ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 11, Nomor 2, Tahun 2022, Halaman 302 - 312
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENGELOMPOKAN PROVINSI DI INDONESIA BERDASARKAN INDIKATOR KESEHATAN LINGKUNGAN MENGGUNAKAN METODE *PARTITIONING AROUND MEDOIDS* DENGAN VALIDASI INDEKS INTERNAL

Diah Aliyatus Saidah[1*], Rukun Santoso[2], Tatik Widiharih[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***Email :** diahaliya27@gmail.com

ABSTRACT

Environmental health is an important aspect in efforts to achieve public health. The condition of environmental health in Indonesia is varies in each province, so the priorities for increasing environmental health are also different. This study aims to grouping provinces in Indonesia based on environmental health indicators in order to know the high/low environmental quality in each province to assist the government in optimizing environmental health efforts. The grouping of provinces is done partitioning around medoids method which is robust to data containing outliers. The measure of similarity objects is calculated using the Euclidean and Manhattan distances, the selection of the best number of clusters is done by validating the internal index, namely the Calinski-Harabasz index, Baker-Hubert index, silhouette index, C-index, and Davies-Bouldin index. The result of this study is that the best number of clusters are two clusters using the Manhattan distance measurement method, with the largest Calinski-Harabasz index value = 24.10072, the largest Baker-Hubert index = 0.8466251, the largest silhouette index = 0.4246581, the smallest C-index = 0.07290109, and the smallest Davies-Bouldin index = 1.094805.

***Keywords**: environmental health, cluster analysis, partitioning around medoids, internal index.*

1. PENDAHULUAN

Rumah atau tempat tinggal merupakan salah satu kebutuhan pokok yang diperlukan untuk keberlangsungan hidup manusia. Sebagai upaya pencegahan gangguan kesehatan dari faktor lingkungan, memperhatikan kesehatan lingkungan merupakan hal yang penting untuk mewujudkan kualitas lingkungan yang sehat dari berbagai faktor (BPS, 2018). Lingkungan sehat yang diharapkan dapat terwujud adalah lingkungan yang tersusun dengan baik dilihat dari berbagai faktor yang ada dalam lingkungan manusia, serta dikelola dengan baik sehingga dapat meningkatkan derajat kesehatan masyarakat.

Kondisi kesehatan lingkungan di tiap-tiap daerah di Indonesia memiliki keragaman yang berbeda apabila dikaitkan dengan indikator kesehatan lingkungan. Oleh sebab itu, prioritas program penyehatan lingkungan di masing-masing daerah juga berbeda. Berdasarkan hal tersebut, maka penulis mencoba membuat pengelompokan provinsi di Indonesia berdasarkan indikator kesehatan lingkungan. Pengelompokan ini dilakukan untuk mengetahui kemiripan atau kesamaan provinsi-provinsi tersebut berdasarkan indikator kesehatan lingkungan.

Analisis klaster merupakan salah satu teknik analisis dalam statistika yang dapat digunakan untuk melakukan pengelompokan objek-objek. Salah satu metode dalam analisis klaster adalah metode *partitioning around medoids* atau sering disebut sebagai metode *k-medoids.* Metode *partitioning* yang sering digunakan adalah *k-means* dan *k-medoids.* Menurut Han dan Kamber (2006), metode *partitioning around medoids* ini dapat digunakan untuk mengelompokkan objek-objek yang mengandung pencilan, sehingga metode ini lebih baik daripada metode *k-means* yang sensitif terhadap pencilan.

Hal yang perlu diperhatikan dalam analisis klaster adalah validasi hasil klaster yang terbentuk. Validasi tersebut dilakukan agar memperoleh jumlah klaster yang paling sesuai dengan data. Penelitian ini akan mengelompokkan provinsi di Indonesia berdasarkan indikator kesehatan lingkungan dengan metode *partitioning around medoids* dengan dua pengukuran jarak, yaitu jarak Euclidean dan Manhattan, dengan menggunakan lima

validasi indeks internal, yaitu validasi Calinski-Harabasz *index,* Baker-Hubert *index, silhouette index, C-index*, dan Davies-Bouldin *index* untuk menentukan jumlah klaster yang optimal.

2. TINJAUAN PUSTAKA

Himpunan Ahli Kesehatan (HAKLI) mendefinisikan kesehatan lingkungan sebagai suatu kondisi lingkungan yang dapat menopang keseimbangan dan hubungan timbal balik yang dinamis antara manusia dengan lingkungannya, sehingga kualitas hidup manusia yang sehat dan bahagia dapat tercapai (Mundiatun dan Daryanto, 2015). Penelitian ini menggunakan beberapa indikator atau ruang lingkup kesehatan lingkungan meliputi: ketersediaan air minum layak, pengelolaan limbah medis, kebijakan Perilaku Hidup Bersih dan Sehat (PHBS), pemenuhan kualitas kesehatan lingkungan di masing-masing wilayah, penyelenggaraan tatanan kawasan sehat, kebersihan tempat pengolahan makanan, akses sanitasi layak, kebersihan tempat-tempat umum, sanitasi total berbasis masyarakat, dan kampanye Gerakan Masyarakat Sehat (Germas). Ruang lingkup ini sesuai dengan fokus dari Kementerian Kesehatan Republik Indonesia dalam upaya penyehatan lingkungan di Indonesia yang tertuang dalam Profil Kesehatan Indonesia tahun 2019. Pada penelitian ini, indikator-indikator tersebut akan digunakan sebagai variabel multivariat yang akan dianalisis dengan analisis klaster menggunakan metode *partitioning araound medoids*.

Analisis klaster merupakan salah satu metode dalam analisis multivariat yang bertujuan untuk melakukan pengelompokan objek-objek berdasarkan karakteristik yang dimilikinya. Analisis klaster melakukan pengelompokan terhadap individu atau objek dalam penelitian, sehingga setiap objek yang mempunyai kemiripan paling dekat dengan objek lain akan berada pada kelompok/klaster yang sama. Objek-objek yang berada dalam satu klaster yang sama memiliki ciri yang relatif sama (homogen), sedangkan antar satu klaster dengan klaster yang lain memiliki karakteristik yang berbeda (heterogen). Pengelompokan tersebut dilakukan berdasarkan variat-variat yang diamati (Usman dan Sobari, 2013).

Menurut Hair et al. (2010), asumsi yang harus terpenuhi dalam analisis klaster adalah:

1. Kecukupan Sampel (Sampel Representatif)

Sampel yang representatif atau sampel yang cukup merupakan sampel yang dapat merepresentasikan atau mewakili populasi yang ada. Pengujian sampel representatif dapat dilakukan dengan uji Kaiser-Mayer-Olkin (KMO). Sampel dapat dikatakan representatif apabila nilai KMO berkisar antara 0,5 sampai 1. KMO dirumuskan sebagai berikut (Widarjono, 2010):

$$KMO = \frac{\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} r_{x_j x_k}^2}{\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} r_{x_j x_k}^2 + \sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} \rho_{x_j x_k, x_l}^2} \quad (1)$$

dengan, p = banyaknya variabel, n = banyaknya objek, x_j = objek pada pengamatan ke- j, $r_{x_j x_k}$ = korelasi antar variabel x_j dan x_k, $\bar{x}_j$ = rata-rata variabel x_j, dan $\rho_{x_j x_k, x_l}$ = korelasi parsial antar variabel x_j dan x_k dengan menjaga agar x_l konstan.

2. Tidak Terjadi Multikolinearitas (Non-Multikolinearitas)

Menurut Gujarati (2009), multikolinearitas didefinisikan sebagai adanya hubungan linear yang pasti atau sempurna antara beberapa atau semua variabel dalam penelitian.

Untuk mendeteksi adanya multikolinearitas salah satunya dengan menggunakan nilai *Variance Inflation Factor (VIF)* yang dirumuskan sebagai berikut:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{2}$$

dengan: R_j^2 = koefisien determinasi variabel ke. Pada analisis klaster, R_j^2 dapat diperoleh dengan meregresikan variabel ke- j dengan variabel lainnya. Apabila nilai $VIF_j \geq 10$, maka dapat dikatakan bahwa terjadi multikolinearitas dalam data.

Pengelompokan objek dalam analisis klaster dapat dilakukan dengan beberapa macam pengukuran jarak. Dalam penelitian ini, akan digunakan metode pengukuran jarak Euclidean *(Euclidean distance)* dan jarak Manhattan *(Manhattan distance).*

1. Jarak Euclidean *(Euclidean Distance)*

Jarak Euclidean mengukur jarak objek degan menghitung akar kuadrat dari penjumlahan kuadrat selisih nilai dari masing-masing variabel pada objek. Jarak Euclidean dirumuskan sebagai berikut (Anderberg, 1973):

$$d_{i,j} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}\ ,\ i = 1,2,3...,n \text{ dan } k = 1,2,3,...,k \tag{3}$$

dengan: $d_{i,j}$ = jarak antara objek i dengan objek j, x_{ik} = nilai objek i pada variat ke-k, x_{jk} = nilai objek j pada variat ke-k, p = banyaknya variat yang diamati.

2. Jarak Manhattan *(Manhattan Distance)*

Jarak *Manhattan* mengukur jarak objek dengan menghitung jumlah absolut perbedaan objek pada masing-masing variabel. Jarak Manhattan dirumuskan sebagai berikut (Anderberg, 1973):

$$d_{i,j} = \sum_{k=1}^{p} \left| x_{ij} - x_{jk} \right|\ ,\ i = 1,2,3...,n \text{ dan } k = 1,2,3,...,k \tag{4}$$

dengan: $d_{i,j}$ = jarak antara objek i dengan objek j, x_{ik} = nilai objek i pada variat ke-k, x_{jk} = nilai objek j pada variat ke-k, p = banyaknya variat yang diamati.

Metode *Partitioning Around Medoids (*PAM*)* merupakan metode *non-hierarchical clustering* yang merupakan pengembangan dari metode *k-means*. Algoritma *partitioning around medoids* menggunakan partisi klastering untuk mengelompokkan sejumlah objek menjadi beberapa klaster. Algoritma ini menggunakan objek untuk mewakili sebuah klaster pada sekumpulan objek. Objek yang dipilih untuk mewakili sebuah klaster disebut dengan *medoids*. Suatu klaster akan dibentuk melalui perhitungan kedekatan jarak yang dimiliki antara *medoid* dengan objek *non-medoid* (Kauffman dan Rousseeuw, 1990).

Metode PAM hadir untuk mengatasi kelemahan dari metode *k-means* yang sensitif terhadap pencilan dikarenakan suatu objek dengan suatu nilai yang besar mungkin menyimpang dari distribusi data (Han dan Kamber, 2006). Menurut Jobson (1992), untuk mendeteksi adanya pencilan pada data multivariat, dapat dilakukan dengan menghitung jarak dari tiap nilai pengamatan ke pusat data (rataan semua variat) menggunakan jarak kuadrat Mahalanobis. Jarak kuadrat Mahalanobis objek ke-i dengan pusat data dari semua objek yang diteliti dapat dihitung dengan rumus berikut:

$$d_{MD}^2(i) = (\mathbf{x_i} - \bar{\mathbf{x}}_\mathbf{p})^\mathbf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x_i} - \bar{\mathbf{x}}_\mathbf{p})\ ,\ i = 1,2,...,n \tag{5}$$

dengan, $d_{MD}^2(i)$ = jarak kuadrat Mahalanobis objek ke-i dengan pusat data dari semua objek yang diteliti, $\mathbf{x_i} = \begin{bmatrix} x_{i1} & x_{2i} & \dots & x_{ip} \end{bmatrix}^T$ adalah vektor objek pada pengamatan ke-i

untuk setiap variat ke- p berukuran $1\times p$, $\bar{\mathbf{x}}_{\mathbf{p}} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \end{bmatrix}^T$ adalah vektor rata-rata dari tiap variat ke- p berukuran $1\times p$, $\boldsymbol{\Sigma}$ = matriks kovarian dari vektor variat berukuran $p\times p$. Jarak kuadrat *Mahalanobis* kemudian dievaluasi dengan distribusi *chi-kuadrat* ($\chi^2_{\alpha,p}$) Apabila sebuah data memiliki nilai jarak kuadrat Mahalanobis (d^2_{MD}) yang lebih besar dari nilai $\chi^2_{\alpha,p}$, maka data tersebut diidentifikasi sebagai data pencilan.

Menurut Han dan Kamber (2006), algoritma dari *partitioning around medoids* adalah sebagai berikut:

1. Menentukan k sebagai jumlah klaster yang ingin dibentuk
2. Memilih k objek pada sekumpulan n objek sebagai *medoid* awal
3. Menghitung jarak objek *non-medoid* dengan *medoid* awal, kemudian menempatkan tiap objek *non-medoid* ke dalam klaster dengan jarak paling dekat dengan *medoid* awal, kemudian menghitung total jaraknya
4. Memilih objek *non-medoid* pada masing-masing klaster secara acak sebagai calon *medoid* baru
5. Menghitung jarak objek *non-medoid* dengan calon *medoid* baru dengan jarak *Euclidean* dan *Manhattan*, kemudian menempatkan objek *non-medoid* tersebut ke calon *medoid* baru yang terdekat, kemudian menghitung total jaraknya
6. Menghitung selisih total jarak ($S_{\text{total jarak}}$), dengan $S_{\text{total jarak}}$ = total jarak pada calon *medoid* baru – total jarak pada *medoid* lama
7. Apabila nilai $S_{\text{total jarak}} < 0$, maka calon *medoid* baru tersebut menjadi *medoid* baru, apabila diperoleh $S_{\text{total jarak}} > 0$, maka iterasi berhenti
8. Mengulangi langkah (4) sampai (7) hingga $S_{\text{total jarak}} > 0$.

Nilai selisih total jarak yang lebih dari 0 memiliki arti bahwa total jarak pada *medoid* lama lebih kecil daripada total jarak pada calon *medoid* baru, sehingga dapat disimpulkan bahwa *medoid* lama tersebut memiliki jarak yang lebih dekat dengan objek *non-medoid* apabila dibandingkan dengan jarak pada calon *medoid* baru ke objek *non-medoid*. Berdasarkan hal tersebut maka penentuan anggota klaster didasarkan pada proses iterasi yang menghasilkan jumlah jarak yang kecil antara objek *medoid* dan objek-objek *non-medoid*.

Hal yang perlu diperhatikan dalam analisis klaster adalah mengevaluasi banyaknya klaster yang diperoleh dari algoritma klastering yang digunakan guna mendapatkan pengelompokan yang sesuai dengan data penelitian. Setiap klaster yang terbentuk mempunyai seperangkat ukuran karakteristik, diantaranya adalah nilai indeks validitas klaster (Brock et al., 2008). Penelitian ini akan menggunakan lima indeks dengan kriteria internal untuk mengevaluasi hasil klaster yang terbentuk, yaitu Calinski-Harabasz *index,* Baker-Hubert *index, silhouette index*, *C-index,* dan Davies-Bouldin *index.*

1. Calinski-Harabasz *Index*

Calinski-Harabasz (CH) *index* memberikan penilaian pada hasil klaster berdasarkan pada perbandingan nilai *sum of square between cluster (SSB)* sebagai *separation* dan *sum of square within cluster (SSW)* sebagai *compactness* yang dikalikan dengan faktor normalisasi, yakni selisih banyaknya data dengan jumlah klaster dan dibagi dengan jumlah klaster dikurangi satu. Semakin besar nilai Calinski-Harabasz *index* maka semakin baik hasil klaster tersebut. (Baarsch dan Celebi, 2012).

Nilai validitas Calinski-Harabasz *index* dapat dihitung dengan persamaan berikut:

$$CH = \frac{trace(SSB)}{trace(SSW)} \times \frac{n-k}{k-1} \tag{6}$$

dengan, n = jumlah semua objek yang diteliti, dan k = jumlah klaster.

2. Baker-Hubert *Index*

Baker-Hubert *index* merupakan validasi hasil klaster yang dikemukakan oleh Baker dan Hubert pada tahun 1976, yang dapat dihitung dengan persamaan berikut:

$$BH(k) = \frac{S^+ - S^-}{S^+ + S^-} \tag{7}$$

dengan, S^+ = jumlah dua pasang objek yang *concordant*, S^- = jumlah dua pasang objek yang *disconcordant.* Dua pasang objek dikatakan *concordant* apabila $d(q,r) < d(s,t)$, dengan objek q dan r berada dalam klaster yang sama, sedangkan objek s dan t berada dalam klaster yang berbeda. Dua pasang objek dikatakan *disconcordant* apabila $d(q,r) < d(s,t)$, objek q dan r berada dalam klaster yang berbeda, objek s dan t berada dalam klaster yang sama. Nilai dari Baker-Hubert *index* berada pada rentang -1 sampai 1, nilai indeks yang terbesar menunjukkan jumlah klaster yang optimal (Charrad, et al., 2010).

3. *Silhouette Index*

Silhouette index menghitung rata-rata masing-masing titik pada sekumpulan data. Adapun rumus yang digunakan untuk menghitung *silhouette index* adalah (Rousseeuw, 1987):

$$SI = \frac{1}{k}\sum_{j=1}^{k} SI_j \tag{8}$$

dengan, $SI_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}}$, $a_i^j = \frac{1}{n_j - 1}\sum_{r=1, r\neq i}^{n_j} d(x_i^j, x_r^j)$, $b_i^j = \min\left\{\frac{1}{n_l}\sum_{r=1, r\neq i}^{n_l} d(x_i^j, x_r^l)\right\}, l \neq j, l = 1,...,k$

a_i^j = rata-rata objek ke-i dengan semua objek lainnya dalam satu klaster j

b_i^j = nilai minimum rata-rata objek ke-i dengan semua data dari klaster selain j

Nilai *silhouette coefficient* (SC) berada pada rentang –1 sampai 1, dan didapatkan dengan persamaan berikut:

$$SC = \max_k SI(k) \tag{9}$$

semakin besar nilai *silhouette coefficient*, maka semakin baik hasil klaster yang terbentuk.

4. *C-Index*

C-Index merupakan indeks validasi hasil klaster dengan kriteria internal yang ditemukan oleh Hubert dan Levin pada tahun 1976. Nilai *C-Index* berada pada rentang 0 – 1, nilai minimum dari indeks tersebut menunjukkan jumlah klaster yang optimal. *C-Index* dapat dihitung dengan persamaan (Charrad et al., 2014):

$$C - Index = \frac{S_w - S_{\min}}{S_{\max} - S_{\min}}, S_{\min} \neq S_{\max}, C - Index \in (0,1) \tag{10}$$

dengan, S_w = jumlah jarak objek dalam satu klaster, $S_{\min}$ = jumlah jarak terkecil antara semua pasangan objek data dalam satu klaster dan antar klaster, $S_{\max}$ = jumlah jarak terbesar antara semua pasangan objek data dalam satu klaster dan antar klaster.

5. Davies-Bouldin *Index*

Davies-Bouldin (DB) *index* menghitung rata-rata nilai setiap objek pada sekumpulan data. Nilai setiap objek dapat dihitung dengan menjumlahkan nilai *compactness* yang dibagi dengan jarak antara kedua objek pusat klaster sebagai *separation.* Nilai indeks DB yang terkecil menunjukkan jumlah klaster yang terbaik (Davies dan Bouldin, 1979). Indeks validitas Davies-Bouldin (DB) *index* dapat dihitung dengan persamaan berikut:

$$DB = \frac{1}{k}\sum_{p=1}^{k} R_p \tag{11}$$

dengan, $R_p = \max\left(\frac{(S_p + S_q)}{M_{pq}}\right), p \neq q$, $S_p = \frac{1}{n_p}\sum_{i=1}^{n_p} d(x_i, c_p)$, $S_q = \frac{1}{n_q}\sum_{i=1}^{n_q} d(y_j, c_q)$, $M_{pq} = d(c_p, c_q)$, S_p = rata-rata jarak setiap objek pada klaster p ke titik pusat klaster, S_q = rata-rata jarak setiap objek pada klaster q ke titik pusat klaster, c_p = titik pusat klaster p , c_q = titik pusat klaster q , n_p = jumlah objek pada klaster p , n_q = jumlah objek pada klaster q

3. METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder indikator kesehatan lingkungan pada tahun 2019 di 34 provinsi di Indonesia yang diperoleh dari publikasi Profil Kesehatan Indonesia tahun 2019 oleh Kementerian Kesehatan Republik Indonesia.

Variabel yang digunakan dalam penelitian ini adalah: persentase kabupaten/kota yang memiliki kebijakan PHBS (X_1), persentase rumah sakit dengan pengelolaan limbah medis sesuai standar (X_2), persentase kabupaten/kota yang memenuhi kualitas kesehatan lingkungan (X_3), persentase kabupaten/kota yang menyelenggarakan tatanan kawasan sehat (X_4), persentase tempat pengolahan makanan yang memenuhi syarat kesehatan (X_5), persentase keluarga dengan akses terhadap fasilitas sanitasi yang layak (X_6), persentase tempat-tempat umum yang memenuhi syarat kesehatan(X_7), persentase desa yang melaksanakan sanitasi total berbasis masyarakat (X_8), persentase rumah tangga yang memiliki akses terhadap air minum layak (X_9), dan persentase kabupaten/kota yang melakukan minimal lima kampanye Germas (X_{10}).

Langkah-langkah analisis data yang dilakukan adalah sebagai berikut:

1. Melakukan pendeteksian terhadap pencilan dengan kuadrat jarak Mahalanobis.
2. Melakukan pengujian asumsi analisis klaster, yaitu:
   a. Uji asumsi kecukupan sampel dengan uji KMO.
   b. Uji asumsi non-multikolinieritas menggunakan nilai VIF.
3. Menentukan k sebagai jumlah klaster yang ingin dibentuk, nilai k yang digunakan dalam penelitian ini adalah k = 2, 3, 4, 5, dan 6.
4. Melakukan analisis klaster dengan algoritma *partitioning around medoids.*
5. Melakukan validasi hasil klaster dengan menggunakan indeks internal, yaitu:
   a. Menghitung nilai Calinski-Harabasz *index,* Baker-Hubert *index, silhouette coefficient, C-index,* dan Davies-Bouldin *index* pada masing-masing k klaster.
   b. Membandingkan nilai masing-masing indeks validasi. Hasil klaster yang terbaik ditunjukkan dengan nilai Calinski-Harabasz *index,* Baker-Hubert *index, dan silhouette coefficient* yang besar, serta memiliki nilai *C-index* dan Davies-Bouldin *index* yang kecil.
6. Melakukan interpretasi karakteristik daerah berdasarkan hasil pengklasteran yang terbaik.

4. HASIL DAN PEMBAHASAN

Pendeteksian pencilan dilakukan dengan metode jarak kuadrat *Mahalanobis*, objek diidentifikasi sebagai pencilan apabila memiliki nilai jarak kuadrat *Mahalanobis* lebih besar dari nilai distribusi *chi-kuadrat* ($\chi^2_{\alpha,p}$). Penelitian ini menggunakan 10 variat pengamatan dengan α = 5%, sehingga nilai distribusi *chi-kuadrat* ($\chi^2_{0,05;10}$) adalah sebesar 18,307038. Berdasarkan perhitungan jarak kuadrat Mahalanobis, dapat diketahui bahwa terdapat 2 provinsi yang memiliki nilai jarak kuardat Mahalanobis yang lebih besar

daripada nilai kritis distribusi *chi-kuadrat* ($\chi_{0,05;10} = 18,307038$), yaitu Provinsi Nusa Tenggara Timur dan Provinsi Papua dengan nilai jarak kuadrat Mahalanobis berturut-turut sebesar 21,633519 dan 23,652925, sehingga kedua provinsi tersebut diidentifikasi sebagai pencilan. Oleh sebab itu, dapat disimpulkan bahwa data indikator kesehatan lingkungan memiliki pencilan secara multivariat, sehingga metode *partitioning around medoids* yang *robust* terhadap pencilan merupakan metode yang tepat untuk melakukan klasterisasi provinsi di Indonesia berdasarkan indikator kesehatan lingkungan.

Analisis klaster memiliki dua asumsi yang harus dipenuhi, yaitu asumsi sampel representatif, dan asumsi non-multikolinearitas. Berdasarkan *output* uji KMO, diperoleh nilai KMO sebesar 0,7616654. Nilai KMO tersebut lebih besar dari 0,5, sehingga dapat disimpulkan bahwa sampel yang digunakan merupakan sampel yang representatif, sehingga asumsi kecukupan sampel telah terpenuhi. Pengujian asumsi non-multikolinieritas dilakukan dengan menghitung nilai VIF dengan hasil perhitungan ditampilkan pada **Tabel 1** berikut:

Tabel 1. Nilai Koefisien Determinasi dan VIF

Variabel	VIF	Variabel	VIF
X_1	2,1533	X_6	3,4746
X_2	2,2810	X_7	1,5246
X_3	3,2862	X_8	2,0255
X_4	3,8241	X_9	1,9639
X_5	1,7425	X_{10}	1,2719

Berdasarkan **Tabel 1** tersebut, dapat diketahui bahwa dari variabel X_1 sampai X_{10} tidak terdapat variabel yang memiliki nilai VIF yang lebih besar dari 10, sehingga dapat disimpulkan bahwa asumsi non-multikolinearitas pada data indikator kesehatan lingkungan telah terpenuhi sehingga pengolahan data dapat dilanjutkan.

Berdasarkan analisis pengklasteran dengan metode *partitioning around medoids* dengan k = 2,3,4,5, dan 6 yang diolah menggunakan *software* RStudio 4.0.3, dapat diketahui objek *medoid* dan jumlah anggota masing-masing klaster yang ditampilkan pada **Tabel 2** berikut:

Tabel 2. Hasil Klasterisasi dengan k = 2,3,4,5, dan 6.

K	Jarak	Klaster ke-	Jumlah Anggota	*Medoid*
2	Euclidean	1	12	Objek ke-2
		2	22	Objek ke-4
	Manhattan	1	10	Objek ke-1
		2	24	Objek ke-12
3	Euclidean	1	5	Objek ke-28
		2	25	Objek ke-4
		3	4	Objek ke-33
	Manhattan	1	10	Objek ke-1
		2	17	Objek ke-5
		3	7	Objek ke-9
4	Euclidean	1	5	Objek ke-28
		2	15	Objek ke-4
		3	10	Objek ke-22
		4	4	Objek ke-33
	Manhattan	1	9	Objek ke-1

		2	17	Objek ke-5
		3	7	Objek ke-9
		4	1	Objek ke-34
5	Euclidean	1	5	Objek ke-28
		2	11	Objek ke-4
		3	10	Objek ke-22
		4	4	Objek ke-13
		5	4	Objek ke-33
	Manhattan	1	9	Objek ke-13
		2	14	Objek ke-5
		3	4	Objek ke-13
		4	6	Objek ke-23
		5	1	Objek ke-34
6	Euclidean	1	5	Objek ke-28
		2	9	Objek ke-4
		3	9	Objek ke-22
		4	4	Objek ke-13
		5	4	Objek ke-33
		6	3	Objek ke-26
	Manhattan	1	5	Objek ke-32
		2	4	Objek ke-28
		3	14	Objek ke-5
		4	4	Objek ke-13
		5	6	Objek ke-23
		6	1	Objek ke-34

Penentuan jumlah klaster terbaik dilakukan dengan melihat nilai masing-masing indeks validasi untuk jumlah klaster k = 2,3,4,5, dan 6, yang ditampilkan pada **Tabel 3** berikut ini:

Tabel 3. Nilai Indeks Validasi Jumlah Klaster

k	Jarak	Indeks Validasi				
		Calinski-Harabasz *Index*	Baker-Hubert *Index*	*Silhouette Index*	*C-Index*	Davies-Bouldin *Index*
2	Euclidean	22,08642	0,6988318	0,3496990	0,1365175	1,188206
	Manhattan	24,10072	0,8466251	0,4246581	0,07290109	1,094805
3	Euclidean	16,09828	0,8273901	0,3327904	0,07773793	1,246011
	Manhattan	15,77764	0,6633112	0,1962943	0,1298273	1,54937
4	Euclidean	13,98702	0,6775355	0,1790045	0,1264020	1,523059
	Manhattan	12,78521	0,7188274	0,1854562	0,1022985	1,309955
5	Euclidean	13,16582	0,7135815	0,1827292	0,1088631	1,430638
	Manhattan	11,29940	0,6440024	0,1501440	0,1220172	1,307501
6	Euclidean	12,37859	0,7511998	0,1934178	0,09392544	1,305382
	Manhattan	10,64936	0,7123629	0,1326853	0,09241654	1,425537

Berdasarkan **Tabel 3** tersebut dapat diketahui bahwa klaster terbaik yang dapat dibentuk berjumlah dua klaster dengan pengukuran jarak Manhattan, dikarenakan memiliki nilai Calinski-Harabasz *index,* Baker-Hubert *index,* dan *silhouette index* terbesar, serta

memiliki nilai *C-index* dan Davies-Bouldin *index* terkecil. Profilisasi hasil pengklasteran yang terbaik ditampilkan pada **Tabel 4** berikut:

Tabel 4. Hasil Klasterisasasi Terbaik (k = 2)

Klaster	Objek *Medoid*	Jumlah Anggota	Provinsi
1	Objek ke-1 (Provinsi Aceh)	10	Aceh, Sumatera Utara, Nusa Tenggara Timur, Kalimantan Barat, Sulawesi Utara, Sulawesi Tenggara, Maluku, Maluku Utara, Papua Barat, dan Papua.
2	Objek ke-12 (Provinsi Jawa Barat)	24	Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Bali, Nusa Tenggara Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Tengah, Sulawesi Selatan, Gorontalo, dan Sulawesi Barat.

Interpretasi hasil klaster dilakukan dengan melihat nilai rata-rata masing-masing variabel pada tiap klaster yang dapat dilihat pada **Tabel 5** berikut:

Tabel 5. Nilai Rata-Rata Masing-Masing Variabel pada Setiap Klaster

Variabel	Klaster 1	Klaster 2
Persentase kabupaten/kota yang memiliki kebijakan PHBS (X_1)	63,2490	95,0883
Persentase rumah sakit dengan pengolahan limbah medis sesuai standar (X_2)	13,6840	50,1396
Persentase kabupaten/kota yang memenuhi kualitas kesehatan lingkungan (X_3)	43,8500	96,9750
Persentase kabupaten/kota yang menyelenggarakan tatanan kawasan sehat (X_4)	38,7310	88,5317
Persentase tempat pengolahan makanan yang memenuhi syarat kesehatan (X_5)	28,7030	41,9600
Persentase keluarga dengan akses terhadap fasilitas sanitasi yang layak (X_6)	79,1980	87,9979
Persentase tempat-tempat umum yang memenuhi syarat kesehatan (X_7)	53,7490	61,6400
Persentase desa yang melaksanakan sanitasi total berbasis masyarakat (X_8)	10,0140	31,9342
Persentase rumah tangga yang memiliki akses terhadap air minum layak (X_9)	65,5670	74,8837
Persentase kabupaten/kota yang melakukan minimal lima kampanye GERMAS (X_{10})	20,2050	19,8533

Berdasarkan **Tabel 4** dan **Tabel 5** di atas, diperoleh informasi sebagai berikut:

a. Klaster 1

Anggota klaster satu terdiri dari 10 provinsi, yaitu: Provinsi Aceh, Sumatera Utara, Nusa Tenggara Timur, Kalimantan Barat, Sulawesi Utara, Sulawesi Tenggara, Maluku,

Maluku Utara, Papua Barat, dan Papua. Klaster ini memiliki rata-rata 9 variabel/indikator kesehatan lingkungan yang lebih rendah daripada klaster dua, yaitu variabel X_1 sampai X_9, sedangkan variabel X_{10} pada klaster satu memiliki rata-rata yang lebih tinggi daripada klaster dua.

b. Klaster 2

Anggota klaster dua terdiri dari 24 provinsi, yaitu: Provinsi Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Bali, Nusa Tenggara Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Tengah, Sulawesi Selatan, Gorontalo, dan Sulawesi Barat. Klaster ini memiliki rata-rata 9 variabel/indikator kesehatan lingkungan yang lebih tinggi daripada klaster satu, yaitu variabel X_1 sampai X_9, sedangkan variabel X_{10} pada klaster kedua memiliki rata-rata yang lebih rendah daripada klaster satu.

5. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan, diperoleh kesimpulan bahwa jumlah klaster terbaik untuk melakukan klasterisasi data indikator kesehatan lingkungan tahun 2019 menggunakan metode *partitioning around medoids* adalah dua klaster ($k = 2$) dengan pengukuran jarak Manhattan, karena memiliki nilai Calinski-Harabasz *index*, Baker-Hubert *index*, dan *silhouette index* yang terbesar, serta nilai *C-index* dan Davies-Bouldin *index* yang terkecil jika dibandingkan dengan jumlah *k* yang lain. Klaster satu terdiri dari 10 provinsi, klaster dua terdiri dari 24 provinsi. Klaster satu merupakan klaster untuk kelompok daerah yang memiliki tingkat kesehatan lingkungan yang masih rendah jika dibandingkan dengan klaster dua, sehingga diperlukan adanya kerja sama antara pemangku kebijakan dan masyarakat yang berada pada provinsi-provinsi di klaster satu untuk meningkatkan kualitas kesehatan lingkungan.

Berdasarkan penelitian yang telah dilakukan, masih terdapat beberapa perbaikan dan pengembangan yang dapat dilakukan untuk penelitian selanjutnya, seperti menggunakan pendekatan pengelompokan objek dengan pengukuran jarak yang lain. Selain itu, dapat juga melakukan klasterisasi dengan metode yang *robust* terhadap data pencilan menggunakan metode selain *partitioning around medoids*, seperti metode CLARA (*Clustering Large Application*), atau dapat membandingkan hasil klasterisasi dari kedua metode tersebut.

DAFTAR PUSTAKA

Anderberg, M. 1973. *Cluster Analysis for Application*. New York: Academic Press.

Baarsch, J., & Celebi, M. E. 2012. Investigastion of Internal Validity Measures for K-Means Clustering. *International Multiconference of Engineers and Computer Scientist 1*. Los Angeles: Louisiana Board of Regents. 14–16.

Baker, F., & Hubert, L. 1975. Measuring the Power of Hierarchical Cluster Analysis. *Journal of the American Statistical Association*. Vol. 70, 31-38.

[BPS] Badan Pusat Statistik. 2018. *Indikator Perumahan dan Kesehatan Lingkungan Tahun 2018*. Jakarta: Badan Pusat Statistik.

Brock, G., Vasyl, P., Susmita, D., & Somnath, D. 2008. CValid: An R Package for Cluster Validation. *Journal of Statistical Software*. Vol. 25, No.4, 1–22.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. 2014. Nbclust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*. Vol. 61, No.6, 1–36.

Charrad, M., Lechevallier, Y., Ahmed, M. B., & Saporta, G. 2010. On the Number of Clusters in Block Clustering Algorithms. *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23*. 392–397

Davies, D. L., & Bouldin, D. W. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gujarati, D. 2009. *Dasar-Dasar Ekonometrika*. Jakarta: Erlangga.

Hair, J. F., Anderson, R. E., Thatham, R. L., & Black, W. C. 2010. *Multivariate Data Analysis Seventh Edition*. New Jersey: Pearson Education Inc.

Han, J., & Kamber, M. 2006. *Data Mining: Concepts and Techniques*. San Fransisco: Elsevier Inc.

Jobson, J. D. 1992. *Applied Multivariate Data Analysis-Second Volume: Categorical and Multivariate Method*. New York: Springer-Verlag Inc.

Kaufman, L., & Rousseeuw, P. J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics)*. New York: Wiley.

[KEMENKES RI] Kementerian Kesehatan Republik Indonesia. 2019. *Profil Kesehatan Indonesia Tahun 2019*. Jakarta: Kementerian Kesehatan Republik Indonesia.

Mundiatun, & Daryanto. 2015. *Pengelolaan Kesehatan Lingkungan*. Yogyakarta: Gava Media.

Rousseeuw, P. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Jornal of Computational and Applied Mathematics*. Vol.20, 53–65.

Usman, H., & Sobari, N. 2013. *Aplikasi Teknik Multivariat*. Jakarta: Rajawali Pers.

Widarjono, A. 2010. *Analisis Statistika Multivariat Terapan*. Yogyakarta: UPP STIM YKPN.