IJSA (2021), Vol. 5, No. 1, 141-160
DOI: https://doi.org/10.29244/ijsa.v5i1p141-160

Indonesian Journal of Statistics and Its Applications
Available at https://ijsa.stats.id
e-ISSN: 2599-0802

Comparison of Soft and Hard Clustering: A Case Study on Welfare Level in Cities on Java Island*

Analisis Cluster dengan Menggunakan Hard Clustering dan Soft Clustering untuk Pengelompokkan Tingkat Kesejahteraan Kabupaten/Kota di Pulau Jawa

Nurafiza Thamrin[1] and Arie Wahyu Wijayanto[2‡]

[1,2]Politeknik Statistika STIS, Indonesia
‡corresponding author: ariewahyu@stis.ac.id

Abstract

The National Medium Term Development Plan 2020-2024 states that one of the visions of national development is to accelerate the distribution of welfare and justice. Cluster analysis is analysis that grouping of objects into several smaller groups where the objects in one group have similar characteristics. This study was conducted to find the best clustering method and to classify cities based on the level of welfare in Java. In this study, the cluster analysis that used was hard clustering such as K-Means, K-Medoids (PAM and CLARA), and Hierarchical Agglomerative as well as soft clustering such as Fuzzy C Means. This study use elbow method, silhouette method, and gap statistics to determine the optimal number of clusters. From the evaluation results of the silhouette coefficient, dunn index, connectivity coefficient, and Sw/Sb ratio, it was found that the best cluster analysis was Agglomerative Ward Linkage which produced three clusters. The first cluster consists of 27 cities with moderate welfare, the second cluster consists of 16 cities with high welfare, the third cluster consists of 76 cities with low welfare. With the best clustering results, the government of cities in Java shall be able to make a better policies of welfare based on the dominant indicators found in each cluster.

Keywords: agglomerative ward linkage, fuzzy c means, k-means, k-medoids, welfare.

* Received: Des 2020; Reviewed: Jan 2021; Published: Mar 2021

1. Pendahuluan

Sasaran utama dari pembangunan di setiap negara adalah peningkatan kesejahteraan masyarakat. Kesungguhan pemerintah dalam usaha mencapai target kesejahteraan tercantum dalam Rencana Pembangunan Jangka Menengah Nasional (RPJMN) 2020-2024 yang menyatakan bahwa salah satu visi pembangunan nasional adalah mempercepat keadilan dan pemerataan (BAPPENAS, 2019). Menurut (BPS, 2018), jumlah penduduk miskin di Indonesia tahun 2018 diperkirakan sebesar 26,58 juta penduduk atau sebesar 10,12% dari total penduduk. Pada Agustus 2018 tingkat pengangguran terbuka pada semua provinsi di pulau Jawa naik sebesar 0,395% secara rata-rata. Jumlah ini akan terus meningkat jika strategi penciptaan lapangan kerja tidak berubah dan mempengaruhi kesejahteraan secara tidak langsung. Permasalahan ini mengindikasikan kebijakan yang kurang tepat sasaran dalam pelaksanaan pembangunan di pulau Jawa. Sehingga perlu adanya identifikasi karakteristik pembangunan berdasarkan tingkat kesejahteraan masyarakat masing–masing daerah untuk terciptanya kebijakan dan strategi yang tepat guna dalam proses percepatan pembangunan. Pengukuran kesejahteraan membutuhkan beberapa indikator yang mendetail dari berbagai aspek. Indikator kesejahteraan rakyat adalah cerminan kualitas sumber daya manusia dari suatu negara yang mencakup indikator pola konsumsi, perumahan dan lingkungan, kependudukan, pendidikan, gizi, kesehatan, pengeluaran dan ketenagakerjaan (BPS, 2018).

Beberapa penelitian telah dilakukan untuk memilih pengelompokkan terbaik. Soemartini & Supartini (2017) melakukan *clustering* dengan metode *K-means clustering* dengan menggunakan indikator kepadatan penduduk, angkatan kerja, laju pertumbuhan penduduk, rata-rata pengeluaran per kapita, angka harapan hidup dan rata-rata lama sekolah. Penelitian lain dari Alwi & Hasrul (2018) menggunakan metode *Average Linkage clustering* dengan indikator PDRB tiap kabupaten/kota, jumlah penduduk miskin, kepadatan penduduk, daya beli, jumlah angkatan kerja, angka melek huruf, angka harapan hidup, angka harapan lama sekolah, rata-rata lama sekolah, kepemilikan rumah sendiri, dan tingkat pengangguran terbuka. Hidayatullah (2014) juga menggunakan metode *Average Linkage* dengan indikator pengeluaran riil perkapita, PDRB perkapita, kepadatan penduduk, angka harapan hidup, dan rata-rata lama sekolah yang mempengaruhi tingkat kesejahteraan.

Analisis komponen utama dilakukan untuk mereduksi data yang awalnya memiliki dimensi yang tinggi (banyak variabel penyusun) menjadi lebih sedikit dengan tetap meminimalisasi resiko kehilangan informasi. Banyak penelitian yang telah membuktikan bahwa penggunaan analisis komponen utama dapat meningkatkan dan mempercepat kinerja metode *clustering*. Hal ini disebabkan reduksi dimensi dapat menghilangkan fitur yang tidak sesuai serta mengurangi *noise* dan *curse of dimensionality* (Izzuddin, 2015; Rahayu & Mustakim, 2017).

Penelitian ini bertujuan untuk mengelompokkan kabupaten/kota di pulau Jawa berdasarkan karakteristik indikator tingkat kesejahteraan rakyat. Dengan adanya pengelompokkan ini, diharapkan pemerintah dapat menerapkan kebijakan yang tepat untuk terciptanya pemerataan kesejahteraan. Selain itu, penelitian ini juga bertujuan untuk membandingkan beberapa metode *clustering* yang dapat digunakan untuk mengelompokkan wilayah berdasarkan tingkat kesejahteraan sehingga didapat metode *clustering* terbaik untuk mengelompokkan kabupaten/kota di pulau Jawa.

2. Pendahuluan

2.1 Bahan dan Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari Badan Pusat Statistik. Penelitian ini menggunakan indikator kesejahteraan rakyat yang didapat dari publikasi indikator kesejahteraan rakyat tahun 2018 dari enam provinsi yang ada di pulau Jawa. Berdasarkan literatur rujukan, maka diputuskan untuk menggunakan sebelas variabel (Tabel 1) yang mempengaruhi tingkat kesejahteraan masyarakat di suatu daerah.

Tabel 1: Variabel Penelitian

No	Variabel	Tipe
1.	Angka Harapan Hidup	Kuantitatif
2.	Angka Harapan Lama Sekolah	Kuantitatif
3.	Rata-rata Lama Sekolah	Kuantitatif
4.	Daya Beli (didekati dengan pengeluaran per kapita yang disesuaikan)	Kuantitatif
5.	Angka Melek Huruf	Kuantitatif
6.	PDRB Kabupaten/Kota	Kuantitatif
7.	Jumlah Angkatan Kerja	Kuantitatif
8.	Tingkat Pengangguran Terbuka	Kuantitatif
9.	Jumlah Penduduk Miskin	Kuantitatif
10.	Kepadatan Penduduk	Kuantitatif
11.	Kepemilikan Rumah Sendiri	Kuantitatif

2.2 Metode Penelitian

Analisis data yang digunakan untuk pengelompokkan kabupaten/kota di pulau Jawa adalah beberapa analisis *clustering* yang terdiri dari *hard clustering* (*K-Means*, *K-Medoids*, dan *Hierarchical Agglomerative*) dan *soft clustering* (*Fuzzy C Means*). Sebelum melakukan clustering, dilakukan pembentukan komponen utama untuk mereduksi variabel yang digunakan.

a. Analisis Komponen Utama

Principal Component Analysis (PCA) adalah analisis yang berguna untuk reduksi variabel dan sering kali digunakan dalam analisis multivariat. Tujuan dari pembuatan komponen utama ini adalah untuk mereduksi data yang berdimensi tinggi menjadi lebih sedikit dengan tetap meminimalisasi resiko kehilangan informasi. Maka untuk meningkatkan efisiensi, pada penelitian ini akan digunakan metode PCA terhadap dataset asli, sehingga variabel yang berkorelasi akan diubah menjadi komponen utama yang saling independen. Sebelum melakukan PCA, dataset harus dinormalisasi, sehingga skala dari semua variabel sama dan tidak terjadi dominasi antar variabel. Analisis ini juga berperan mengurangi keberadaan *outlier* dan melepaskan asumsi multikolinieritas (Clayman et al., 2020; Muntaner et al., 2012). Pembentukan komponen didasarkan pada dua cara yaitu matriks korelasi dan matriks kovarian (Johnson & Wichern, 2002). Proses menentukan komponen utama dengan menggunakan matriks korelasi adalah sebagai berikut:

1. Membuat matriks $Z_{(n \times p)}$ yang merupakan hasil standarisasi variabel X.
2. Membuat matriks korelasi dari Z (Z'Z). Hal pertama yang harus dilakukan pada proses reduksi komponen adalah mencari nilai *eigen* $(\lambda_1, \lambda_2, \dots, \lambda_p)$ dari persamaan:

$$\left|Z'_{(p\times n)} Z_{(n\times p)} - \lambda_{(p\times p)} I_{p\times p)}\right| = 0$$

Keterangan :

Z : matriks hasil standarisasi variabel X.

λ : matriks *eigen* yang merupakan matriks diagonal terssusun oleh *eigen value.*

I : matriks identitas berukuran n x p.

Jumlah komponen utama dipilih berdasarkan *eigen value* (λ) dimana jika λ > 1 maka komponen tersebut akan dipilih sebagai komponen utama (Supranto, 2010). Setelah asumsi terpenuhi selanjutnya dilakukan analisis *cluster*.

b. Analisis *Cluster*

Analisis *cluster* merupakan pengelompokan beberapa objek menjadi beberapa kelompok dimana setiap kelompok yang terbentuk terdiri dari objek yang mirip di dalam kelompok (Supranto, 2010). Dalam *clustering,* digunakan ukuran yang menjelaskan kemiripan antar objek (data) untuk menerangkan struktur kelompok yang lebih sederhana dan berasal dari data yang lebih kompleks, yaitu ukuran similaritas dengan menggunakan jarak *Euclidean* (Johnson et al., 2002). Analisis *cluster* dibagi menjadi beberapa metode yaitu *hard clustering* yang terdiri dari metode hirarki dan non hirarki, serta *soft clustering* dengan menggunakan *Fuzzy C Means*.

1. Analisis *Hierarchical Clustering*

Hierarchical clustering adalah analisis *clustering* dengan pengelompokan yang mereduksi kelompok secara sistematis dari n kelompok, $(n-1)$ hingga satu kelompok (metode *Divisive*) atau sebaliknya secara berjenjang dari satu kelompok, lalu dua kelompok hingga *n* kelompok (metode *Agglomerative*). Terdapat beberapa teknik metode *Agglomerative* diantaranya metode *Single Linkage, Complete Linkage, Average Linkage*, dan *Ward Linkage* (Johnson et al., 2002).

Single Linkage

Single Linkage adalah teknik *clustering* hierarki dengan menggabungkan objek pengamatan yang memiliki kesamaan terdekat. Metode ini sangat baik untuk melihat *cluster* bentuk *non-elliptical,* tapi sangat sensitif terhadap *outlier* (Govender & Sivakumar, 2020). Jika terdapat matriks jarak D = $\{d_{ij}\}$ dan objek koresponden adalah U, untuk membentuk *cluster* (UV), maka jarak antara (UV) dengan *cluster* lain misalnya W (Johnson et al., 2002) dapat dirumuskan dengan :

$$d_{(UV)W} = \min\{d_{UW}, d_{VW}\}$$

Keterangan :

d_{UW} : jarak antara objek U dan *cluster* W.

d_{VW} : jarak antara objek V dan *cluster* W.

$d_{(UV)W}$: jarak minimum antara *cluster* UV dan *cluster* W.

Average Linkage

Average linkage adalah teknik *clustering* hierarki yang menggunakan jarak rata-rata antara seluruh pasangan objek dimana salah satu pasangan objek milik masing-masing *cluster*. Jika matriks D = $\{d_{ik}\}$ digunakan untuk menentukan objek. Misalnya objek U dan V akan bergabung ke *cluster* (UV) dengan jarak antara *cluster* UV dan *cluster* W (Johnson et al., 2002) adalah :

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}$$

Keterangan :

$N_{(UV)}$: jumlah objek dalam *cluster* (UV).

N_W : jumlah objek dalam *cluster* (W).

Complete Linkage

Metode *complete linkage* merupakan metode *clustering* hierarki yang mengelompokkan variabel berdasarkan jarak/kesamaan terjauh (berlawanan dengan *single linkage*), sehingga dua variabel yang memiliki kemiripan terkecil akan ditempatkan pada kelompok pertama dan seterusnya (Johnson et al., 2002). Metode ini menghasilkan *cluster* yang lebih padat dan lebih robust dibanding Single Linkage (Govender & Sivakumar, 2020).

$$d_{(UV)W} = \max\{d_{UW}, d_{VW}\}$$

Ward Linkage

Dalam metode *Ward Linkage*, cluster yang terbentuk memiliki varians internal sekecil mungkin karena jarak antara dua cluster didasarkan pada *sum of square* dari kedua *cluster* yang terbentuk. Metode ini juga efektif sehingga sering digunakan dibanding metode hierarki lainnya (Govender & Sivakumar, 2020). Proses *clustering* didasarkan pada varians minimum dalam cluster (Johnson et al., 2002). Metode ini biasa digunakan untuk objek dengan jumlah *cluster* kecil. Proses dalam *ward linkage clustering* adalah sebagai berikut :

1) Menghitung *sum square error* antara dua cluster dengan rumus :

$$SSE_{ij} = \frac{1}{2}\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2$$

Keterangan :

SSE_{ij} : *Sum Squares Error* antar pasangan objek i dan j.

x_{ik} : nilai objek i pada variabel ke-k.

x_{jk} : nilai objek j pada variabel ke-k.

2) Mencari nilai SSE terkecil antara kedua cluster lalu menggabungkan menjadi satu cluster. Sehingga dari sebanyak n klaster secara sistematis akan berkurang menjadi n-1 dan seterusnya.
3) Mengulangi langkah (2) sampai diperoleh jumlah klaster minimum.

Pemilihan Metode Hierarki Terbaik

Pemilihan metode hierarki pada penelitian ini menggunakan koefisien *agglomerative* sebagai ukuran struktur dari *cluster* hierarki. Nilai koefisien *agglomerative* yang mendekati 1 menunjukkan struktur yang lebih seimbang dan kuat (Kaufman & Rousseeuw, 2009). Nilai yang mendekati 0 menunjukkan *cluste*r yang terbentuk lebih buruk. Namun, nilai koefisien *agglomerative* cenderung meningkat seiring bertambahnya jumlah sampel, sehingga koefisien ini tidak bisa digunakan untuk membandingkan data dengan ukuran yang jauh berbeda.

2. Analisis *Non Hierarchical Clustering*

Analisis *clustering* non hierarki adalah metode pengelompokan objek ke dalam kelompok sejumlah k yang telah ditentukan sebelum melakukan analisis. Pada penelitian kali ini menggunakan analisis *K-means clustering* dan *K-medoids clustering* untuk melihat kesesuaian metode dengan data yang digunakan.

K-Means Clustering

K-Means clustering adalah metode *clustering* non hierarki yang paling sederhana dan membagi data ke dalam beberapa kelompok dimana data dengan kesamaan karakteristik dimasukkan ke dalam satu kelompok dan data dengan karakteristik yang lebih berbeda dikelompokkan ke dalam *cluster* lain (Johnson et al., 2002). Proses dalam *K-Means clustering* adalah sebagai berikut :

1. Menentukan jumlah *cluster* (k).
2. Inisiasi secara *random* nilai *centroid* (pusat *cluster*) sejumlah k.
3. Menghitung jarak setiap data terhadap masing–masing *centroid cluster* dengan menggunakan jarak *Euclidean*. Jarak *euclidean* antara objek i dan j dirumuskan dengan :

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

Keterangan :

x_{ik} : nilai objek ke-*i* pada variabel ke-*k.*

x_{jk} : nilai objek ke-*j* pada variabel ke-*k.*

p : banyak variabel yang diamati.

4. Mengelompokkan data berdasarkan jarak terdekat antara setiap data terhadap *centroid*.
5. Menentukan *centroid* baru dengan menghitung rata-rata seluruh data pada pusat *cluster* yang sama.

$$C_{kj} = \frac{x_{1j}+x_{2j}+\cdots+x_{aj}}{a}, j = 1,2,\ldots,p$$

Keterangan :

C_{kj} : Pusat *cluster* ke- *k* variabel ke- *j.*

a : banyak data pada *cluster* ke- *k*.

6. Kembali ke tahap 3 dan terus melakukan iterasi hingga *centroid cluster* tetap dan anggota *cluster* tidak berpindah (konvergen).

K-Medoids Clustering

Metode *K-Medoids* adalah pengembangan dari metode *K-Means*. Seperti yang diketahui, ukuran *mean* bersifat sangat rentan terhadap keberadaan *outlier*. Nilai ekstrim pada *outlier* dapat menggeser rata-rata sehingga distribusinya menjadi tidak normal. Menurut Kaufman & Rousseeuw (1990) metode *K-means* sensitif terhadap data yang memiliki *outlier* akibat penggunaan *mean* sebagai ukuran pemusatannya. Tidak seperti metode *K-Means* yang sensitif terhadap adanya pencilan, algoritma *K-Medoids* dapat mengatasi kelemahan tersebut (Arora et al., 2016) karena penggunaan jarak *Manhattan* yang lebih robust dibanding jarak *Euclidian* (Gupta & Panda, 2018). Kedua metode ini menghasilkan sebanyak k *cluster* yang dibentuk dengan mengukur jarak setiap objek dengan titik pusat, lalu objek dikelompokkan dalam satu *cluster* berdasarkan titik pusat terdekat.

Prinsip dasar algoritma *K-Medoids clustering* adalah menentukan k *cluster* dari n objek dengan menginisiasi objek random yang dianggap representatif untuk setiap *cluster*. Titik pusat setiap *cluster* pada metode ini disebut dengan medoid. Klaster disusun berdasarkan ukuran *similarity* antara *medoid* dengan objek selain *medoid* menggunakan jarak *Manhattan*. Jarak *Manhattan* antara objek i dan objek j dirumuskan dengan :

$$d_{ij} = \sum_{k=1}^{p}|x_{ik} - x_{jk}|$$

Proses di dalam algoritma *K-Medoids* (Gupta & Panda, 2018) yaitu :

1. Inisiasi secara *random* k objek dari n objek sebagai *medoid*. Misal medoid dinyatakan dengan m_i.
2. Menghitung jarak antara setiap objek terhadap *medoid* pada *cluster* terdekat dan menempatkan setiap objek ke *cluster* dengan *medoid* terdekat.
3. Inisiasi secara *random* objek selain *medoid* sebagai o_i.
4. Menghitung total biaya (*cost*) dari pertukaran *medoid* O_i dan O_{random}. Dimana
$$\text{Total cost} = \sum \mathrm{d_{ij}}$$
5. Jika S < 0 , maka tukar m_i dengan o_i untuk dijadikan sebagai medoid baru. Hal ini dilakukan secara iteratif sampai S bernilai konstan (0).
Dimana $S = \text{Total cost}_{\text{baru}} - \text{Total cost}_{\text{lama}}$

Algoritma yang digunakan dalam *K-medoids clustering* adalah PAM (*Partitioning Around Medoid*) dan CLARA (*Clustering Large Application*). Dimana kedua algoritma ini memiliki prinsip yang mirip. CLARA merupakan pengembangan dari PAM dengan mengandalkan proses pengambilan sampel. CLARA mengambil sampel dari *dataset*, menerapkan PAM dalam sampel dan mencari *medoid* dari sampel. Untuk pendekatan lebih baik, CLARA mengambil banyak sampel dan memberikan *output* terbaik (Gupta & Panda, 2018).

3. Fuzzy C-Means Clustering

Fuzzy C-Means menghubungan derajat keanggotaan dari suatu objek terhadap jarak antara pusat kelompok terhadap objek tersebut. Kelemahan FCM adalah sensitif terhadap *noise* dan mudah terjebak pada lokal optimum (Izakian & Abraham, 2011). Keunggulan dari *Fuzzy C-Means* adalah metode ini *robust* dalam meminimumkan fungsi objektif (Hadi, 2017; Izakian & Abraham, 2011). Fungsi objektif pada *Fuzzy C-Means* dirumuskan sebagai berikut.

$$J_m(\widetilde{U}, v) = \sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{ik})^m (d_{ik})^2$$

$$d_{ik} = d(x_k - v_i) = \left[\sum_{j=1}^{m} (x_{kj} - v_{ij})^{\frac{1}{2}} \right]$$

Keterangan

μ_{ik} : nilai keanggotaan data ke- k *cluster* ke-*i.*

d_{ik} : jarak antara titik data ke-k dengan pusat *cluster* ke-*i.*

x_k : data ke-*k.*

v_i : pusat *cluster* ke-*i.*

m : *fuzziness*, parameter ukuran kesamaran hasil *clustering* (m > 1).

Fuzzy C-Means dapat dievaluasi dengan beberapa ukuran yaitu *Partition Coefficient*, *Classification Entropy Index*, *Xie and Beni's Index*, *Separation Index*, dan *Dunn Index* (Wijayanto & Takdir, 2014).

Partition Coefficient

Koefisien ini digunakan untuk mengukur jumlah *overlapping* antar *cluster* (Grekousis & Thomas, 2012; Wijayanto & Takdir, 2014). μ_{ij} adalah derajat keanggotaan titik data ke-j di dalam kelompok ke-i. Model terbaik dinyatakan dengan nilai PC maksimal. Persamaannya adalah sebagai berikut :

$$PC = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij}^2$$

Classification Entropy Index

Indeks ini mengukur kesamaran dari partisi dalam *cluster* (Grekousis & Thomas, 2012; Wijayanto & Takdir, 2014). Model terbaik adalah model yang memiliki nilai CE minimum. Persamaannya dirumuskan sebagai berikut :

$$CE = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log_a(\mu_{ij})$$

Xie and Beni's Index

Indeks ini mengukur rasio total varians *within cluster* terhadap *separation* pada *cluster*. Model terbaik adalah model yang memiliki nilai XB minimum. Indeks ini dirumuskan sebagai berikut (Grekousis & Thomas, 2012) :

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^{m}\|x_j - v_i\|^2}{N\min\limits_{i,j}\|x_j - v_i\|^2}$$

Separation Index

Indeks ini mengukur pemisahan jarak minimum sebagai validitas partisi (Wijayanto & Takdir, 2014). Model terbaik dinyatakan dengan nilai indeks S yang minimum. Persamaannya dirumuskan sebagai berikut :

$$SI = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^{2}\|x_j - v_i\|^2}{N\min\limits_{i,k}\|v_k - v_i\|^2}$$

4. Penentuan jumlah *cluster*

Ada beberapa metode yang dapat digunakan untuk menentukan jumlah *cluster* optimal pada proses *clustering*. Dalam penelitian ini, metode yang akan digunakan adalah *silhouette method*, *elbow method* dan *gap statistic* (Clayman et al., 2020). Semua metode ini akan menampilkan *plot* untuk menentukan jumlah *cluster* optimal.

Silhouette Method

Metode *Silhouette* digunakan untuk memilih jumlah *cluster* optimal dengan menggunakan data skala rasio. Ketika diterapkan, algoritma *silhouette* akan mengukur jarak rata-rata dari suatu objek terhadap seluruh objek yang terdapat pada *cluster* yang sama dengan objek di *cluster* lainnya. Nilai *Silhoutte* yang mendekati 1 menunjukkan jumlah *cluster* yang optimal.

Elbow Method

Metode *Elbow* digunakan untuk memilih jumlah *cluster* berdasarkan siku yang terbentuk pada suatu titik di grafik SSE dan didasarkan pada penurunan SSE yang besar. Jika nilai *cluster* sebelumnya (k-1) dengan nilai *cluster* selanjutnya (k) mengalami penurunan terbesar maka jumlah *cluster* tersebut yang tepat (k). Metode ini menggunakan nilai *Sum of Square Error* (SSE) dari masing-masing jumlah *cluster*. Semakin besar jumlah *cluster*, maka SSE akan terus mengecil, sehingga jumlah *cluster* terbaik adalah jumlah *cluster* yang mengalami penurunan terbesar. Rumus SSE dapat dituliskan sebagai :

$$SSE = \sum_{K=1}^{K}\sum_{X_i}|x_i - c_k|^2$$

Keterangan:

K : *cluster.*

x_i : data ke-*i.*

c_k : pusat *cluster* ke-*k.*

***Gap* Statistik**

Gap statistik merupakan ukuran yang paling konstan untuk menentukan jumlah *cluster* jika dibandingkan ukuran yang lain (Silvi, 2018). Jarak antara objek berpasangan di dalam *cluster* dirumuskan sebagai :

$$D_r = \sum_{i,i' \epsilon C_r} d_{ii'}$$

Dimana d adalah kuadrat dari jarak *euclidean*. Jumlah kuadrat di dalam *cluster* dirumuskan sebagai berikut.

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

Nilai gap merupakan hasil estimasi jumlah *cluster* optimum dengan menggunakan pendekatan standarisasi W_k. Dimana $En*$ adalah ekspektasi dari distribusi jumlah sampel. Kriteria jumlah *cluster* optimal merupakan jumlah cluster yang memiliki nilai *gap* statistik tertinggi atau jika nilai *gap* selalu naik maka jumlah *cluster* optimum adalah nilai yang mengindikasi kenaikan gap minimum (Silvi, 2018).

5. Evaluasi *Cluster*

Evaluasi pada *cluster* terdiri dari evaluasi eksternal dan internal. Dimana evaluasi eksternal dapat dilakukan saat *cluster* tersebut memiliki label (*supervised*). Penelitian ini hanya menggunakan evaluasi internal karena data yang digunakan tidak berlabel. Evaluasi internal menggunakan informasi internal pada data untuk menilai hasil *clustering*. Evaluasi internal mencerminkan kepadatan, hubungan dan pemisahan partisi *cluster*. Evaluasi internal yang akan digunakan adalah *Connectivity coefficient*, *Silhouette coefficient*, dan *Dunn Index*. Selain evaluasi ini, digunakan juga rasio rata-rata simpangan baku *within cluster* terhadap simpangan baku *between cluster*.

Internal : *Silhouette Coefficient*

Silhouette coefficient merupakan ukuran derajat kepercayaan dalam pengelompokkan suatu pengamatan dengan *cluster. Cluster* yang terbentuk akan dikategorikan baik jika koefisien yang dihasilkan mendekati 1 dan sebaliknya jika koefisien mendekati angka -1. *Silhouette coefficient* dihitung dengan cara :

- Menghitung rata-rata jarak antara sebuah objek (*i*) terhadap setiap objek pada *cluster* yang sama (kohesi).

  $$a(i) = \frac{1}{|A|-1} \sum_{j \in A, j \neq i} d(i,j)$$

  Dimana *j* adalah objek selain *i* dalam *cluster* yang sama yaitu *A* dan |A| adalah banyaknya anggota cluster *A*.
- Menghitung rata-rata jarak antara objek *i* dengan setiap objek pada *cluster* lainnya lalu ambil jarak terkecil (*cluster* tetangga terdekat).

  $$d(i,C) = \frac{1}{|A|} \sum_{j \in C} d(i,j))$$

  Dimana d(i,C) adalah jarak rata-rata antara objek *i* terhadap setiap objek pada *cluster* lain (*C*) dimana $A \neq C$.
- Selanjutnya menghitung separation yang merupakan jarak dengan cluster tetangga terdekat.

  $$b(i) = \min_{C \neq A} d(i,j)$$
- Hasil koefisien *silhouette* dirumuskan dengan :

  $$s(i) = \frac{b(i) - a(i0}{\max(a(i), b(i))}$$

Internal : *Dunn Index*

Dunn Index merupakan ukuran validasi hasil *clustering* yang didapat dengan mengukur jarak antara dua *cluster* dan diameter *cluster*. *Dunn Index* merupakan rasio dari jarak terbesar antara dua *cluster* terhadap jarak terkecil di dalam suatu *cluster. Cluster* yang terbentuk akan semakin baik saat nilai *Dunn index* semakin tinggi (Brock et al., 2011). *Dunn Index* dirumuskan sebagai :

$$DI = \min_{i=1,..,k}\{ \min_{j=i+1,...,k} \left(\frac{diss(c_i,c_j)}{\max_{m=1,...,k}(diam(c_m))} \right)$$

Keterangan :

DI : *Dunn Index.*

(c_i , c_j) : jarak antara *cluster i* dan *cluster j.*

(c_m) : diameter *cluster i.*

Internal : *Connectivity Coefficient*

Kepadatan berhubungan dengan mengevaluasi homogenitas dari *cluster*, biasa dilihat menggunakan varians intra-*cluster*. Hubungan ini menunjukkan posisi dari data observasi dalam sebuah *cluster*, yang disebut sebagai tetangga terdekat. Nilai kepadatan tersebut diukur dengan koefisien konektivitas. Nilai *connectivity* berada di antara nilai nol sampai ∞. *Cluster* yang terbentuk akan semakin baik saat nilai koefisien *connectivity* semakin rendah (Brock et al., 2011). *Connectivity coefficient* dirumuskan sebagai berikut :

$$CC = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}}$$

$nn_{i(j)}$: tetangga terdekat objek *j* dari objek di *i*.

$x_{i,nn_{i(j)}}$: mendekati 0 bila objek *i* dan *j* dalam satu *cluster* dan 1 bila sebaliknya.

L : parameter ukuran jumlah tetangga.

Rasio Rata-rata Simpangan Baku *Within Cluster* terhadap Simpangan Baku *Between Cluster*.

Metode *clustering* terbaik juga dapat ditentukan dengan menghitung rasio dari rata-rata simpangan baku *within* terhadap simpangan baku *between* untuk melihat homogenitas pada *cluster* terbentuk (Silvi, 2018). Rata-rata simpangan baku di dalam *cluster* (S_w) dan antar *cluster* (S_b) dirumuskan dengan :

$$S_w = \frac{1}{c} \sum_{k=1}^{c} S_k$$

$$S_b = \left[\frac{1}{c-1} \sum_{k=1}^{c} (\bar{X}_k - \bar{X})^2 \right]^{\frac{1}{2}}$$

Dimana c adalah jumlah *cluster* dan k adalah *cluster* yang akan dihitung. Metode *clustering* dapat dikatakan baik jika nilai S_w semakin kecil dan nilai S_b semakin besar. Metode terpilih memiliki rasio S_w/S_b terkecil sehingga metode tersebut memiliki homogenitas yang tinggi di dalam *cluster*.

3. Hasil dan Pembahasan

3.1 Pembentukan Komponen dengan PCA

Pembentukan komponen dengan PCA akan menghasilkan beberapa komponen yang merupakan kombinasi linear dari varabel-variabel yang digunakan dalam analisis. Komponen yang dapat dilanjutkan dalam proses analisis adalah komponen yang memiliki nilai *eigen value* > 1.

Pada penelitian kali ini, terbentuk tiga komponen dengan *eigen value* lebih dari 1 (Tabel 2). Kemudian analisis dilanjutkan dengan pemodelan *clustering* berbagai metode dan menggunakan hasil komponen utama sebagai variabel yang akan

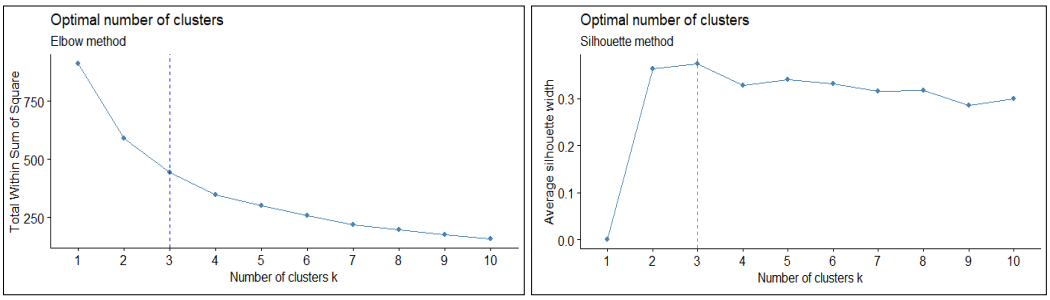
dianalisis.

Tabel 2: Hasil Komponen PCA

Komponen	*Eigen value*
Komponen 1	1.7976787
Komponen 2	1.4722719
Komponen 3	1.1421801

3.2 Analisis *Clustering* dengan metode Hierarki.

Tahap pertama adalah menentukan jumlah *cluster* (k) optimal yang akan digunakan untuk *clustering* data dengan metode hierarki *agglomerative*. Penelitian ini menggunakan *Elbow method* dan *Silhouette method* untuk menentukan jumlah k optimal. Sehingga didapatkan *plot* seperti disajikan pada Gambar 1.



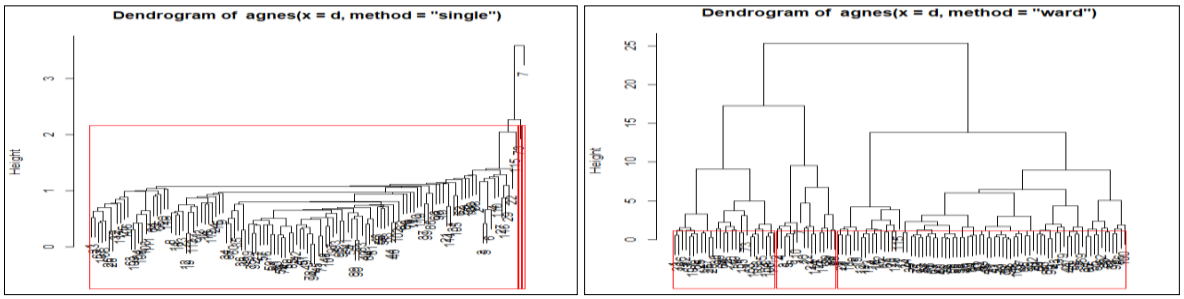
Gambar 1. *Elbow method* dan *Silhouette Method* untuk Hierarki *Agglomerative*

Berdasarkan *plot* dari *Elbow Method,* dapat dilihat bahwa patahan gradient terbesar *terjadi* saat jumlah kaster sebesar tiga. *Silhouette method* juga memberikan hasil yang sama. Sehingga selanjutnya dapat dilakukan perbandingan metode hierarki terbaik antara metode *Single Linkage*, *Complete Linkage*, *Average Linkage*, dan *Ward Linkage* dengan beberapa ukuran evaluasi (*Silhouette Coefficient*, *Connectivity Coefficient*, *Dunn Index*, dan *Agglomerative Coefficient*).

Tabel 3: Evaluasi Metode Hierarki

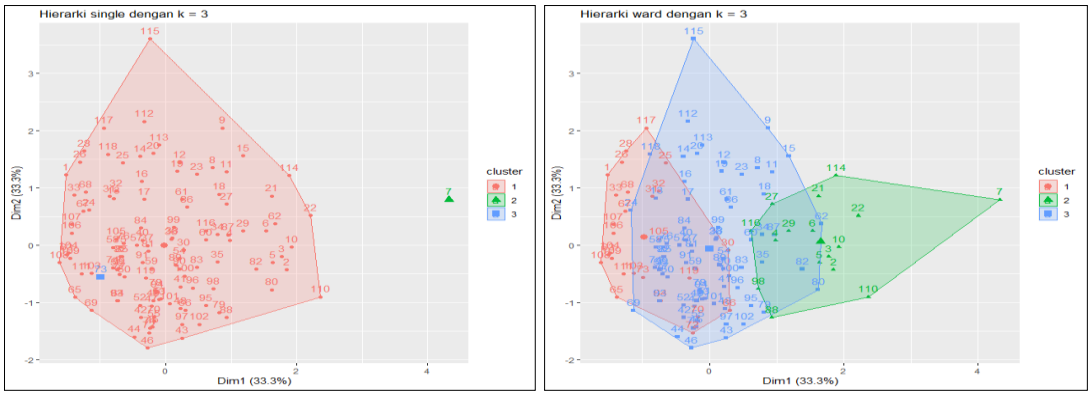
Evaluasi	Ward	Single	Complete	Average
SC	0.3741	0.3078	0.3689	0.3737
CC	22.6563	6.0579	17.731	16.168
DI	0.1386	0.2234	0.1399	0.1450
AC	0.9687	0.8116	0.9182	0.8828

Dapat dilihat dari Tabel 3, Metode *Ward Linkage* unggul pada *Silhouette Coefficient* dan *Agglomerative Coefficient*. Sehingga dapat dikatakan bahwa metode *Ward Linkage* adalah metode *clustering* yang memiliki struktur paling seimbang dan kuat dibanding metode lainnya. Sedangkan menurut ukuran *Dunn Index* dan *Connectivity Coefficient*, metode terbaik adalah *Single Linkage*. Sehingga perlu dilihat bentuk *cluster* yang dihasilkan untuk mendapatkan model terbaik.



Gambar 2. *Dendogram Aglomerrative* Metode *Single Linkage* dan *Ward Linkage*

Dapat dilihat bahwa struktur *dendogram* seperti yang disajikan pada Gambar 2, metode *Single Linkage* memang tidak stabil seperti metode *Ward Linkage*. Sehingga untuk memperjelas metode terbaik, dilakukan visualisasi hasil *cluster* yang terbentuk.



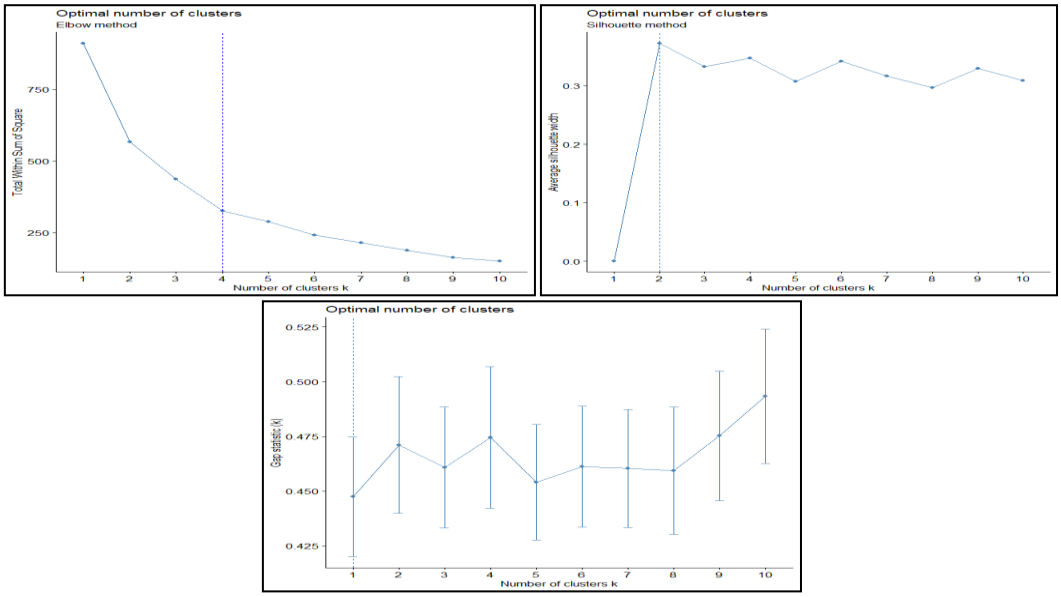
Gambar 3. Clustering *Aglomerrative Single Linkage* dan *Ward Linkage* dengan k = 3

Menurut hasil *clustering*, dapat terlihat lebih jelas bahwa *Agglomerative* metode *Ward Linkage* mengelompokkan wilayah berdasarkan tingkat kesejahteraan dengan lebih baik dibandingkan metode single karena ketiga *cluster* tergambar lebih baik dan jelas batasannya (Gambar 3). Oleh karena itu dapat disimpulkan bahwa metode hierarki *Agglomerative* terbaik yang dapat mengelompokkan kesejahteraan berdasarkan kabupaten/kota di pulau Jawa adalah Metode Hierarki *Agglomerative Ward Linkage*.

3.3 *Analisis Clustering* dengan metode *K-Means*.

Sama seperti analisis *clustering* dengan metode hierarki, langkah pertama yang harus dilakukan adalah menentukan nilai k yang tepat untuk proses *clustering*. Pada *clustering* ini, peneliti menggunakan *elbow method*, *silhouette method* dan *gap statistic* untuk menentukan nilai k optimal.

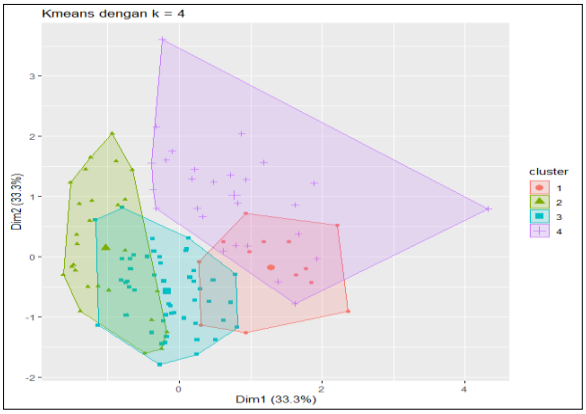
Berdasarkan plot dari *Elbow method* (Gambar 4), dapat dilihat bahwa patahan gradien terbesar saat jumlah *cluster* sebesar empat. *Silhouette method* menunjukkan bahwa jumlah *cluster* optimal adalah dua, tetapi jumlah ini terlalu sedikit sehingga digunakan nilai terbesar setelahnya yaitu empat *cluster*. Berdasarkan *plot* dari *gap statistic* ditunjukkan bahwa jumlah *cluster* optimal sebesar satu, namun jumlah ini tidak sesuai tujuan penelitian sehingga nilai terbesar setelahnya yaitu empat *cluster* dipilih (Gambar 5).



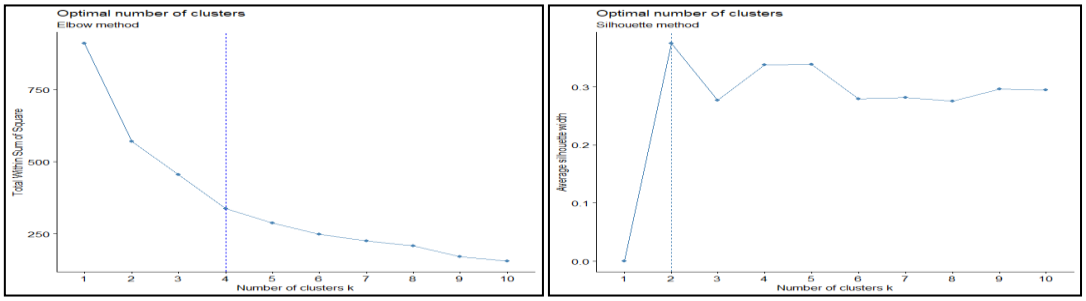
Gambar 4. *Elbow*, *Silhouette*, dan *Gap Statistic Method* untuk *K-Means*

3.4 Analisis *Clustering* dengan metode *K-Medoids*.

Algoritma *K-Medoids* yang digunakan adalah PAM dan CLARA. Seperti metode sebelumnya, hal pertama yang dilakukan adalah menentukan jumlah *cluster* optimal dengan metode *Elbow* dan *Silhouette* untuk algoritma PAM.

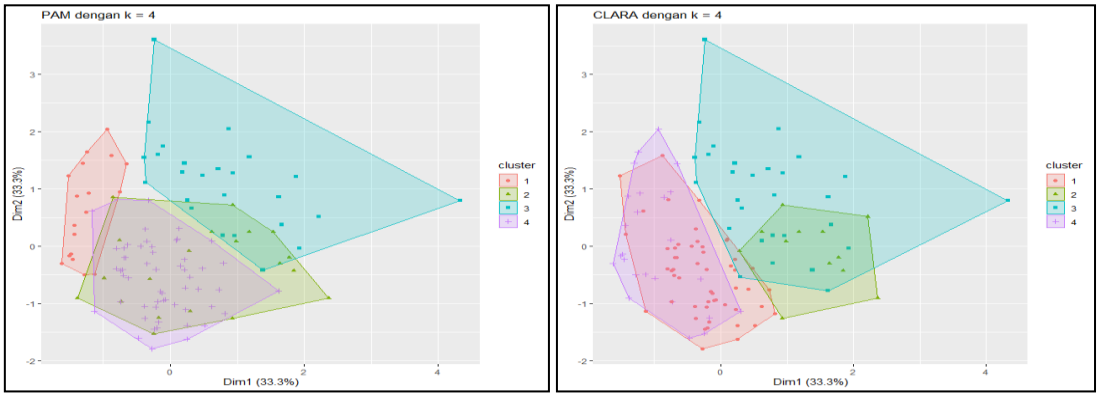


Gambar 5. *Clustering K-Means* dengan k = 4.



Gambar 6. *Elbow Method* dan *Silhouette Method* untuk PAM.

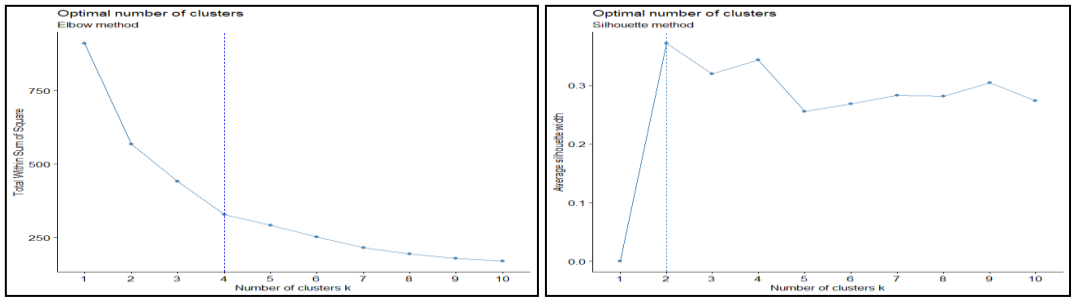
Plot dari *Elbow Method* (Gambar 6) menunjukkan bahwa patahan gradien terbesar terjadi saat jumlah *cluster* sebesar empat. Menurut metode *silhouette* didapatkan bahwa jumlah *cluster* terbaik adalah dua, namun jumlah ini terlalu sedikit sehingga jumlah *cluster* yang digunakan adalah nilai terbesar selanjutnya yaitu saat jumlah *cluster* sebesar empat. Jumlah *cluster* optimal untuk algoritma CLARA juga menunjukkan hasil yang sama dengan visualisasi yang sama persis seperti algoritma PAM sehingga tidak ditampilkan. Selanjutnya dilakukan proses *clustering* untuk melihat hasil *cluster* yang terbentuk dengan *K-Medoids* algoritma PAM dan CLARA (Gambar 7).



Gambar 7. *Clustering K-Medoids* algoritma PAM dan CLARA dengan k = 4.

3.5 Analisis *Clustering* dengan metode *Fuzzy C-Means*

Langkah pertama yang harus dilakukan dalam analisis *clustering* pada umumnya adalah menentukan nilai k. Pada metode *Fuzzy C-Means*, sama seperti metode–metode sebelumnya akan digunakan *elbow method* dan *silhouette* untuk menentukan jumlah *cluster* optimal.



Gambar 8. *Elbow Method* dan *Silhouette Method* untuk *Fuzzy C-Means*

Berdasarkan *output* dari *plot elbow method* (Gambar 8), dapat dilihat bahwa patahan gradien terbesar terjadi saat jumlah *cluster* empat sehingga jumlah *cluster* terbaik menurut elbow *method* adalah empat. Sementara *silhouette method* menunjukkan bahwa jumlah *cluster* terbaik adalah dua, namun seperti yang dijelaskan sebelumnya jumlah *cluster* ini terlalu sedikit sehingga dipilih jumlah *cluster* dengan nilai *silhouette* terbesar setelahnya yaitu empat. Menentukan jumlah *cluster* optimal juga dapat dilakukan dengan menggunakan ukuran-ukuran evaluasi untuk jumlah *cluster* terbaik. Ukuran yang digunakan adalah *Partition Coefficient*, *Classification Entropy*, *Xie and Beni Index*, *Separation*, *Silhouette Coefficient* dan *Dunn Index*.

Berdasarkan ukuran evaluasi seperti disajikan pada Tabel 4, dapat dilihat bahwa jumlah *cluster* optimal pada metode *Fuzzy C Means* adalah tiga *cluster*. Jika kembali melihat plot *Silhouette*, jumlah cluster tiga adalah nilai ketiga tertinggi setelah dua dan empat. Sehingga jumlah *cluster* yang digunakan untuk metode *Fuzzy C Means* adalah tiga.

Tabel 4: Perbandingan Validitas FCM untuk k = 3 sampai k= 6 dengan m = 1.5

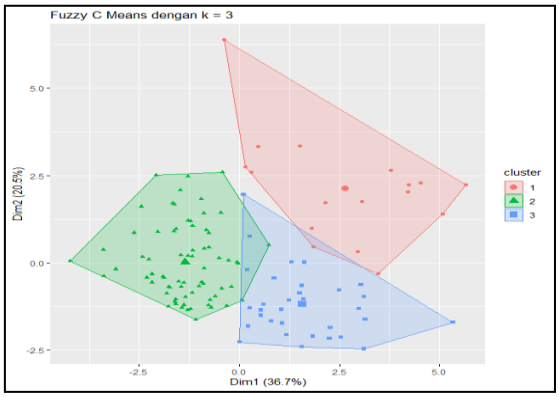
	k			
	3	4	5	6
PC	0.810	0.780	0.719	0.720
CE	0.353	0.422	0.550	0.564
XB	0.315	0.376	0.672	0.577
S	0.255	0.289	0.490	0.411
DI	0.082	0.062	0.056	0.072

Berbeda dengan metode lain yang langsung dilanjutkan dengan proses clustering. Pada penelitian ini, digunakan *threshold* sebesar 10^{-6}. Pada *Fuzzy C-Means* ini terdapat parameter m yang harus ditentukan. Sehingga akan digunakan beberapa nilai indeks untuk mengevaluasi m terbaik untuk model *Fuzzy C-Means*.

Tabel 5: Perbandingan Validitas FCM untuk m = 1.5 sampai m = 3.5 dengan k = 3

	k				
	1.5	2	2.5	3	3.5
PC	0.810	0.565	0.453	0.398	0.370
CE	0.353	0.755	0.924	1.004	1.045
XB	0.315	0.453	0.369	0.306	0.267
S	0.255	0.453	0.623	0.919	1.421
DI	0.082	0.062	0.062	0.051	0.051

Menurut Tabel 5, dapat dilihat bahwa semakin tinggi nilai parameter m, ukuran PC, CE, S, dan *Dunn index* memburuk. Walaupun pada XB *index* model terbaik yang diidentifikasi adalah model *Fuzzy C Means* dengan m = 3,5. Tapi secara keseluruhan dapat disimpulkan bahwa model terbaik adalah *Fuzzy C Means* dengan m = 1,5 (Gambar 9), sehingga model ini akan dilanjutkan untuk membentuk *cluster*.



Gambar 9: Hasil *Clustering* Fuzzy *C Means* dengan k = 3 dan m = 1.5

3.6 Hasil Evaluasi Seluruh Metode

Selanjutnya, evaluasi semua metode yang digunakan untuk mengelompokkan tingkat kesejahteraan rakyat berdasarkan wilayah kabupaten/kota di pulau Jawa. Ukuran evaluasi yang akan digunakan untuk membandingkan semua metode yang diajukan adalah *sillhoute coefficient*, *dunn index*, *connectivity coefficient*, dan rasio simpangan baku *within cluster* terhadap simpangan baku *between cluster*.

Tabel 6: Perbandingan Evaluasi Internal Berdasarkan Metode

Metode	SC	DI	CC	S_w/S_b
Ward Linkage (k = 3)	0.374	0.139	22.656	0.548
K-Means (k=4)	0.348	0.062	43.578	0.549
PAM (k=4)	0.338	0.073	40.963	0.515
CLARA (k=4)	0.325	0.058	51.849	0.524
FCM (m= 1,5 dan k = 3)	0.543	0.082	41.695	0.552

Menurut hasil evaluasi pada Tabel 6, dapat disimpulkan bahwa metode terbaik untuk pengelompokkan tingkat kesejahteraan rakyat berdasarkan wilayah kabupaten/ kota di pulau Jawa adalah metode *Agglomerative Ward Linkage* dengan nilai *Dunn Index* maksimum sebesar 0,139 dan *Connectivity Coefficient* minimum sebesar 22,656. Sedangkan berdasarkan ukuran *Silhouette Coefficient*, metode terbaik yang bisa digunakan adalah metode *Fuzzy C Means* dengan m = 1,5. Rasio varians *within* dan *between* menunjukkan bahwa metode terbaik adalah metode *K-Medoids* dengan algoritma PAM. Sehingga interpretasi selanjutnya akan menggunakan hasil *clustering* dari metode *Agglomerative Ward Linkage*.

3.7 Interpretasi Hasil *Cluster* Metode Terbaik.

Hasil *clustering* dari metode *Agglomerative Ward Linkage* membentuk tiga *cluster* dengan kabupaten/kota yang terkelompok seperti disajikan pada Tabel 7. Berdasarkan *output* Tabel 7 dapat disimpulkan bahwa *cluster* 1 terdiri dari 32 kabupaten/kota dengan tingkat kesejahteraan yang sedang. *Cluster* 2 terdiri dari 17 kabupaten/kota yang memiliki tingkat kesejahteraan tinggi. *Cluster* 3 terdiri dari 70 kabupaten/kota dengan tingkat kesejahteraan rendah (*cluster* terbesar diantara ketiga *cluster* yang terbentuk). Kategori ini disimpulkan berdasarkan komposisi kabupaten/kota yang terkelompok di dalam *cluster*. Untuk memastikan hal ini, maka harus dilihat secara univariat dengan menggunakan ukuran rata-rata (Tabel 8).

Cluster pertama adalah kelompok kabupaten/kota dengan harapan lama sekolah yang tinggi dan penduduk miskin yang rendah, tidak ada indikator lainnya yang dominan di *cluster* ini. Berdasarkan hal ini, dapat dikatakan bahwa *cluster* ini adalah kelompok kabupaten/kota dengan tingkat kesejahteraan masyarakat sedang.

Cluster kedua adalah kelompok kabupaten/kota dengan angka harapan hidup, rata-rata lama sekolah, daya beli, angka melek huruf, PDRB dan angkatan kerja yang tinggi dimana semua ini adalah ciri daerah dengan kesejahteraan tinggi. Namun *cluster* ini juga memiliki tingkat pengangguran terbuka, penduduk miskin, dan kepadatan penduduk yang tinggi (tipikal wilayah perkotaan padat penduduk). Berdasarkan hal ini dapat disimpulkan bahwa *cluster* ini adalah kelompok kabupaten/kota dengan tingkat kesejahteraan masyarakat tinggi.

Tabel 7: Hasil Clustering *Agglomerative Ward Linkage*

***Cluster* 1**		***Cluster* 2**		***Cluster* 3**		
Kepulauan Seribu	Bantul	Kota Jakarta Selatan	Sukabumi	Kebumen	Pekalongan	Pasuruan
Purwakarta	Sleman	Kota Jakarta Timur	Cianjur	Purworejo	Pemalang	Brebes
Kota Bogor	Kota Yogyakarta	Kota Jakarta Pusat	Garut	Wonosobo	Situbondo	Jombang
Kota Sukabumi	Kota Kediri	Kota Jakarta Barat	Tasikmalaya	Magelang	Mojokerto	Nganjuk
Kota Cirebon	Kota Blitar	Kota Jakarta Utara	Ciamis	Boyolali	Gunung Kidul	Madiun
Kota Cimahi	Kota Malang	Bogor	Kuningan	Klaten	Bojonegoro	Magetan
Kota Tasikmalaya	Kota Probolinggo	Bandung	Cirebon	Wonogiri	Ponorogo	Ngawi
Kota Banjar	Kota Pasuruan	Karawang	Majalengka	Sragen	Trenggalek	Pacitan
Sukoharjo	Kota Mojokerto	Bekasi	Sumedang	Grobogan	Tulungagung	Tuban
Karanganyar	Kota Madiun	Kota Bandung	Indramayu	Blora	Pamekasan	Malang
Kudus	Kota Batu	Kota Bekasi	Subang	Rembang	Bangkalan	Gresik
Kota Magelang	Serang	Kota Depok	Bandung Barat	Pati	Lamongan	Kediri
Kota Surakarta	Kota Cilegon	Kota Semarang	Pangandaran	Jepara	Lumajang	Sampang
Kota Salatiga	Kota Serang	Sidoarjo	Cilacap	Demak	Temanggung	Blitar
Kota Pekalongan	Kota Tangerang Selatan	Kota Surabaya	Banyumas	Semarang	Banyuwangi	Sumenep
Kota Tegal		Tangerang	Purbalingga	Jember	Bondowoso	Tegal
Kulonprogo		Kota Tangerang	Banjarnegara	Kendal	Pandeglang	Lebak
				Batang	Probolinggo	

Tabel 8: Rata-rata Variabel Hasil *Clustering Agglomerative Ward Linkage*

Cluster	X1	X2	X3	X4	X5	X6
1	72.84	13.87	9.48	12784.28	97.92	24790.28
2	73.35	13.47	10.04	14387.94	98.34	206614.71
3	71.98	12.52	7.00	9548.44	92.90	27579.43
Jawa	72.41	13.02	8.10	11109.94	95.02	52405.87

Cluster	X7	X8	X9	X10	X11
1	281898.00	5.66	38.65	5253.42	74.02
2	1303440.24	7.18	133.39	10310.07	65.25
3	632154.46	4.74	140.49	843.42	90.25
Jawa	633865.82	5.34	112.09	3249.63	82.31

Cluster ketiga adalah kabupaten/kota dengan indikator kesejahteraan rakyat yang rendah. Walaupun variabel kepemilikan rumah dan tingkat pengangguran terbuka menunjukkan hasil yang baik, variabel rata-rata lama sekolah, angka harapan hidup, harapan lama sekolah, daya beli, angka melek huruf, PDRB, angkatan kerja berada di bawah rata-rata pulau Jawa. Hal ini menunjukkan capaian yang buruk dalam kesejahteraan masyarakat di kabupaten/kota pada kelompok ini. Sehingga dapat

disimpulkan bahwa *cluster* tiga adalah kelompok kabupaten/kota dengan tingkat kesejahteraan masyarakat yang rendah.

Terdapat beberapa kabupaten/kota yang cukup maju dan memiliki industri yang cukup banyak seperti kota Cilegon, kota Yogyakarta, dan kota Pekalongan serta kota yang metropolitan seperti Tangerang Selatan di *cluster* 1. Pada penelitian ini, adanya overlapping cukup berhasil diatasi dengan metode *Fuzzy C Means* tetapi berdasarkan hasil evaluasi model, metode terbaik adalah *Agglomerative Ward Linkage.* Berdasarkan capaian indikator, kota Cilegon memiliki AHH (66.43 tahun), jumlah angkatan kerja (198809), dan persentase rumah milik sendiri (81.88%) di bawah rata-rata pulau Jawa serta indikator TPT (9.33%) jauh di atas rata-rata pulau Jawa. Sedangkan kota Yogyakarta memiliki PDRB (26129 ribu rupiah), jumlah angkatan kerja (239542), dan persentase rumah milik sendiri (39.93%) jauh di bawah rata-rata pulau Jawa serta indikator kepadatan penduduk (13359.31 penduduk per km^2) dan TPT (6.22%) diatas rata-rata pulau Jawa. Kota Pekalongan memiliki capaian harapan lama sekolah, jumlah angkatan kerja, persentase kepemilikan rumah milik sendiri dan PDRB di bawah rata–rata pulau Jawa serta indikator kepadatan penduduk dan TPT di atas rata-rata pulau Jawa. Berbeda dengan tiga kota sebelumnya, capaian buruk kota Tangerang Selatan adalah AHH (72.26) yang lebih rendah dari rata-rata pulau Jawa dan kepadatan penduduk (11525 penduduk per km^2) di atas rata–rata dan mengindikasikan capaian baik pada indikator lainnya tapi terkena efek *overlapping* yang ada pada penelitian ini. Berdasarkan hal ini, wajar jika kota Cilegon, kota Yogyakarta dan kota Pekalongan dikelompokkan pada *cluster* pertama (tingkat kesejahteraan sedang) tetapi kota Tangerang Selatan merupakan pengecualian yang disebabkan adanya *overlapping* pada hasil clustering (kurang *crisp*) ditandai dengan tumpang tindihnya beberapa kota pada cluster terbentuk.

4. Kesimpulan

Metode pengelompokkan yang sesuai untuk indikator tingkat kesejahteraan rakyat kabupaten/kota di Pulau Jawa adalah metode *Agglomerative Ward Linkage*. Hal ini disebabkan metode ini memiliki struktur *cluster* yang kuat dan ukuran evaluasi yang lebih baik dibanding metode *cluster* lainnya (*K-means*, *K-Medoids* algoritma PAM dan CLARA, dan *Fuzzy C Means*). Metode *Agglomerative Ward Linkage* membentuk tiga *cluster* dengan rincian :*Cluster* 1 terdiri dari 32 kabupaten/kota yang memiliki tingkat kesejahteraan sedang dengan angka harapan lama sekolah tinggi dan penduduk miskin rendah, *Cluster* 2 terdiri dari 17 kabupaten/kota yang memiliki tingkat kesejahteraan tinggi dengan rata-rata lama sekolah, daya beli, angka harapan hidup, angka melek huruf, PDRB dan angkatan kerja yang tinggi namun tingkat pengangguran terbuka, penduduk miskin, dan kepadatan penduduk yang tinggi pula, dan *Cluster* 3 terdiri dari 70 kabupaten/kota yang memiliki tingkat kesejahteraan rendah dengan variabel AHH, daya beli, AMH, rata-rata lama sekolah, PDRB, angkatan kerja, dan harapan lama sekolah berada di bawah rata-rata pulau Jawa.

Daftar Pustaka

Alwi, W., & Hasrul, M. (2018). Analisis Klaster Untuk Pengelompokkan Kabupaten/Kota Di Provinsi Sulawesi Selatan Berdasarkan Indikator Kesejahteraan Rakyat. *Jurnal MSA (Matematika Dan Statistika Serta Aplikasinya)*, *6*(1), 35.

Arora, P., Varshney, S., & others. (2016). Analysis of K-Means and K-Medoids algorithm for big data. *Procedia Computer Science*, *78*, 507–512.

BAPPENAS. (2019). *Narasi RPJMN 2020-2024*. BAPPENAS. https://www.bappenas.go.id/files/rpjmn/Narasi%20RPJMN%20IV%202020-2024_Revisi%2018%20Juli%202019.pdf

BPS. (2018). *Indikator Kesejahteraan Rakyat 2018*. BPS. https://www.bps.go.id/publication/2018/11/28/f6adb407ea72d9b66776a270/indikator-kesejahteraan-rakyat-2018.html

Brock, G., Pihur, V., Datta, S., Datta, S., & others. (2011). ClValid, an R package for cluster validation. *Journal of Statistical Software (Brock et al., March 2008)*.

Clayman, C. L., Srinivasan, S. M., & Sangwan, R. S. (2020). K-means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes. *Procedia Computer Science*, *168*, 97–104.

Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, *11*(1), 40–56.

Grekousis, G., & Thomas, H. (2012). Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods. *Applied Geography*, *34*, 125–136.

Gupta, T., & Panda, S. P. (2018). A comparison of k-means clustering algorithm and clara clustering algorithm on iris dataset. *International Journal of Engineering & Technology*, *7*(4), 4766–4768.

Hadi, B. S. (2017). *Pendekatan Modified Particle Swarm Optimization dan Artificial Bee Colony pada Fuzzy Geographically Weighted Clustering (Studi Kasus pada Faktor Stunting Balita di Provinsi Jawa Timur)* [PhD Thesis]. Institut Teknologi Sepuluh Nopember.

Hidayatullah, K. H. (2014). Analisis Klaster Untuk Pengelompokan Kabupaten/Kota di Provinsi Jawa Tengah Berdasarkan Indikator Kesejahteraan Rakyat. *Jurnal Statistika Universitas Muhammadiyah Semarang*, *2*(1).

Izakian, H., & Abraham, A. (2011). Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. *Expert Systems with Applications*, *38*(3), 1835–1838.

Izzuddin, A. (2015). Optimasi Cluster pada Algoritma K-Means dengan Reduksi Dimensi Dataset Menggunakan Principal Component Analysis untuk Pemetaan Kinerja Dosen. *Energy*, *5*(2), 41–46.

Johnson, R. A., Wichern, D. W., & others. (2002). *Applied multivariate statistical analysis* (Vol. 5). Prentice hall Upper Saddle River, NJ.

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding Groups in Data: An Introduction to Cluster Analysis*, *344*, 68–125.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Muntaner, C., Chung, H., Benach, J., & Ng, E. (2012). Hierarchical cluster analysis of labour market regulations and population health: A taxonomy of low-and middle-income countries. *BMC Public Health*, *12*(1), 286.

Rahayu, G., & Mustakim, M. (2017). Principal Component Analysis untuk Dimensi Reduksi Data Clustering Sebagai Pemetaan Persentase Sertifikasi Guru di Indonesia. *Seminar Nasional Teknologi Informasi Komunikasi Dan Industri*, 201–208.

Silvi, R. (2018). Analisis Cluster dengan Data Outlier Menggunakan Centroid Linkage dan K-Means Clustering untuk Pengelompokan Indikator HIV/AIDS di Indonesia. *JURNAL MATEMATIKA MANTIK*, *4*(1), 22–31.

Soemartini, S., & Supartini, E. (2017). *Analisis K-Means Cluster Untuk Pengelompokan Kabupaten/Kota di Jawa Barat Berdasarkan Indikator Masyarakat*.

Supranto, J. (2010). Analisis Multivariat Arti dan Interprestasi, cet. Kedua. *Jakarta: Rineka Cipta*.

Wijayanto, A. W., & Takdir. (2014). Fighting cyber crime in email spamming: An evaluation of fuzzy clustering approach to classify spam messages. *2014 International Conference on Information Technology Systems and Innovation (ICITSI)*, 19–24.