. García-Escudero, L.; Gordaliza, A.; Matrán, C.; Mayo, A. A General Trimming Approach to Robust Cluster Analysis. *Ann. Stat.* **2008**, *36*, 1324–1345, doi:10.1214/07-AOS515.
25. García-Escudero, L.; Gordaliza, A.; Matrán, C.; Mayo, A. A Review of Robust Clustering Methods. *Adv. Data Anal. Classif.* **2010**, *4*, 89–109, doi:10.1007/s11634-010-0064-5.
26. Nisa, K.; Herawati, N.; Setiawan, E.; Nusyirwan Robust Principal Component Analysis Using Minimum Covariance Determinant Estimator.; November 30 2006.
27. Rousseeuw, P.J.; Driessen, K.V. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* **1999**, *41*, 212–223, doi:10.1080/00401706.1999.10485670.
28. Hubert, M.; Debruyne, M.; Rousseeuw, P.J. Minimum Covariance Determinant and Extensions. *WIREs Comput. Stat.* **2018**, *10*, e1421, doi:10.1002/wics.1421.
29. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65, doi:10.1016/0377-0427(87)90125-7.
30. Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of Internal Clustering Validation Measures. *2010 IEEE Int. Conf. Data Min.* **2010**, doi:10.1109/ICDM.2010.35.
31. Lam, B.S.Y.; Choy, S.K. A Trimmed Clustering-Based L1-Principal Component Analysis Model for Image Classification and Clustering Problems with Outliers. *Appl. Sci.* **2019**, *9*, 1562, doi:10.3390/app9081562.

LAMPIRAN

LAMPIRAN

Lampiran 1. Aricle ULICOSTE 2021

Implementation of the Trimmed *k*-means clustering method in mapping the distribution of Covid-19 in Indonesia

Netti Herawati[1, a)] ,Khorin Nisa[2, b)], and Subian Saidi[3, c)]

[1, 2, 3)]Department of Mathematics, University of Lampung, Bandar Lampung, Indonesia

Author Emails

[a)] Corresponding author: netti.herawati@fmipa.unila.ac.id

[b)] khoirin.nisa@fmipa.unila.ac.id

[c)] subian.saidi@fmipa.unila.ac.id

Abstract. The Coronavirus that appeared in (COVID-19), caused by SARSCoV-2, started at Wuhan in the Hubei province of China and has spread with great speed around the world; it has caused a severe health crisis all around the world, including Indonesia. This study aims to use a clustering technique to assess the risk of the COVID-19 pandemic in Indonesia, based on data obtained between March 2020 and July 2021 in that country (http://www.covid19.go.id). Provinces in Indonesia were grouped based on COVID-19 infection rates and mortality data. Since the data contained some outliers, i.e. provinces with a very high number of cases, we used a robust clustering method; this method is sensitive to outliers. The analysis was performed using the Trimmed *k*-means clustering method. Based on the results of this study, with four provinces detected as outliers in the data, there were three optimal clusters with the maximum separation index. Cluster 1 consisted of 14 provinces, and clusters 2 and 3 consisted of 10 and 6 provinces, respectively. The four outliers, i.e. Jakarta, West Java, Central Java and East Java, formed a separate cluster..

Introduction

The global Covid-19 in Indonesia is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2). As of 23 July 2021, Indonesia has reported 3,082,410 cases, the highest in Southeast Asia, ahead of the Philippines. With 80,598 deaths, Indonesia ranks third in Asia and 15th in the world.

The corona virus pandemic has had a huge impact on Indonesia. Not only has it impacted the health sector, but it also hit other sectors such as the economy, industry, education, social, and tourism [1–3]. There is not a single province in Indonesia that was not affected by the coronavirus on a small or large scale. Various policies have been carried out by the government to anticipate and minimize the impact of the pandemic. These include setting up a task force team to accelerate the handling of COVID-19, enforcement a Large-Scale Social Restrictions (LSSR) in various regions, purchasing medical equipment, upgrading a number of referral hospitals for Covid-19 patients, distributing social assistance in the form of basic necessities and cash for people in need, exempting from electricity costs for certain groups, and providing tax incentives for the industrial sector.

Indonesia is a country with a very large population. In fact, Indonesia ranks fourth in the world as the country with the largest population after China, India and the United States. As an archipelagic country, Indonesia's population is spread across various provinces in Indonesia. The population in each province is different and continues to change. Large population growth and uneven distribution are a source of problems in Indonesia. This problem also applies when it comes to handling Covid-19 [4]. The spread of Covid-19 in Indonesia was rapid and varied from region to region. Due to the large area and large population in Indonesia, it is necessary to anticipate the transmission of the Covid-19 virus quickly, one of the efforts undertaken was to study the data of the Covid-19 cases in 34 provinces in Indonesia separately.

Various studies on Covid-19 cases in Indonesia have been carried out by numerous authors to provide good statistical analysis for the government as a reference in policy making. Some of those studies were modelling the data using a generalized linear model approach [5], forecasting using nonparametric analysis [6], and classification using cluster analysis [7–9]. Cluster analysis seems to be the most popular technique for analyzing the Covid-19 data. The technique is also predominately used for Covid-19 data analysis in worldwide countries; one can see [10–14]

Cluster analysis is divided into two types, namely hierarchical clustering and non-hierarchical clustering. The difference between the two types lies in determining the number of clusters to be produced. The hierarchical clustering is used when the desired number of clusters is unknown, while the non-hierarchical clustering is used when the desired number of clusters has been predetermined. When data contains outliers, a robust method is required in order to avoid a biased analysis result due to the influence of the outliers, see [15] for details. One of the robust methods for cluster analysis is the trimmed *k*-means method. The advantage of the trimmed *k*-means method is its resistance to outliers [16,17]. Moreover, cluster analysis is slightly different from other multivariate methods. This method does not assume a certain distribution of data as generally required by other multivariate methods. Cluster analysis is based on a distance matrix which represents a measure of similarity, and the most commonly used measure is Euclidean distance. There is, however, an assumption that must be fulfilled when using Euclidean distance for cluster analysis, namely that all variables are uncorrelated, and this assumption is often ignored, causing the clustering results to be not optimal. When the variables in the data are correlated with each other, principal component analysis (PCA) can be performed. The cluster analysis is then carried out based on the principal component scores of the observations. For more information on clustering analysis through PCA, the reader is urged to consult[18–20].

In this study, we conducted a cluster analysis of the Covid-19 data based on the number of confirmed cases, death cases, recovered cases and mortality to incidence ratio (MIR) from 34 provinces in Indonesia using the robust trimmed k-means method and principal component scores to obtain a cluster map that was more optimal and efficient than the k-means clustering. The results of the analysis could be used as an overview of the severity and level of risk of the spread of Covid-19 in the provinces in Indonesia using the clusters obtained

Materials and Methods

The data used in this study are Covid-19 data in Indonesian, obtained between 20 March 2020 and 23 July 2021 and provided by the Covid-19 Acceleration Task Force (Indonesian: SATGAS) (https://covid19.go.id/). There were three variables obtained from the web, i.e. numbers of Covid-19 confirmed recovered, and death cases, respectively. The MIR of Covid-19 was included to show the measure of severity of the disease. The MIR as a proxy for survival rate was calculated by dividing the number of deaths cases in each province by the confirmed cases in the same province then multiplied by 100. The MIR ranges from 0 to 100%, 100% indicating the worst situation where all confirmed cases have died [10].

We present the summary of Covid-19 data from 34 provinces of Indonesia in Table 1. There are three variables in the data that have significantly different mean and median values, i.e. confirmed, death and recovered cases; this shows that the three variables have an asymmetric or skewed distribution. The asymmetry of the data can be caused by the natural structure of the variable or it can also be caused by the existence of outliers. The MIR of the 34 provinces in in Indonesia on 23 July 2021 was 2.531%, this value decreased compared to the previous year but it is known to be the highest among ASEAN countries.

Table 1. Summary of COVID-19 cases of 34 Province in Indonesia

Descriptive Statistics	Confirmed Cases	Death Cases	Recovered Cases	MIR
Minimum	7103	149	5930	0.897
1st Quartile	17790	314	13393	1.710
Median	31069	775	22580	2.250
Mean	90655	2363	71084	2.531
3rd Quartile	73331	1706	57341	2.766
Maximum	778521	17512	678992	6.568
Standard deviation	164523.3	4292.167	136554.595	1.180

Clustering is a broad set of techniques for finding subgroups of observations within a data set. When we cluster observations, we want observations in the same group to be similar and observations in different groups to be dissimilar. Clustering allows us to identify which observations are alike, and potentially categorize them therein. The k-means algorithm is the simplest and the most commonly used clustering method for splitting a dataset into a set of *k* groups, including for

the clustering analysis of Covid-19 data [7,11,21–23]. The algorithm aims to minimize the collective squared Euclidean distances between the observation and the centroid of cluster to which they belong, however, the *k*-means algorithm does not give the best results; it is sensitive to outliers (i.e. a point which is different from the rest of data points). The outliers in the data will cause the results of cluster analysis to be inefficient. In this cases, a more robust method is needed and is described below.

Trimmed *k*-means was introduced by [16]. Using this method, one is allowed to remove a certain proportion of the possible outliers when grouping the results [24,25]; the proportion to be removed is the magnitude of the data rate to be trimmed in this method. The trimmed *k*-means method is used to overcome the outliers contained in a cluster of data to be grouped. The main concept of the trimmed *k*-means method is to form new *k* clusters by removing, or trimming the outliers contained in the data. The method belongs to a class of procedures based on "impartial trimming" (determined by the data) with the aim of making the classical hierarchical clustering technique namely *k*-means, more robust. Furthermore, the generalized *k* means method is based on the minimization of the discrepancy between a random variable (or a sample of this random variable) and a set with *k* points measured, using a penalty function Φ [16].

Let α be contained in an open unit interval, i.e. $\alpha \in (0,1)$, k a natural number, and Φ a penalty function. For every set A such that $P(A) \geq 1-\alpha$ and every k-set $M = \{ m_1, m_2, \ldots, m_k \}$ in R^p, let us consider the variation about *M* given *A:*

$$V_{\Phi}^{A}(M) = \frac{1}{P(A)} \int_A \Phi\left(\inf_{i=1,\ldots,k} \|X - m_i\| \right) dP \tag{1}$$

The variation $V_{\Phi}^{A}(M)$ measures how well the set M represents the probability mass of P defined on A, and our job is to choose a set of given probability mass such that the variation is minimized. This is done by minimizing $V_{\Phi}^{A}(M)$ on A and M in the following way:

1. obtain the *k*-variation given A, $V_{k\Phi}^{A}(M)$, by minimizing over *M*:

$$V_{k,\Phi}^{A}(M) = \inf_{M \subset R^p, \#M \neq k} V_{\Phi}^{A}(M) \tag{2}$$

2. obtain the trimmed *k*-variation, $V_{k,\Phi,\alpha}$ by minimizing over *A*:

$$V_{k,\Phi,\alpha} := V_{k,\Phi,\alpha}(X) := V_{k,\Phi,\alpha}(P_X) := \inf_{A \in \beta^p, P(A) \geq 1-\alpha} \quad V_{k,\Phi}^{A} \tag{3}$$

We wish to obtain a trimmed set A_0 if it exists, and a k-set $M_0 = \{m_1^0, m_2^0, \ldots, m_k^0\}$ if it exists, using the condition $V_{\Phi}^{A_0}(M_0) = V_{k,\Phi,\alpha}$ [16]

Principal Component Analysis (PCA) is a statistical multivariate technique that aims to reduce the dimensions of the data to obtain new uncorrelated variables and retain most of the information contained in the original variables. The resulting variable is a linear combination of the original variables and is called the principal component (PC). The sum of the squares of the coefficients in the linear combination is equal to unity, and the PCs are orthogonal.

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ be a random vector of a multivariate Gaussian distribution with mean vector $\mu = (\mu, \mu, \ldots, \mu)$ and covariance matrix $\boldsymbol{\Sigma}$ and *k* independent eigenvectors, a_k, $k = 1, 2, \ldots, k$. The principal component then can be written as follows:

$$Y = A(X - \mu) \tag{4}$$

where **A** is a *p* by *p* coefficient matrix that carries the *p*-element variable $\boldsymbol{X}$ into the *n*-element variable y. The column vectors in **A** are eigenvectors **Σ**, i.e. $\boldsymbol{A} = [\boldsymbol{a_1} \mid \boldsymbol{a_2} \mid \ldots \mid \boldsymbol{a_k}]$. The i-th principal component can be written as:

$$y_i = a_{1i} X_1 + a_{2i} X_2 + \cdots + a_{ki} X_k = a_i^T X \tag{5}$$

A problem arises in PCA when the data contain outliers, since the covariance matrix is very sensitive to the presence of outliers. To overcome this, a robust principal component analysis is needed, namely by replacing the classical sample covariance matrix **S** with a robust estimator. The use of a robust covariance matrix estimate when calculating the principal components is the same as performing a robust PCA [26]. One of the robust estimators for the covariance matrix is the

Minimum Covariance Determinant (MCD). The MCD estimator is a pair $(\overline{X}_{MCD}, S_{MCD})$, where $\overline{X}_{MCD}$ is the mean vector and S_{MCD} is the covariance matrix that minimizes the determinant of the sample covariance matrix **S** in the subsample containing exactly *h* members of *n* observations [27,28].

It is often necessary to validate the number of clusters obtained from the partitioning by using the cluster validity index. The validity index is a method for evaluating the results of the clustering algorithm in order to get the best number of clusters [29]. The validity index is calculated based on the following two criteria:

1. Compactness is the level of similarity of objects in the same cluster, and
2. Separation is the level of difference between objects in different clusters.

The silhouette analysis measures how well an observation is clustered, and it estimates the average distance between clusters. It contains the average value of each point in the data set, more specifically, the calculation of the value of each point is the difference between the values of separation and compactness, divided by the maximum between the two. The best number of clusters is indicated by the silhouette value which is as close to 1 as possible. Suppose there are *N* points in a data set. There exist clusters, *p* and *q*, whereby x_i is a point in cluster *p* and y_j is a point in cluster *q*, such that $a_{p,i}$ is the average distance of point x_i to each point in cluster *p*, and $d_{q,i}$ is the average distance of point x_i to every point in cluster *q*. For each observation, *i*, the silhouette width s_{x_i} is calculated as follows:

1. Calculate the average dissimilarity $a_{p,i}$ between *i* and all other points of the cluster to which i belongs, as follows:

$$a_{p,i} = \frac{1}{n_p} \sum_{k=1}^{n_p} d(x_i, x_k) \tag{6}$$

2. For all other clusters *q*, to which i does not belong, calculate the average dissimilarity $d_{q,i}$ of *i* to all observations of *q*, i.e. $d_{q,i} = \frac{1}{n_q} \sum_{j=1}^{n_q} d(x_i, x_j)$. The smallest of these $d_{q,i}$ is defined as $b_{q,i} = \min d_{q,i}, q = 1, \ldots, k$. The value of $b_{q,i}$ can be understood as the dissimilarity between *i* and its "neighbor" cluster, i.e., the nearest one to which it does not belong.
3. Finally, the silhouette width of the observation *i*, is defined by the formula:

$$s_{x_i} = \frac{\left(b_{q,i} - a_{p,i}\right)}{\max\left\{b_{q,i} - a_{p,i}\right\}}, p \neq q \tag{7}$$

The silhouette width can be interpreted as follows:

$s_{x_i} > 0$, means that the observation is well grouped

$s_{x_i} < 0$, means that the observation has been placed in the wrong cluster.

$s_{x_i} = 0$, means that the observation is between two clusters.

The average silhouette width is then given by:

$$SIL = \frac{1}{n} \sum_{i=0}^{N} s_{x_i} \tag{8}$$

Observations with a large s_{x_i}, almost 1, are well clustered. For a maximum value of the average silhouette, SIL, the optimal clustering is obtained [30].

Results

We performed a robust cluster analysis using the trimmed *k*-means method to the data by firstly addressing correlated variables and outliers using robust principal component analysis. Then, we used the robust principal component score for cluster analysis using the trimmed *k*-means method.

Outliers Detection

In this study, the method used to detect outliers is the Mahalanobis squared distance method where the *i*-th observation is identified as an outlier if $d^2_{MD}(i) > \chi^2_{p,1-a}$. Based on this case, there are 4 variables studied, and we used α = 5%. Therefore, we have $\chi^2_{p,} = 4,(1-0{,}05) = 9{,}488$. Outliers are defined by comparing the results of the Mahalanobis squared distance of each object with the value of $\chi^2_{4,0,95}$; as a result, 4 outlier provinces are found, namely DKI Jakarta, West Java, Central Java, and East Java. The quantile-quantile (Q-Q) plot of the chi-square of the data is presented in Figure 1. The result of the outlier detection is also confirmed by the panel plot of each pair between variables as shown in Figure 2.

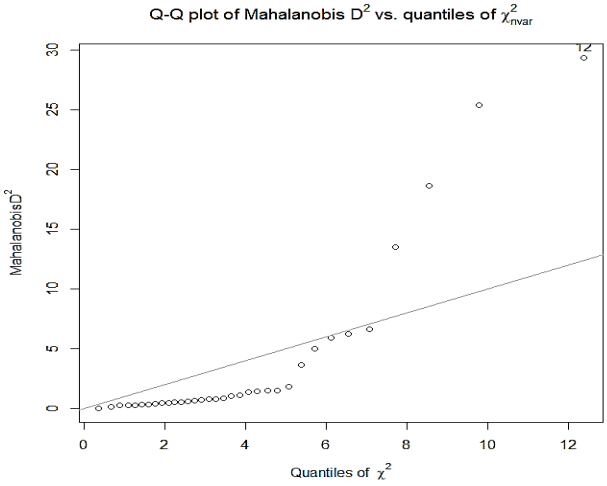


Figure 1. Q-Q plot of squared Mahalanobis distance vs. chi-squared quantiles

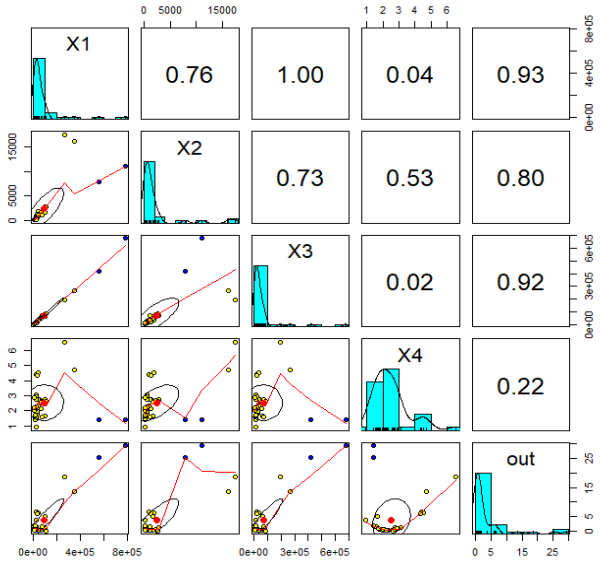


Figure 2. Pair panels plot of Covid-19 Data in Indonesia

The four provinces identified as outliers with their attribute presented in Table 2. The data in the table show that DKI Jakarta had the highest confirmed cases, while East Java had the lowest confirmed cases. However, the MIR of East Java was the highest among the four provinces.

Table 2. List of outliers in the Covid-19 Data in Indonesia

Province	Confirmed Cases	Death Cases	Recovered Cases	MIR

DKI Jakarta	778521	11131	678992	1.430
West Java	556181	7917	421977	1.423
Central Java	343210	16195	267511	4.719
East Java	266638	17512	194233	6.568

The correlation coefficients between variables in the data are shown in Table 3. It can be seen that there were correlations between variables, with the maximum value is the correlation between the confirmed cases and the recovered cases, i.e. 0.997. Some other correlation coefficients were moderate, with the values between 0.53 and 0.76.

Table 3. Correlation matrix between variables

	Confirmed Cases	**Death Cases**	**Recovered Cases**	**MIR**
Confirmed Cases	1.000	0.758	0.997	0.036
Death Cases	0.758	1.000	0.732	0.533
Recovered Cases	0.997	0.732	1.000	0.015
MIR	0.015	0.533	0.015	1.000

Robust Cluster Analysis using Robust Principal Scores

A robust PCA using MCD estimator was performed to obtain new uncorrelated variables from the data containing outliers. The coefficients of the linear combinations of each PC are shows in Table 4

Table 4. Robust Principal Component Coefficients

Variables	**PC1**	**PC2**	**PC3**	**PC4**
Confirmed Cases	-0.561	-0.253	-0.099	0.781
Death Casess	-0.571	0.153	-0.671	-0.446
Recovered Cases	-0.558	-0.257	0.681	-0.398
MIR	-0.215	0.920	0.274	0.179

The trimmed k-means clustering was done using all principal components to retain all the information in the data. We conducted α-trimmed k-means clustering with α=0.1 and the number of clusters k = 2,3,4,5,6,7, and 8. Using α=0.1 means that 10% of the trimmed data are observations that were not part of the cluster that was formed. The optimal clustering was then selected based on the validation indices resulting from each of the k clusters. The comparison of the average silhouette indexes of all k can be seen through their plots in Figure 3.

In Figure 3, it can be seen that the average silhouette index reached the maximum value at k=3 with the value equals 0.54; this means that the optimal cluster number for the data was k=3.

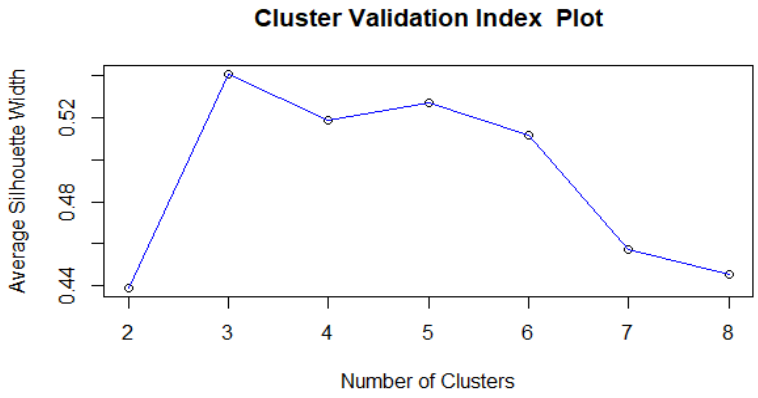


Figure 3. Plot of average silhouette index for *k*=2,…, 8

The three clusters resulting from the trimmed k-means clustering in Figure 4 are Cluster 1, Cluster 2 and Cluster 3 (the other cluster in the figure, i.e. Cluster 0, is containing the four provinces detected as outliers; they will be analyzed further). Since the silhouette width of the observations in each cluster are positif ($s_{x_i} > 0$), then all observations in the three clusters are well grouped. Based on the silhouette index, the optimal number of clusters for the α-trimmed clustering (i.e. the data excluding 4 outliers) is *k*=3. Nevertheless, this result is in accordance with the separation index as shown in Figure 4. The result of the trimmed *k*-means clustering is shown in Figure 5.

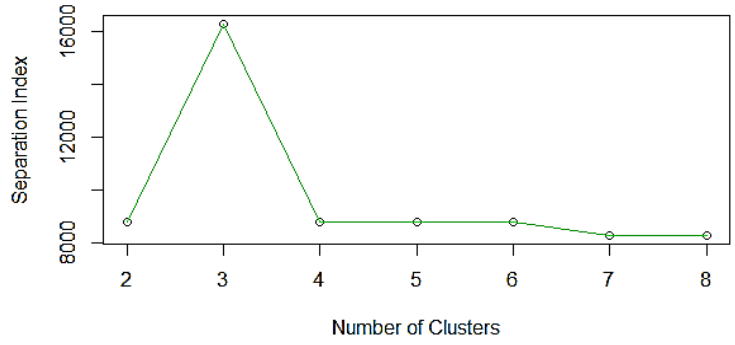


Figure 4. Plot of separation index for k=2, …, 8

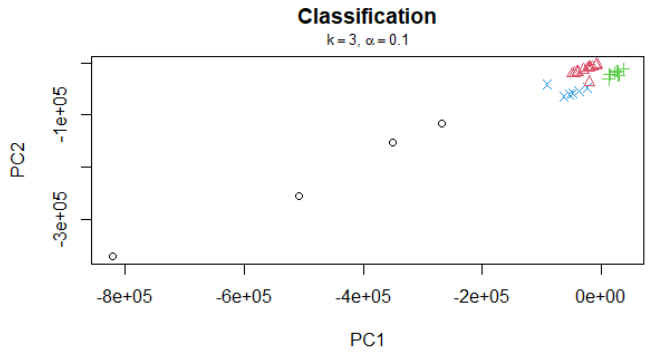


Figure 5. Trimmed *k*-means classification plot

Combining the outlying clusters with the three clusters of trimmed *k*-means clustering, we have 4 clusters, the member of each clusters are presented in Table 5 and grapically in Figure 6.

Table 5. Members of Cluster

Cluster	Cluster size	Province
1	14	North Sumatera, South Sumatera, Lampung, Riau Island, Bali, West Nusa Tenggara, West Kalimantan, South Kalimantan, North Kalimantan, Central Sulawesi, South East Sulawesi, West Sulawesi, Maluku Utara, Papua
2	10	Aceh, Jambi, Bengkulu, Bangka Belitung Island, East Nusa Tenggara, Central Kalimantan, North Sulawesi, Gorontalo, Maluku, West Papua
3	6	West Sumatera, Riau, Banten, DI Yogyakarta, East Kalimantan, South Sulawesi
outliers	4	DKI Jakarta, West Java, Central Java, and East Java

Figure 6. Covid-19 cluster result map in Indonesia between March 2020 and July 2021

To find out the characteristics of each cluster, it was necessary to do further analysis on each cluster by calculating the cluster centroid i.e. the mean value of each variable of the observations in the cluster ($\bar{\boldsymbol{X}}_{(x)k}$). Furthermore, these values are compared to the robust centroid of the entire data ($\boldsymbol{\mu}_{(x)}$) as shown in Table 6. If $\bar{\boldsymbol{X}}_{(x)k} \leq \boldsymbol{\mu}_{(x)}$, where x is the clustering variables, then it could be interpreted that the mean value of the variables in the cluster was "low", on the one hand. On the other hand, if $\bar{\boldsymbol{X}}_{(x)k} \geq \boldsymbol{\mu}_{(x)}$, then the mean value of the variables in the cluster could be interpreted as "high" and "very high" if $\bar{\boldsymbol{X}}_{(x)k} \geq \boldsymbol{\mu}_{(x)}$ compare to the other 3 clusters. The result is shown in Table 7.

Table 6. Cluster Centroid

Cluster	Confirmed Cases	Death Cases	Recovered Cases	MIR
1	28708.8	787.9	21238.2	2.58
2	20519.3	453.7	15341.4	2.24
3	88434.7	2004.8	67230.7	2.24
Outliers	371262.6	14269.5	341566.9	3.54
Data $(\boldsymbol{\mu}_{(x)})$	26391.3	648.9	20053.9	2.40

Table 7. Cluster Characteristics

Cluster	Confirmed Cases	Death Cases	Recovered Cases	MIR
1	High	High	High	High
2	Low	Low	Low	Low
3	High	High	High	Low

Outliers	Very High	Very High	Very High	Very High

Discussion

From the results of clustering the data Covid-19 in Indonesia between March 2020 and July 2021, it can be seen that the spread of the pandemic caused by coronavirus 2 (SARS-Cov-2) varies between provinces. Of the 34 provinces in Indonesia, four provinces based on Mahalanobis squared distance method were detected as outliers which means the confirmed, death, recovered cases and MIR are much higher in these four provinces than other provinces. To overcome the outliers, the trimmed *k*-means clustering method was applied. From the results, it can be seen that the trimmed k-means method is able to separate the outliers in the data and provide an optimal clusters. These results are consistent with preceded studies that have reported the trimmed *k*-means clustering method was robust to outliers [16-17,24-25, 31]

Regarding the Covid-19 pandemic data in Indonesia, the trimmed k-means clustering method forms 3 clusters and separates the outliers from the 3 clusters to form a separate cluster. Cluster 1 consisted of 14 provinces and Cluster 2 and 3 consisted of 10 and 6 provinces, respectively. On the one hand, the outlier provinces, namely DKI Jakarta, West Java, Central Java, and East Java formed a separate cluster outside the 3 clusters above. This proves that the trimmed k-means clustering method works very well in grouping data containing outliers and can be used on similar types of data.

CONCLUSION

Three optimal clusters of the Covid-19 pandemic in Indonesia between March 2020 and July 2021 were found, based on the maximum separation index using the trimmed *k*-means clustering method. Cluster 1 consisted of 14 provinces and Cluster 2 and 3 consisted of 10 and 6 provinces, respectively. Meanwhile, four provinces, namely DKI Jakarta, West Java, Central Java, and East Java, were outliers and formed a separate cluster outside the 3 clusters above. This shows that the four provinces had higher confirmed cases, recovered deaths, mortality to incidence ratio (MIR) than the other provinces in Indonesia.

References

1. P.H. Livana, R. H. Suwoso, T. Febrianto, D. Kushindarto, and F. Aziz. Dampak Pandemi Covid-19 Bagi Perekonomian Masyarakat Desa. *Indones. J. Nurs. Health Sci*, **1**, 37–48 (2020).
2. A. I. Fahrika, and J. Roy. Dampak pandemi covid 19 terhadap perkembangan makro ekonomi di indonesia dan respon kebijakan yang ditempuh. *INOVASI*, **16**, 206–213 (2020). doi:10.29264/jinv.v16i2.8255.
3. S. N. Rohmah. Adakah Peluang Bisnis Di Tengah Kelesuan Perekonomian Akibat Pandemi Corona? *ADALAH*, **4**, 63–74 (2020).
4. M. Muhyiddin. Covid-19, New Normal, Dan Perencanaan Pembangunan Di Indonesia. *J. Perenc. Pembang. Indones. J. Dev. Plan*, **4**, 240–252 (2020). doi:10.36574/jpp.v4i2.118.
5. S. Saidi, N. Herawati, K. Nisa, Modeling with Generalized Linear Model on Covid-19: Cases in Indonesia. *Int. J. Electron. Commun. Syst*, **1**, 25–32 (2021).
6. S. Saidi, N. Herawati, K. Nisa, and E. Setiawan. Nonparametric Modeling Using Kernel Method for the Estimation of the Covid-19 Data in Indonesia during 2020. *Int. J. Math. Trends Technol. IJMTT*, **67**, 136-144 (2021).
7. D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat. The Application of K-Means Clustering for Province Clustering in Indonesia of the Risk of the COVID-19 Pandemic Based on COVID-19 Data. *Qual. Quant*, 1–9 (2021). doi:10.1007/s11135-021-01176-w.
8. R. A. Indraputra, and R. Fitriana. K-Means Clustering Data COVID-19. *J. Tek. Ind*, **10**, 275–282 (2020). doi:10.25105/jti.v10i3.8428.
9. F. Virgantari, and Y. E. Faridhan. K-Means Clustering of COVID-19 Cases in Indonesia's Provinces, **7** (2020).

10. N. Vahabi, M. Salehi, J. D. Duarte, A. Mollalo, and G. Michailidis. County-Level Longitudinal Clustering of COVID-19 Mortality to Incidence Ratio in the United States. *Sci. Rep,* **11**, 3088 (2021). doi:10.1038/s41598-021-82384-0.
11. F. Rojas, O. Valenzuela, and I. Rojas. Estimation of COVID-19 Dynamics in the Different States of the United States Using Time-Series Clustering. *medRxiv* (2020). 2020.06.29.20142364, doi:10.1101/2020.06.29.20142364.
12. A. Maugeri, M. Barchitta, G. Basile, and A. Agodi. Applying a Hierarchical Clustering on Principal Components Approach to Identify Different Patterns of the SARS-CoV-2 Epidemic across Italian Regions. *Sci. Rep*, **11**, 7082 (2021). doi:10.1038/s41598-021-86703-3.
13. S. Umar. Use of Cluster Analysis to Monitor Novel Coronavirus-19 Infections in Maharashtra, India. *Indian J. Med. Sci*, **72**, 44–48 (2020**)**. doi:10.25259/IJMS_68_2020.
14. Y. J. Choi, M. J. Park, S. J. Park, D. Hong; S. Lee, K. S. Lee, S. Moon, J. Cho, Y. Jang, D. Lee, et al. Types of COVID-19 Clusters and Their Relationship with Social Distancing in the Seoul Metropolitan Area, South Korea. *Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis*, **106**, 363–369 (2021) doi:10.1016/j.ijid.2021.02.058.
15. M. T. Gallegos, and G. Ritter. A Robust Method for Cluster Analysis. *Ann. Stat.*, **33**, 347–380 (2005), doi:10.1214/009053604000000940.
16. J. A. Cuesta-Albertos, A. Gordaliza, and C. Matran. Trimmed K-Means: An Attempt to Robustify Quantizers. *Ann. Stat.* **25**, 553–576 (1997)**.**
17. L. A. Garcia-Escudero, A. Gordaliza. Robustness Properties of k Means and Trimmed k Means. *J. Am. Stat. Assoc*, **94**, 956–969 (1999). doi:10.2307/2670010.
18. S. D. A. Larasati, K. Nisa, and N. Herawati. Robust Principal Component Trimmed Clustering of Indonesian Provinces Based on Human Development Index Indicators. *J. Phys. Conf. Ser*, *1751*, 012021 (2021)**.** doi:10.1088/1742-6596/1751/1/012021.
19. S. Meng, Y. Fu, T. Liu, and Y. Li,. Principal Component Analysis for Clustering Temporomandibular Joint Data. In Proceedings of the 2015 8th International Symposium on Computational Intelligence and Design (ISCID), **1**, 422–425 (2015).
20. A. S. Rahman, and A. Rahman. Application of Principal Component Analysis and Cluster Analysis in Regional Flood Frequency Analysis: A Case Study in New South Wales, Australia. *Water*, **12**, 781 (2020). doi:10.3390/w12030781.
21. M. C. Untoro, L. Anggraini, M. Andini, H. Retnosari, and M. A. Nasrulloh. Penerapan metode k-means clustering data COVID-19 di Provinsi Jakarta. *Teknol. J. Ilm. Sist. Inf.,* **11**, 59–68 (2021). doi:10.26594/teknologi.v11i2.2323.
22. W. Utomo. The Comparison of K-Means and k-Medoids Algorithms for Clustering the Spread of the Covid-19 Outbreak in Indonesia. *Ilk. J. Ilm.*, **13**, 31–35 (2021). doi:10.33096/ilkom.v13i1.763.31-35.
23. J. Hutagalung, N. L. W. S. R. Ginantra, G. W. Bhawika, W. G. S.Parwita, A. Wanto. and P. D. Panjaitan. COVID-19 Cases and Deaths in Southeast Asia Clustering Using K-Means Algorithm. *J. Phys. Conf. Ser.*, **1783**, 012027 (2021). doi:10.1088/1742-6596/1783/1/012027.
24. L. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo. A General Trimming Approach to Robust Cluster Analysis. *Ann. Stat.*, **36**, 1324–1345 (2008)**.** doi:10.1214/07-AOS515.
25. L. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo. A Review of Robust Clustering Methods. *Adv. Data Anal. Classif.*, **4**, 89–109 (2010)**.** doi:10.1007/s11634-010-0064-5.
26. K. Nisa, N. Herawati, E. Setiawan, Nusyirwan. Robust Principal Component Analysis Using Minimum Covariance Determinant Estimator, International Conference on Mathematics and Natural Sciences (ICMNS), November 29-30, 2006, Bandung, Indonesia.
27. P. J. Rousseeuw, and K. V. Driessen. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* , **41**, 212–223 (1999**).** doi:10.1080/00401706.1999.10485670.
28. M. Hubert, M. Debruyne, and P. J. Rousseeuw,. Minimum Covariance Determinant and Extensions. *WIREs Comput. Stat.*, **10**, e1421, (2018)**.** doi:10.1002/wics.1421.
29. P. J. Rousseeuw, and Silhouettes. A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.*, **20**, 53–65 (1987**).** doi:10.1016/0377-0427(87)90125-7.
30. Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu,. Understanding of Internal Clustering Validation Measures. *2010 IEEE Int. Conf. Data Min.* (2010), doi:10.1109/ICDM.2010.35.

31. B. S. Y. Lam, and S. K. Choy. A Trimmed Clustering-Based L1-Principal Component Analysis Model for Image Classification and Clustering Problems with Outliers. *Appl. Sci.*, **9**, 1562 (2019)**.** doi:10.3390/app9081562.

Lampiran 2. Data COVID-19 di Indonesia Kurun Waktu 6 Bulan (Maret 2020-Agustus 2020)

Provinsi	Kasus Positif	Kasus Meninggal	Kasus Sembuh
Aceh	145	3	33
Sumatera Utara	1521	3	398
Sumatera Barat	753	13	158
Riau	1008	21	600
Jambi	465	7	35
Sumatera Selatan	841	27	405
Bengkulu	444	5	208
Lampung	445	11	105
Kep. Bangka Belitung	639	18	411
Kep. Riau	912	27	464
Banten	2626	25	420
DKI Jakarta	8033	157	10631
Jawa Barat	8925	156	6888
Jawa Tengah	4498	446	4094
D I Yogyakarta	1431	97	976
Jawa Timur	6912	349	4085
Bali	1407	0	0
Nusa Tenggara Barat	295	0	226
Nusa Tenggara Timur	807	12	880
Kalimantan Barat	422	14	0
Kalimantan Tengah	348	12	515
Kalimantan Selatan	876	8	112
Kalimantan Timur	1505	58	808
Kalimantan Utara	460	0	146
Sulawesi Utara	333	2	86
Sulawesi Tengah	497	5	135
Sulawesi Selatan	1286	19	480
Sulawesi Tenggara	274	5	141
Gorontalo	20	1	33
Sulawesi Barat	77	2	45
Maluku	121	7	155
Maluku Utara	181	7	330
Papua Barat	239	3	266
Papua	325	1	0
Total	49071	1521	34269

Lampiran 2. Script Analisis Klaster Menggunakan Metode *TRIMMED K-MEANS*

```
## DATA COVID INDONESIA ##

DATA=matrix(c(9541      ,383    ,7812   ,
        24521       ,833    ,20991      ,
        29152       ,634    ,26898      ,
        31397       ,755    ,2944   ,
        5487    ,74     ,4231   ,
        15913       ,746    ,13690      ,
        4933    ,144    ,4573   ,
        12534       ,644    ,10806      ,
        7347    ,110    ,6543   ,
        8667    ,229    ,8272   ,
        29417       ,518    ,21260      ,
        339201 ,5475 ,322080 ,
        211102 ,2327 ,170655 ,
        153018 ,6217 ,102587 ,
        27708       ,676    ,21761      ,
        129410 ,8456 ,108549 ,
        34402       ,544    ,13268      ,
        8728    ,350    ,5990   ,
        9247    ,257    ,6503   ,
        4637    ,31     ,4153   ,
        13848       ,309    ,12073      ,
        21835       ,723    ,19264      ,
        55397       ,1269 ,47767    ,
        9694    ,138    ,6800   ,
        14997       ,501    ,11892      ,
        10025       ,259    ,8227   ,
        56196       ,838    ,51550      ,
        10033       ,195    ,9271   ,
        4854    ,135    ,4359   ,
        5223    ,103    ,3810   ,
        6984    ,105    ,6141   ,
        3982    ,113    ,3380   ,
        7423    ,125    ,6808   ,
        17042       ,175    ,9281   ),34,3, byrow=TRUE)

DATA1=DATA

DATA2=data.frame(DATA1)
DATA2

cor(DATA2)

library(MVN)
```

```
outlier= mvn(DATA2, mvnTest = "hz", univariateTest = "AD", multivariatePlot =
      "qq", multivariateOutlierMethod = "adj", showOutliers = TRUE,
      showNewData = TRUE)
outlier

STANDARISASI=scale(DATA2)
STANDARISASI

library(MASS)
MCD=cov.mcd(STANDARISASI)
MCD

meanMCD=MCD$center
meanMCD

covMCD=MCD$cov
covMCD

Datanew=STANDARISASI-meanMCD
Datanew

nilai_eigen=eigen(covMCD)
nilai_eigen

jumlah_nilai_eigen=(sum(nilai_eigen$values))
jumlah_nilai_eigen

proporsi_eigen1=((nilai_eigen$values[1])/jumlah_nilai_eigen)
proporsi_eigen2=((nilai_eigen$values[2])/jumlah_nilai_eigen)
proporsi_eigen3=((nilai_eigen$values[3])/jumlah_nilai_eigen)

proporsi=data.frame(proporsi_eigen1, proporsi_eigen2, proporsi_eigen3)
proporsi

vektor_eigen1=nilai_eigen$vectors[,1]
vektor_eigen1
vektor_eigen2=nilai_eigen$vectors[,2]
vektor_eigen2
vektor_eigen3=nilai_eigen$vectors[,3]
vektor_eigen3

a1=nilai_eigen$vectors[,1:3]
a1

komponen_utama=Datanew%*%a1
komponen_utama
plot(komponen_utama)
```

```
plot(nilai_eigen$values, xlab = 'Eigenvalue Number', ylab = 'Eigenvalue Size',
        main = 'Scree Graph')
lines(nilai_eigen$values)

### TCLUST ###
library ("tclust")

## KLASTER 2 ##
clus2 = tclust (komponen_utama, k = 2, alpha = 0.35)
clus2
obj2=clus2$obj
obj2
plot (clus2)

## KLASTER 3 ##
clus3 = tclust (komponen_utama, k = 3, alpha = 0.35)
clus3
obj3=clus3$obj
obj3
plot (clus3)

## KLASTER 4 ##
clus4 = tclust (komponen_utama, k = 4, alpha = 0.35)
clus4
obj4=clus4$obj
obj4
plot (clus4)

## KLASTER 5 ##
clus5 = tclust (komponen_utama, k = 5, alpha = 0.35)
clus5
obj5=clus5$obj
obj5
plot (clus5)

## KLASTER 6 ##
clus6 = tclust (komponen_utama, k = 6, alpha = 0.35)
clus6
obj6=clus6$obj
obj6
plot (clus6)


hasilobj=cbind(obj2,obj3,obj4)
hasilobj

plot (ctlcurves (komponen_utama, k = 2:4, alpha = seq (0, 0.4, by = 0.05)))
```

```
## TRIMMED K-MEANS ##

library(trimcluster) #dengan package(trimcluster)#
tkm2 = trimkmeans (komponen_utama, k = 2, alpha = 0.35)
tkm2



library(tclust) #dengan package(tclust)#
## KLASTER 2 ##
tkm2 = tkmeans (komponen_utama, k = 2, alpha = 0.35)
tkm2
plot (tkm2)

## KLASTER 3 ##
tkm3 = tkmeans (komponen_utama, k = 3, alpha = 0.35)
tkm3
plot (tkm3)

## KLASTER 4 ##
tkm4 = tkmeans (komponen_utama, k = 4, alpha = 0.35)
tkm4
plot (tkm4)



## K-MEANS ##
library(cluster)

## KLASTER 2 ##
cl2=kmeans(komponen_utama, 2)
```